

# **Cross-Lingual and Cross-Modal Limitations of Large Language Models**

by

Senyu Li

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science  
University of Alberta

© Senyu Li, 2024

# Abstract

Large Language Models (LLMs), including Vision Large Language Models (VLLMs), herald the coming of a new research epoch in machine learning and computational linguistics. Despite most LLMs being predominantly trained on English, their proficiency in various languages has been confirmed by many studies. Nonetheless, critical questions remain about their performance consistency across different languages. A similar concern is raised for VLLMs regarding their performance disparities across various modalities. Moreover, while the remarkable competence of LLMs in solving downstream tasks is widely acknowledged, they still fall short of satisfactory performance in several tasks, requiring further experimentation for deeper insights. In this thesis, we investigate the phenomenon of cross-language generalization in LLMs by employing a novel prompt back-translation method. We investigate the interactions and comparisons between text and image modalities by introducing a new concept called cross-modal consistency and propose a quantitative evaluation framework based on this concept. Additionally, we evaluate the performance of an LLM on two specific linguistic tasks: Lexicalization Generation and Lexical Gap Detection. We have also developed a novel algorithmic approach for comparative analysis. The findings reveal that LLMs face challenges in providing accurate results for translation-variant tasks, reveal a significant inconsistency between vision and language modalities within GPT, and show that ChatGPT underperforms in the two evaluated downstream tasks, being significantly outperformed by our rule-based method.

# Preface

The work presented in Chapter 2 is published as “Don’t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs” (X. Zhang et al. 2023a) which I presented in person at a parallel oral session at EMNLP 2023. I conducted all the experiments described in the chapter.

The work presented in Chapter 3 is entitled “Cross-Modal Consistency in Large Multimodal Models: An In-Depth Study with GPT-4V” (X. Zhang et al. 2023b) and is now in submission. I conducted all the experiments described in the chapter.

The work presented in Chapter 4 is entitled “Translation-based Lexicalization Generation and Lexical Gap Detection: Application to Kinship Terms” (S. Li et al. 2024) and is now in submission. I conducted all the experiments described in the chapter.

# Acknowledgements

I would like to express my sincere appreciation to my supervisor, Professor Grzegorz Kondrak, for his support and guidance not only throughout the thesis but also for teaching me the spirit and ability of research. I also want to show my great appreciation to Xiang Zhang, Ning Shi and Bradley Hauer for their assistance, guidance, feedback and valuable advice throughout my research experience. Finally, I would like to thank my parents for their constant support throughout my studies.

This thesis was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

# Table of Contents

|          |                                                            |          |
|----------|------------------------------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                                        | <b>1</b> |
| 1.1      | Background . . . . .                                       | 2        |
| 1.2      | Thesis Statement . . . . .                                 | 3        |
| 1.3      | Evaluating LLMs’ Cross-lingual Consistency . . . . .       | 4        |
| 1.4      | Evaluating VLLMs’ Cross-modal Consistency . . . . .        | 4        |
| 1.5      | Evaluating LLMs on LexGen and LexGap tasks . . . . .       | 5        |
| 1.6      | Contributions . . . . .                                    | 6        |
| <b>2</b> | <b>The Evaluation of Multilingual Capabilities of LLMs</b> | <b>7</b> |
| 2.1      | Background . . . . .                                       | 9        |
| 2.2      | Categorizing Language-Dependent Tasks . . . . .            | 10       |
| 2.2.1    | Categorization by Task Properties . . . . .                | 11       |
| 2.2.2    | Categorization by Translatability . . . . .                | 12       |
| 2.3      | Methods . . . . .                                          | 13       |
| 2.3.1    | Prompt Translation . . . . .                               | 14       |
| 2.3.2    | Response Back-Translation . . . . .                        | 14       |
| 2.3.3    | Identifying Multilingual Types . . . . .                   | 16       |
| 2.4      | Experiments . . . . .                                      | 17       |
| 2.4.1    | Datasets . . . . .                                         | 17       |
| 2.4.2    | Metrics . . . . .                                          | 19       |
| 2.4.3    | Results on TE Tasks . . . . .                              | 19       |
| 2.4.4    | Cover Letters . . . . .                                    | 20       |
| 2.4.5    | Results on Puns . . . . .                                  | 22       |
| 2.4.6    | Analysis of Results . . . . .                              | 23       |
| 2.5      | Discussion and Future Directions . . . . .                 | 24       |
| 2.6      | Threats to Validity . . . . .                              | 25       |
| 2.7      | Conclusion . . . . .                                       | 26       |

|          |                                                              |           |
|----------|--------------------------------------------------------------|-----------|
| <b>3</b> | <b>The Evaluation of Multi-Modality Capabilities of LLMs</b> | <b>27</b> |
| 3.1      | Related Work . . . . .                                       | 29        |
| 3.2      | Preliminaries and Key Concepts . . . . .                     | 30        |
| 3.3      | Method . . . . .                                             | 32        |
| 3.3.1    | WorkFlow . . . . .                                           | 33        |
| 3.3.2    | Metrics . . . . .                                            | 33        |
| 3.4      | Experiments . . . . .                                        | 34        |
| 3.4.1    | Data Construction . . . . .                                  | 34        |
| 3.4.2    | Experiment Details. . . . .                                  | 37        |
| 3.4.3    | Main Results . . . . .                                       | 39        |
| 3.4.4    | Ablation Study on Content Extraction from Images . . . . .   | 40        |
| 3.5      | Vision-Depicting-Prompting (VDP) . . . . .                   | 41        |
| 3.5.1    | Prompting Details . . . . .                                  | 42        |
| 3.5.2    | Experiment Results for VDP . . . . .                         | 43        |
| 3.6      | Threats to Validity . . . . .                                | 45        |
| 3.7      | Conclusion . . . . .                                         | 45        |
| <b>4</b> | <b>Lexicalization Generation and Lexical Gap Detection</b>   | <b>46</b> |
| 4.1      | Theoretical Framework . . . . .                              | 48        |
| 4.1.1    | Linguistic Background . . . . .                              | 49        |
| 4.1.2    | Definitions . . . . .                                        | 50        |
| 4.1.3    | Disjunctive Triples . . . . .                                | 52        |
| 4.1.4    | Propositions . . . . .                                       | 53        |
| 4.2      | Taxonomy . . . . .                                           | 54        |
| 4.3      | Methodology . . . . .                                        | 56        |
| 4.3.1    | Candidate Generation . . . . .                               | 57        |
| 4.3.2    | Candidate Filtering . . . . .                                | 57        |
| 4.4      | Experiments . . . . .                                        | 59        |
| 4.4.1    | Data . . . . .                                               | 59        |
| 4.4.2    | Seed Words and Glosses . . . . .                             | 60        |
| 4.4.3    | Evaluation and Comparison Methods . . . . .                  | 61        |
| 4.4.4    | Results . . . . .                                            | 62        |
| 4.4.5    | Error Analysis . . . . .                                     | 63        |
| 4.4.6    | Ablation study . . . . .                                     | 64        |
| 4.5      | Threats to Validity . . . . .                                | 65        |
| 4.6      | Conclusion . . . . .                                         | 65        |

|                                                                 |           |
|-----------------------------------------------------------------|-----------|
| <b>5 Conclusion</b>                                             | <b>66</b> |
| <b>Appendix A: Generations of Large Language Models</b>         | <b>75</b> |
| A.1 Motivation in the designed prompts . . . . .                | 75        |
| <b>Appendix B: Dataset Samples</b>                              | <b>81</b> |
| B.1 Math Equation Solving (Easy) Dataset . . . . .              | 81        |
| B.2 Math Equation Solving (Hard) Dataset . . . . .              | 83        |
| B.3 LogicQA Dataset . . . . .                                   | 85        |
| B.4 MMLU Dataset . . . . .                                      | 87        |
| B.5 Table Understanding Dataset . . . . .                       | 89        |
| B.6 Math Reasoning Dataset . . . . .                            | 92        |
| B.7 State Machine Dataset . . . . .                             | 94        |
| B.8 Motivation in the designed prompts . . . . .                | 95        |
| B.9 Design decision of data instances in each dataset . . . . . | 96        |
| <b>Appendix C: Prompts in Chapter 4</b>                         | <b>97</b> |
| C.1 Motivation in the designed prompts . . . . .                | 97        |
| <b>Appendix D: BN and WN Synsets for Kinship Concepts</b>       | <b>99</b> |

# List of Tables

|     |                                                                                                             |    |
|-----|-------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Accuracy for TE tasks: math reasoning (MR), commonsense reasoning (CSR), and knowledge access (KA). . . . . | 19 |
| 2.2 | The frequency of different sign-offs in 250 different Chinese cover letters generated by ChatGPT. . . . .   | 21 |
| 2.3 | Accuracy on pun detection (P-Acc) and location (L-Acc). . . . .                                             | 22 |
| 3.1 | Test results for vision-language consistency datasets. . . . .                                              | 38 |
| 3.2 | Result of performing OCR on the all images of experimented task instances. . . . .                          | 41 |
| 3.3 | Conditional vision-language consistency score given the OCR results.                                        | 42 |
| 3.4 | Result of VDP prompting. . . . .                                                                            | 44 |
| 4.1 | An incorrect translation from Chinese to English from Google Translate.                                     | 47 |
| 4.2 | The glosses and the number of generated nodes for each root concept.                                        | 55 |
| 4.3 | Results (in %) on the kinship database. . . . .                                                             | 61 |
| 4.4 | Gloss-context templates used to obtain candidate lexicalizations. . . .                                     | 65 |
| A.1 | BERT similarity scores for explanations in TE tasks. . . . .                                                | 76 |
| A.2 | Cover letters generated by ChatGPT with the same prompt in English                                          | 77 |
| A.3 | Cover letters generated by ChatGPT with the prompt in Chinese. . .                                          | 78 |
| A.4 | ChatGPT explanations of a single English pun in different languages.                                        | 79 |
| A.5 | A sample response from Llama2 to a knowledge access question. . . .                                         | 80 |
| B.1 | Sample 1 of Math Equation Solving (Easy) Dataset in The Modality of Text . . . . .                          | 81 |
| B.2 | Sample 2 of Math Equation Solving (Easy) Dataset in The Modality of Text . . . . .                          | 82 |
| B.3 | Sample 1 of Math Equation Solving (Hard) Dataset in The Modality of Text . . . . .                          | 83 |
| B.4 | Sample 2 of Math Equation Solving (Hard) Dataset in The Modality of Text . . . . .                          | 84 |

|      |                                                                     |     |
|------|---------------------------------------------------------------------|-----|
| B.5  | Sample 1 of LogicQA Dataset in The Modality of Text . . . . .       | 85  |
| B.6  | Sample 2 of LogicQA Dataset in The Modality of Text . . . . .       | 86  |
| B.7  | Sample 1 of MMLU Dataset in The Modality of Text . . . . .          | 87  |
| B.8  | Sample 2 of MMLU Dataset in The Modality of Text . . . . .          | 88  |
| B.9  | Sample 1 of Table Understanding Dataset in The Modality of Text .   | 89  |
| B.10 | Sample 2 of Table Understanding Dataset in The Modality of Text .   | 91  |
| B.11 | Sample 1 of Math Reasoning Dataset in The Modality of Text . . . .  | 92  |
| B.12 | Sample 2 of Math Reasoning Dataset in The Modality of Text . . . .  | 93  |
| B.13 | Sample 1 of State Machine Dataset in The Modality of Text . . . . . | 94  |
| B.14 | Sample 2 of State Machine Dataset in The Modality of Text . . . . . | 95  |
| D.1  | Mapping from kinship concept to WordNet synset and BabelNet synset  | 102 |

# List of Figures

|     |                                                                                                                                           |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | The three types of bilingual lexical representations. . . . .                                                                             | 8  |
| 2.2 | Three categories of NLP tasks. . . . .                                                                                                    | 10 |
| 2.3 | A TE task (common sense reasoning) and a TV task (pun detection). . . . .                                                                 | 15 |
| 2.4 | An overview of our prompt translation and response back-translation methods. . . . .                                                      | 15 |
| 2.5 | Flowchart for detecting multilingual types. . . . .                                                                                       | 17 |
| 3.1 | Visualization of the performance gap between the modality of text and image in seven different tasks. . . . .                             | 28 |
| 3.2 | Illustration of the concept of cross-modal consistency. . . . .                                                                           | 32 |
| 3.3 | An Overview of the Components of Our Vision-Language Consistency Dataset. . . . .                                                         | 34 |
| 3.4 | Overview of the VDP Method. . . . .                                                                                                       | 40 |
| 4.1 | An illustration of a hierarchical structure of kinship concepts (top), and a concept-language matrix of lexicalizations (bottom). . . . . | 49 |
| 4.2 | Types of concept triples. Distinct lexicalization are represented by different variables. . . . .                                         | 55 |
| 4.3 | The algorithm for generating a concept graph. . . . .                                                                                     | 56 |
| 4.4 | Pseudo-code of the algorithm for lexicalization generation and lexical gap detection. . . . .                                             | 59 |
| 4.5 | Averaged evaluation results across 10 test languages with an increasing number of filters. . . . .                                        | 64 |
| B.1 | Sample 1 of Math Equation Solving (Easy) Dataset in The Modality of Image . . . . .                                                       | 81 |
| B.2 | Sample 2 of Math Equation Solving (Easy) Dataset in The Modality of Image . . . . .                                                       | 82 |
| B.3 | Sample 1 of Math Equation Solving (Hard) Dataset in The Modality of Image . . . . .                                                       | 83 |

|      |                                                                                    |    |
|------|------------------------------------------------------------------------------------|----|
| B.4  | Sample 2 of Math Equation Solving (Hard) Dataset in The Modality of Image. . . . . | 84 |
| B.5  | Sample 1 of LogicQA Dataset in The Modality of Image . . . . .                     | 85 |
| B.6  | Sample 2 of LogicQA Dataset in The Modality of Image . . . . .                     | 86 |
| B.7  | Sample 1 of MMLU Dataset in The Modality of Image . . . . .                        | 87 |
| B.8  | Sample 2 of MMLU Dataset in The Modality of Image . . . . .                        | 88 |
| B.9  | Sample 1 of Table Understanding Dataset in The Modality of Image .                 | 89 |
| B.10 | Sample 2 of Table Understanding Dataset in The Modality of Image .                 | 90 |
| B.11 | Sample 1 of Math Reasoning Dataset in The Modality of Image . . .                  | 92 |
| B.12 | Sample 2 of Math Reasoning Dataset in The Modality of Image . . .                  | 93 |
| B.13 | Sample 1 of State Machine Dataset in The Modality of Image . . . .                 | 94 |
| B.14 | Sample 2 of State Machine Dataset in The Modality of Image . . . .                 | 95 |

# Chapter 1

## Introduction

In the realm of artificial intelligence, large language models (LLMs), including Vision Large Language Models (VLLMs) have emerged as a cornerstone, bringing a new era of machine learning and computational linguistics. These sophisticated models trained on expansive datasets, have not only redefined the boundaries of natural language processing but also have profoundly impacted various industries and academic fields. LLMs have demonstrated superior performance in applications ranging from text editing like writing assistance to medical study. LLMs and VLLMs, such as GPT-3.5 and GPT-4V, serve users from diverse countries which possess a broad spectrum of linguistic/cultural backgrounds, bringing queries in various modalities.

However, the training of LLMs is mainly focused on English, and it is a widely recognized concern that current LLMs exhibit inconsistent capabilities across different languages. However, there is a noticeable absence of comprehensive research that systematically identifies and categorizes these inconsistencies. This gap in research creates challenges in finding directions for enhancement for researchers who aim to advance these models. Similar challenges happen to VLLMs. Researchers have raised concerns about the varying performance of VLLMs when processing inputs of different modalities. However, prior studies have predominantly concentrated on evaluating performance within specific domains, with minimal investigation into the capability differences between processing information from text and image modalities. Also,

despite it has been extensively demonstrated that LLMs exhibit robust performance across a broad range of downstream tasks, they still possess insufficient comprehension of languages like the knowledge regarding lexicons.

In this thesis, a thorough exploration and analysis of the multilingual capabilities of LLMs are presented, along with the introduction of a novel evaluation protocol designed to assess the multilingualism of LLMs. Employing a similar approach, a systematic analysis was conducted to examine the consistency across different modalities in VLLMs. In order to investigate the understanding of languages and lexicons by LLMs, experimental research was carried out on two linguistic tasks. Ultimately, we developed a novel rule-based method that significantly surpassed the performance of LLMs.

The remainder of the chapter is organized as follows. First, we provide some background information related to our work. Then we present and briefly explain the thesis statement of this thesis. Finally, we briefly introduce the works that are presented in the next 3 chapters followed by the main contributions of this thesis.

## 1.1 Background

In this section, we provide an overview of some key concepts. Specifically, we introduce Large Language Models and Vision-Large Language Models. Large language models represent a groundbreaking advancement in the field of artificial intelligence, specifically within natural language processing. These models, trained on extensive datasets comprising vast types of text, have the ability to understand, generate, and interact with human language in a remarkably coherent and contextually relevant manner. Large language models are characterized by their immense number of parameters – often in the billions – these models leverage deep learning techniques, particularly neural networks, to capture the details and subtleties of language. The advantages of large language models, such as OpenAI’s GPT series, lies in their capacity to perform a wide range of language tasks, from translation and summarization

to question-answering and creative writing, thereby pushing the boundaries of machine understanding and the generation of human language. Vision-large language models, exemplified by systems like DALL-E, CLIP, and their successors, have revolutionized the way machines understand and interpret the visual world in conjunction with textual information. VLLMs are trained on vast datasets containing millions of images and text pairs, allowing them to generate, modify, or interpret visual content based on textual descriptions with remarkable accuracy. This synthesis of visual and linguistic understanding opens up unprecedented applications in diverse fields such as digital art creation.

## 1.2 Thesis Statement

The thesis statement of this thesis is: *“Current LLMs and VLLMs, such as GPT-3.5 and GPT-4V, exhibit inconsistent capabilities across different languages and modalities, leading to potential limitations in performance even in language/modality-independent tasks, making LLMs more effective as assistants rather than problem solvers”.*

To grasp the concept of “inconsistent capabilities across different languages” in a LLM, consider this example: when prompted in English with “Brandon sold 86 geckos last year. He sold twice that many the year before. How many geckos has Brandon sold in the last two years?”, with the correct answer being “258”, the LLM responds with “258”. However, if the same question is posed in French, the response shifts to “129”<sup>1</sup>.

Similarly, the notion of “inconsistent capabilities across different modalities” in a VLLM can be illustrated through the following scenario. When asking “Give only the answer, no steps. John is twice as old as Mary and half as old as Tonya. If Tanya is 60, what is their average age?” in the modality of text, with the correct answer being “35”, the VLLM identifies the answer as “40”. Yet, if the same question is conveyed via the modality of image, like a screenshot containing the prompt, the

---

<sup>1</sup>The experiments were conducted on May 8th, 2023

VLLM answers “45”<sup>2</sup>.

The idea of using LLMs as assistants rather than problem solvers can be approached by using them not for direct task execution but as resources for information or raw/refined input. Subsequently, other methods, such as rule-based approaches, could be employed to tackle the task using the insights provided by the LLMs.

### 1.3 Evaluating LLMs’ Cross-lingual Consistency

In the work that is presented in Chapter 2, we propose a systematical method for evaluating the multilingual capabilities of LLMs. To enable a thorough analysis, we categorize language-dependent abilities into three distinct categories based on how the choice of language affects performance: Reasoning (least impact), Knowledge Access, and Articulating (most impact). we explore a set of carefully selected tasks from these categories, evaluating the multilingual capability of an LLM through a novel prompting technique named Response Back Translation (RBT). By analyzing the responses generated, we can not only assess the LLM’s multilingual performance but also identify the specific type of multilingualism they exhibit. For instance, we investigate the performance of LLMs on the task of pun detection, which is a task heavily dependent on the target language. The experiment results indicate that the well-known LLM “GPT” conveys a noticeable bias toward English and is characterized by a “translator-like” behaviour when the queries are from languages other than English.

### 1.4 Evaluating VLLMs’ Cross-modal Consistency

In the research that is presented in Chapter 3, we move beyond the conventional scope of assessing multimodal systems, which typically involves evaluating them through independent downstream tasks and presenting their respective scores. This study’s

---

<sup>2</sup>The experiments were conducted on October 15th, 2023

primary focus centers on the consistencies in the capabilities of various modalities, with particular emphasis on vision and language. To facilitate a thorough analysis, this work has introduced the concept of *cross-modal consistency*. Following this, we developed an evaluation framework and a comprehensive vision-language parallel dataset that covers seven tasks. Each task is specifically designed to highlight different aspects of vision and language skills. Our experimentation with the GPT-4V model on this dataset reveals notable inconsistencies between its abilities to solve questions in vision and language modality. The findings demonstrate that GPT-4V’s performance varies significantly depending on whether the same task instance is presented in one modality versus the other.

## 1.5 Evaluating LLMs on LexGen and LexGap tasks

The work introduced in Chapter 4 includes 2 tasks, Lexicalization Generation (LexGen) and Lexical Gap detection (LexGap). Given a language-independent concept and a target language, the output of a LexGen system is either a lexicon that lexicalizes the given concept or indicates the concept is a lexical gap. LexGap is a binary identification task which can be reduced from LexGen. Given the same input as LexGen, the output of a LexGap system is either the provided concept is a gap or not. To employ GPT-3.5-turbo on these tasks, We use *in-context learning* Brown et al. 2020, a technique allowing large language models to execute tasks based on examples included in their input instructions, without the need for external updates or specific model training. We prompt GPT-3.5-turbo with the template shown in Table 4.4 in the Appendix. What can be used to do the comparison is a generally applicable algorithmic method we propose to automatically generate concept lexicalizations, which is based on machine translation and hypernymy relations between concepts. An absence of a lexicalization implies a lexical gap. We conducted the experiments on kinship terms, which makes a suitable case study due to their explicit definitions and clear structure. Empirical evaluations demonstrate that our approach yields higher

accuracy than GPT-3.5-turbo. This indicates that even one of the top LLMs, GPT-3.5-turbo, falls short in these two downstream tasks which can be done better by algorithmic approaches grounded in robust linguistic theories and propositions.

## 1.6 Contributions

In this thesis, we propose a systematic way of qualitatively and quantitatively evaluating the multilingual capabilities of LLMs. we investigate the phenomenon of cross-language generalization in LLMs, wherein limited multilingual training data leads to advanced multilingual capabilities. Under the same thought behind cross-lingual consistency, we thus introduce a new concept: cross-modal consistency. Following this concept, we propose a quantitative evaluation framework accompanied by a cross-modal dataset which is used to evaluate the cross-modal consistency of vision-large language models.

Furthermore, we evaluate the capabilities of LLMs on two downstream linguistic tasks, LexGap and LexGen, to investigate LLMs’ depth of understanding of languages. this study reveals the deficient ability of LLMs to solve language-specific lexical questions. To progress further on these two tasks, we developed a novel rule-based method that leverages theories built based on the hypernymy/hyponymy relationships between language-independent concepts. An empirical experiment showed that our method can outperform GPT-3.5-turbo by a large margin on these tasks.

## Chapter 2

# The Evaluation of Multilingual Capabilities of LLMs

The study of bilingualism has long been a topic of interest among linguists (Yu et al. 2022; Hoffmann 2014), as it provides insight into the mechanisms of language acquisition and processing. Furthermore, research on multilingualism has contributed to the development of more effective machine learning models, such as neural translation systems (Zou et al. 2013). With the rise of large language models (LLMs), researchers have discovered many emergent properties (Wei et al. 2022a) in these models, and have used them for a variety of purposes (Wei et al. 2022b). However, the multilingual ability of these models has not been extensively studied.

Previous research has shown that large language models, such as GPT, are capable of performing a wide variety of language tasks when the task is presented in English (Qin et al. 2023). However, investigations into the multilingual language abilities of these models have been limited. F. Shi et al. explore this topic by applying the models to multilingual datasets, and measuring performance differences across languages. However, they do not explore the underlying mechanisms of how LLMs perform different tasks, nor how this affects the results. Moreover, most LLMs (Brown et al. 2020; Touvron et al. 2023) are trained on datasets that are heavily skewed towards English, which leaves open the question of how multilingual abilities in such models are acquired.

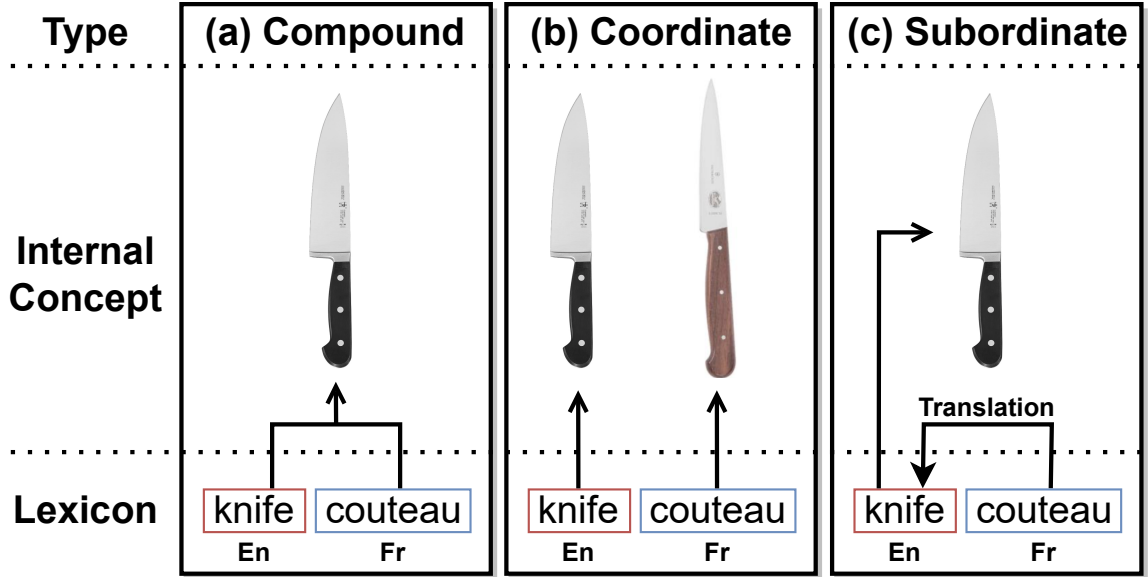


Figure 2.1: The three types of bilingual lexical representations.

In this study, we present a systematic approach to analyzing the multilingual capabilities of LLMs. To facilitate a comprehensive analysis, we propose categorizing language-dependent abilities into three distinct categories which vary in the impact of language choice on the performance: Reasoning (least impact), Knowledge Access, and Articulation (most impact). We investigate a carefully selected set of tasks from these three categories by evaluating the multilingual abilities of an LLM using a novel prompting method which we call response back-translation (RBT). By comparing the generated answers, we can both measure multilingual performance of the LLM, but also determine the type of multilinguality they exhibit. For example, we examine the capabilities of LLMs on pun detection, a highly language-dependent task.

The results of our experiments show that the popular LLM “GPT”: (1) achieves higher performance when the task is presented in English; (2) achieves higher performance on tasks that can be translated without altering the correct output; and (3) exhibits a mixture of *coordinate* and *subordinate* bilingualism.

Our main contributions are:

- We present a first-of-its-kind quantitative and qualitative analysis of the multi-

lingual abilities of LLMs.

- We propose two novel task categorizations to facilitate the multilingual ability analysis.
- Our work is the first to investigate LLMs with respect to a linguistic typology of bilingualism and multilingualism.

## 2.1 Background

There are three types of bilingual lexical representation: compound, coordinate, and subordinate lexical representation (D’Acierno 1990; DONG et al. 2005; Pavlenko 2009). Figure 2.1 illustrates these representations, showing how individuals with different types of English-French bilingual lexical representation might internally represent the concept of “knife”.

*Compound* bilingual lexical representation mostly emerges among individuals who learn two languages simultaneously from birth. In this case, both languages are equally dominant and integrated, blurring any clear distinction between them and giving the impression of a single unified language (Moradi 2014). Compound bilingual lexical representation entails a shared mental representation of lexicons across both languages they acquire, and bilinguals that carry compound bilingual lexical representation are the most flexible in their use of multiple languages, exhibiting the ability to switch between languages without losing consistency in linguistic tasks (De Groot et al. 1991).

In contrast, *coordinate* bilingual lexical representation maintains separate mental representations for the lexicon of each language. This separation leads to differences when tasks are performed under different language settings (Jakobovits 1968).

Finally, *subordinate* bilingual lexical representations is characterized by a “translator” behaviour (Marcos 1976). This type of bilingual lexical representation is characterized by a single lexicon representation that is linked to their dominant lan-

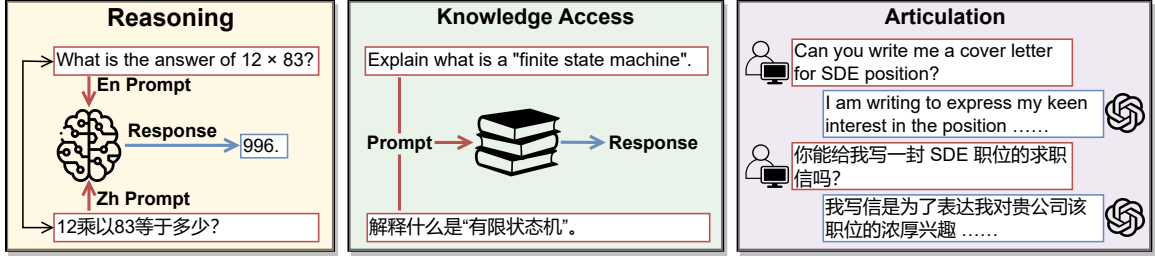


Figure 2.2: Three categories of NLP tasks.

guage (Lörscher 2012). When performing tasks in languages other than their dominant one, bilinguals that have subordinate bilingual lexical representation tend to rely on translating the task into their dominant language, formulating an answer in the dominant language, and then translating that answer back into the language of the task. As a result, bilinguals with subordinate bilingual lexical representation may experience lower proficiency in communicating and completing tasks in the second, subordinate language.

Despite the demonstration in prior work of consistent multilingual performance in many large language models (F. Shi et al. 2023), it remains unclear what are the multilingual lexical representations inside LLMs. It is an open question whether the LLMs exhibit a representation of knowledge shared across both languages (compound), separate representations for each language (coordinate), or whether they rely on a translation processes (subordinate). We develop an experimental framework aimed at using performance on various natural language processing tasks to determine how the multilingual abilities of LLMs relate to these categories.

## 2.2 Categorizing Language-Dependent Tasks

Language ability is a multifaceted concept encompassing various tasks and aspects (Wei et al. 2022a). It is therefore difficult to assess a model’s capabilities with respect to a given language. To facilitate such assessment, researchers have often classified tasks into distinct categories (Khurana et al. 2023), such as parsing and summarization.

However, the delineation of such categories often lacks systematic criteria, particularly in the context of multilingual analysis.

In this section, we propose a novel approach to categorizing NLP tasks, which is better suited to analysis of multilingual abilities. The categorization is two dimensional: one dimension is based on the linguistic knowledge necessary to complete the task (Section 2.2.1), the other on how the task is impacted by the language in which it is presented (Section 2.2.2).

### 2.2.1 Categorization by Task Properties

We classify NLP tasks into three distinct categories: Reasoning, Knowledge Access, and Articulation. This division is based on the extent to which performance on each task is influenced by the model’s capabilities with the language used. Figure 2.2 provides an overview of this categorization.

**Reasoning** The first category includes tasks that are minimally influenced by language, on which consistent performance is expected across languages. Reasoning tasks involve logical and rational thinking to solve problems based on available information and logical principles. Examples include mathematical problem-solving (Lu et al. 2023), coding (J. Li et al. 2023), and common sense reasoning (Sap et al. 2020). These tasks can be performed using universal language elements, such as mathematical symbols, or rely on general life experience and common sense, which can be acquired without language. For example, answering the question *“If I drop an apple, which direction will it go?”* relies more on understanding gravity than on language-specific knowledge.

**Knowledge Access** LLMs have the capability to function as knowledge bases (KBs) by storing knowledge extracted from training data (Heinzerling et al. 2021). Knowledge Access tasks depend on the ability to access this knowledge and formulate accurate responses based on it. While the underlying knowledge may not be language

dependent, models may be less reliable in retrieving and utilizing knowledge learned in a language other than the one used to formulate the task. Examples of Knowledge Access tasks include factual knowledge checking (De Cao et al. 2021), knowledge-focused question answering (Zhen Wang 2022), and named entity recognition (Malmasi et al. 2022).

**Articulation** Much of everyday human conversation is highly language-dependent, as it involves the pragmatics and cultural nuances of the spoken language. For instance, writing a cover letter in English significantly differs from writing one in Japanese, due to the distinct social norms and conventions associated with those languages. The Articulation category includes tasks that are heavily influenced by the language choice, such as summarization (Nenkova et al. 2012), dialogue generation (Ni et al. 2023), paraphrasing (Zhou et al. 2021), and style writing (Jin et al. 2022). These tasks require an extensive understanding of not only language, but the associated culture, as they involve capturing and reproducing the appropriate style, tone, and manner of expression specific to a given language.

## 2.2.2 Categorization by Translatability

The second dimension of our task classification scheme involves translatability. We introduce the concepts of Translation Equivariant (TE) and Translation Variant (TV) tasks.

A function is considered *equivariant* if it commutes with a symmetry transformation. That is, applying a transformation before or after computing the function yields the same result. Formally,  $f(\cdot)$  is said to be equivariant under  $g(\cdot)$  if:

$$\forall x \in \mathcal{D}, \quad g(f(x)) = f(g(x)) \quad (2.1)$$

where  $\mathcal{D}$  represents the domain of both  $f$  and  $g$ .

We denote translation as a transformation  $g$  that converts a given text in language A to an equivalent text in language B. In practice,  $g$  can be implemented by a machine

translation system. We further use  $f$  to denote a function which solves a given task, given an instance of that task as input. A task is considered Translation Equivariant between languages A and B if the correct output can be obtained by translating the input, and then applying a method for solving the task, or by solving the task, and then translating the output; in other words, if  $g(f(x)) = f(g(x))$ . Most of the tasks in the Reasoning and Knowledge Access categories are regarded as Translation Equivariant since the correct output does not depend on the chosen language. Figure 2.3 shows an example where the answer to the question posed in English remains the same in Chinese, regardless of in which order the translation system and the question answering system are applied.

A task which is not Translation Equivariant is Translation Variant. For such tasks, translating the input may change the correct output. TV tasks rely heavily on the language used, and include many tasks in the Articulation category. Representative TV tasks that we investigate in our experiments are *letter writing* and *pun understanding*. The former is subject to the conventions of the specific language and culture, while the latter involves word polysemy, which is often sensitive to translation. Figure 2.3 shows an example where a pun is present in the original English input, but not in the Spanish translation, making the classification dependent upon the order in which translation is applied.

## 2.3 Methods

In this section, we present our approach to analyzing the multilingual ability of LLMs. Our methods involve *prompt translation* (PT) and *response back-translation* (RBT). They are designed to measure performance of an LLM, and its consistency across languages. In our experiments, we apply these methods to both TE and TV tasks, with the aim of determining the type of bilingualism (compound, coordinate, or subordinate) exhibited by an LLM.

### 2.3.1 Prompt Translation

Multilingual datasets are unavailable for many tasks. However, with state-of-the-art machine translation (MT) systems and LLMs, we can translate monolingual datasets for TE tasks to generate parallel multilingual parallel data with minimal loss of information (Whitehouse et al. 2023; F. Shi et al. 2023). This is the key intuition behind prompt translation (PT); an example is shown in Figure 2.4a, where an English multiple choice question, and its possible answers, are translated to Chinese. The LLM is then prompted, and the response is given and evaluated, in Chinese. Prompting in distinct languages is performed in independent LLM sessions.

We measure the differences in multilingual task performance by comparing the answers given by the LLM in each language. Assuming that the LLM successfully learns to solve a TE task in a language-independent way, the pairwise responses for each instance should be the same after the translation (regardless of whether it is correct or incorrect). This is because TE tasks, such as mathematical problem solving, do not depend on the language used to query the LLMs, as the solution does not depend on the language used to express the problem.

### 2.3.2 Response Back-Translation

One of the goals of our work is to understand what the consistency of LLM output across languages tells us about the model, and to determine the type of bilingualism an LLM exhibits. This is crucial for individuals who use LLMs for multilingual tasks, as it can impact the way task results are generated, and affect the quality and consistency of the results. For example, a network exhibiting subordinate bilingualism would produce output that appears to be the result of translation, rather than resembling text generated by a native speaker of the output language.

To quantitatively measure how reasoning is performed, we propose a prompting method based on back-translation, as illustrated in Figure 2.4b. Similar to prompt translation (Section 2.3.1), we begin by translating the instance to the target lan-

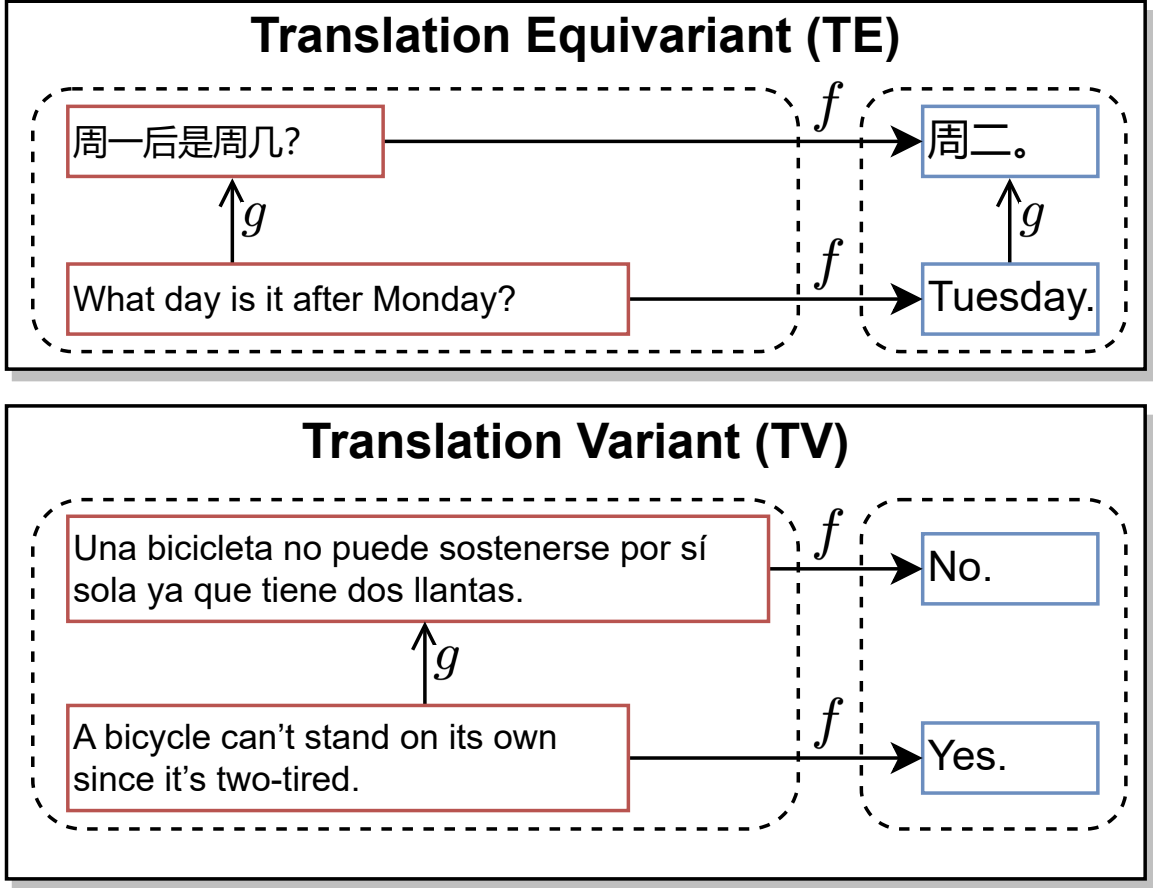


Figure 2.3: A TE task (common sense reasoning) and a TV task (pun detection). Translation is denoted by  $g$ , and  $f$  is the solver function.

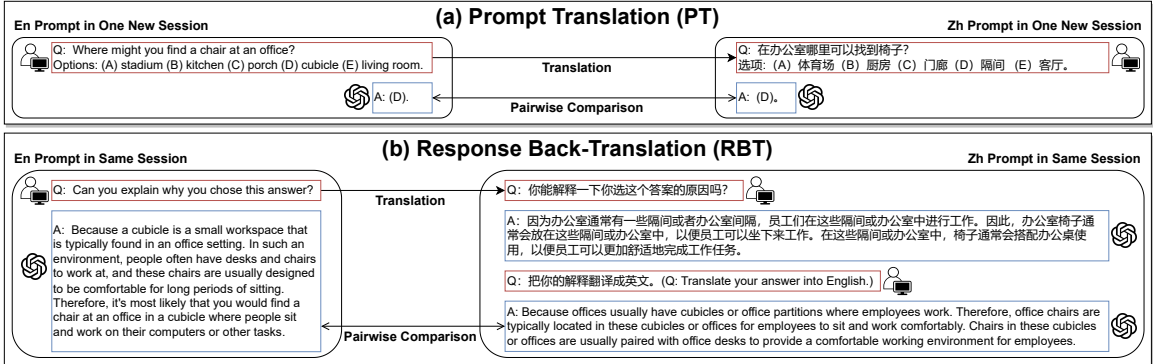


Figure 2.4: An overview of our prompt translation and response back-translation methods.

language, and prompting the LLM to produce a response in that language. After obtaining output from the LLM, regardless of the language, we further prompt the

LLM to generate an explanation for its output (e.g., “*Explain how you obtain this result*”); and then translate the output of the LLM back to the original language. We then compare the explanation given in the source language to the explanation back-translated from the target language.

If the LLM is performing translation-based reasoning, the reasoning process is conducted in one language and then translated into another. Since the internal reasoning of the LLM can be partially observed through the output explanation, back-translating such explanations into the source language allows us to compare the internal reasoning used to solve the problem in each language. High similarity of explanations should indicate homogeneity in using the same internal reasoning process to perform the task in both languages. On the other hand, dissimilarity in the reasoning process across languages should be reflected in a lower explanation similarity.

### 2.3.3 Identifying Multilingual Types

In our investigation, we employ both Prompt Translation (PT) and Response Back-Translation (RBT) to analyze how an LLM solves TE and TV tasks in different languages. As depicted in the first two steps in Figure 2.5, a compound LLM should exhibit consistent results on TE tasks with both methods. This is because a compound model performance does not depend on the language in which a question is presented. Conversely, subordinate and coordinate types of networks are expected to yield somewhat different results on TE tasks. A coordinate model accesses distinct representations in different languages, which may result in different reasoning and answers. Finally, a subordinate model heavily depends on an internal translation process, which we expect to lead to some deterioration of output quality across languages.

Testing on TV tasks provide additional information, which can be used to distinguish between coordinate and subordinate models. A coordinate LLM is expected to reason differently for each language, which may yield different outputs, whether

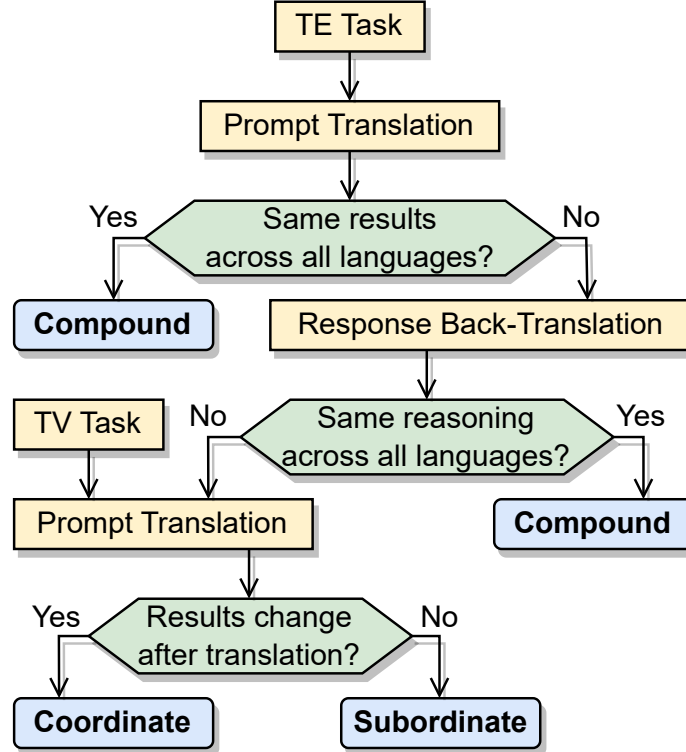


Figure 2.5: Flowchart for detecting multilingual types.

correct or not. In contrast, a pure subordinate model is expected to reason only in the dominant language, producing relatively similar results in different languages, regardless of whether the correct output is preserved after translation.

## 2.4 Experiments

We apply the methodology proposed in Section 2.3 to TE and TV tasks. As our LLM, we use ChatGPT, via the official web application<sup>1</sup>, due to its availability.

### 2.4.1 Datasets

**Reasoning** We use 50 instances selected at random from each of two datasets: GSM8K (Cobbe et al. 2021), which contains 7,500 training and 1,000 test problems, and CommonsenseQA (Talmor et al. 2019), which contains 12,247 questions. We used ChatGPT to translate these instances into French, Spanish, German, Japanese, and

<sup>1</sup><https://chat.openai.com/>

Chinese. GSM8K is a dataset of grade-school math problems. Each problem consists of a question and a multiple-choice answer. CommonsenseQA is a question answering dataset for testing logic and common sense. Each instance consists of a question and five answer choices, only one of which is considered correct.

**Knowledge Access** WebQuestions is a dataset of 6,642 question-answer pairs extracted from Freebase (Bordes et al. 2014). An example question is “*Where is the Thames River located?*” to which the correct answer is *London*. To simplify the evaluation, and avoid the issue of extracting answers from ChatGPT’s often verbose responses, we manually converted 50 randomly selected instances into the multiple-choice format used by CommonsenseQA. To create plausible incorrect answers (*distractors*), we randomly selected four incorrect candidate answers from sets of world city names<sup>2</sup> and celebrity names<sup>3</sup> (correct answers in this dataset are all either city names or celebrity names). This yielded a set of 50 multiple choice questions with five possible answers each (one correct, four incorrect). We translated the English instances into five other languages via ChatGPT.

**Puns** We randomly selected 80 positive and 80 negative instances each from the English, French, and Spanish instances in the JOKER@CLEF 2022 dataset (Ermakova et al. 2022). Each instance is annotated with a yes/no classification as to whether it contains a pun, and the pun location, if a pun is present. An example English instance is “*Astronauts work in a nice atmosphere*” for which the pun location is the word *atmosphere*. We used ChatGPT to translate the French and Spanish instances into English, and the English instances into French, Spanish, German, Japanese, and Chinese. This yields 10 balanced sets of 160 instances each (three original and seven translated).

---

<sup>2</sup><https://simplemaps.com/data/world-cities>

<sup>3</sup>[https://github.com/janester/mad\\_libs](https://github.com/janester/mad_libs)

| Task | En          | Fr          | De   | Es   | Ja   | Zh   |
|------|-------------|-------------|------|------|------|------|
| MR   | <b>0.90</b> | 0.80        | 0.78 | 0.80 | 0.82 | 0.78 |
| CSR  | <b>0.68</b> | 0.58        | 0.52 | 0.54 | 0.48 | 0.52 |
| KA   | <b>0.96</b> | <b>0.96</b> | 0.94 | 0.94 | 0.80 | 0.68 |

Table 2.1: Accuracy for TE tasks: math reasoning (MR), commonsense reasoning (CSR), and knowledge access (KA).

**Articulation** To test the Articulation abilities of an LLM, we prompt the model to generate a cover letter for a job application, with randomized specifications. For each prompt, we first generate the name and background of an individual, including information such as level of education, specialties, and hobbies. We then randomly select one well-known company to which cover letter is to be addressed. Finally, we select a set of topics such as “What skills would you want to develop in this role?”. Each of these randomized prompts is then provided to the LLM. The output is then manually evaluated by a native speaker of the language of the prompt. We generate 50 prompts each in English and Chinese. An example is provided in Table A.2 and Table A.3 in the appendix.

### 2.4.2 Metrics

Since ChatGPT can give different answers to the same question, we present each multiple-choice question to ChatGPT five times, and use the most frequent output for evaluation. For computing similarity between explanations, we use appendix(Devlin et al. 2019). Specifically, we translate all non-English output to English via ChatGPT, and compute the cosine similarity of the BERT embeddings of the two explanations.

### 2.4.3 Results on TE Tasks

As shown in Table 2.1, the results on TE tasks in English are on average much higher in English than in other languages. In math reasoning (MR), the least language-dependent task, the gap between English and other languages is over 10% on aver-

age. In common sense reasoning (CSR), the difference is over 15% on average. In knowledge access (KA), there is no substantial difference between English and other European languages, but accuracy on Japanese and Chinese is 16% and 28% lower, respectively. To confirm that the accuracy gap is not due to instance translation quality, we manually compared all 50 Chinese MR questions with their original English counterparts, and found no translation errors. Taken together, these results provide strong evidence that GPT is better able to reason and retrieve knowledge given an English prompt, compared to prompting in other languages. In terms of multilinguality type, the evidence is against compound multilingualism in GPT (cf., Figure 2.5), as a compound model would be expected to exhibit no substantial difference in performance across languages.

We also analyzed the BERT similarity values between explanations in different languages (cf., Table A.1 in the Appendix). In commonsense reasoning, which relies on logic and conceptual distinctions, we observe that the average BERT similarity of German, Spanish, Japanese and Chinese to French is substantially lower than the corresponding average similarity to English (0.849 vs. 0.868), while French itself is substantially more similar to English than to German (0.871 vs 0.857). We interpret this as additional evidence of the GPT’s dependence on its strong English model.

On the other hand, we observe no such trend in knowledge access questions. We hypothesize that since these problems are mostly about named entities, they tend to be more language-independent. Indeed, we observe higher performance on French, German, and Spanish, which use the Latin script, and therefore can represent named entities as English does, compared to Japanese and Chinese, which use different orthographies.

#### **2.4.4 Cover Letters**

Cover letter writing is an example of a TV articulation task. We found that cover letters generated by ChatGPT with the same set of instructions in different lan-

| Chinese     | English Translation              | Frequency |
|-------------|----------------------------------|-----------|
| 诚挚地         | Sincerely                        | 54.0%     |
| 致意          | Regards                          | 38.4%     |
| 祝愿          | Best Wishes                      | 3.6%      |
| 此致敬礼        | Salute (Proper Chinese Sign-off) | 0.8%      |
| No sign-off |                                  | 3.2%      |

Table 2.2: The frequency of different sign-offs in 250 different Chinese cover letters generated by ChatGPT.

languages exhibit relatively high BERT similarity to their English versions, ranging from 0.818 Japanese to 0.865 for German. To provide some comparison, we also computed pairwise BERT similarities between English cover letters generated with the same prompts by ChatGPT and two other LLMs, Claude and Instant-Claude, which yielded the values of 0.618 and 0.643, respectively. This indicates that the letters generated in different languages by ChatGPT are more similar to each other than the letters in English generated by different LLMs. Cover letters generated in languages other than English exhibit a written style which is closer to English than to the target language. For example, consider the cover letter shown in Table A.2 and Table A.3 in the Appendix. The expressions 阁下 “from what I have gathered” and 狂热的户外运动爱好者 “avid outdoor enthusiast” are very unnatural in Chinese, and appear characteristic of literal translations from English. The sign-off phrase 真诚的 “Sincerely” is similarly inappropriate in formal Chinese, although it is usual in English. Table 2.2 shows that less than 1% of the letters have a proper Chinese sign off.

| Language | P-Acc | L-Acc |
|----------|-------|-------|
| Es       | 0.488 | 0.697 |
| Es-En    | 0.507 | 0.714 |
| Fr       | 0.500 | 0.886 |
| Fr-En    | 0.513 | 0.813 |
| En       | 0.506 | 0.965 |
| En-Fr    | 0.500 | 0.646 |
| En-De    | 0.519 | -     |
| En-Es    | 0.488 | 0.607 |
| En-Ja    | 0.519 | -     |
| En-Zh    | 0.550 | 0.511 |

Table 2.3: Accuracy on pun detection (P-Acc) and location (L-Acc). X-Y means the puns were translated from language X to language Y before prompting.

### 2.4.5 Results on Puns

Table 2.3 shows the results on the translation-variant tasks of pun detection and location. The accuracy of pun detection is close to what we would expect from a random baseline, as ChatGPT strongly favors positive pun classifications. The sole exception is a slightly higher accuracy of 0.55 when English puns are translated into Chinese, due to a higher proportion of negative classifications.

Since few conclusions can be drawn from the pun detection results, we conducted an evaluation of the pun location results in most datasets, which required manual extraction of the location information from ChatGPT’s explanations. The results are shown in Table 2.3. The pun location accuracy on the original English puns is very high at 96.5%, but drops dramatically when the sentences are translated into other languages. When French puns are translated to English, there is likewise a drop in performance, though it is much smaller than what is observed when English puns are translated to French. However, the situation is different for Spanish puns, where

the location accuracy increases slightly after the puns are translated into English. This is surprising, as puns are often language-specific, and tend to disappear after translation.

When the prompt is not in English, evidence suggests that ChatGPT relies, at least partly, on its English capabilities for semantic interpretation. Consider the homonymous English word *bat* which has two unrelated senses, translated by different words in Chinese: 蝙蝠 for the “animal” sense, and 球拍 for the “club” sense (Hauer et al. 2020). When the original English prompt is “*What is the famous bat brand for baseball?*”, ChatGPT appears unable to distinguish between these two translations of *bat* within a Chinese prompt. Although the choice of the Chinese translation of *bat* greatly affects the meaning of the question, it does not seem to impact ChatGPT’s response. However, when we replace 蝙蝠 “animal bat” with 老虎 “tiger”, ChatGPT correctly responds that the question makes no sense. We interpret the inability of ChatGPT to differentiate between the two distinct Chinese translations of *bat* as strong evidence of subordinate bilingualism.

#### 2.4.6 Analysis of Results

The results of our experiments provide evidence that GPT exhibits a substantial degree of subordinate multilingualism. Many of its responses are what we would expect from a system which translates all input into English, formulates a response in English, and then translates this response into the input language. Since translation is an error-prone process, the resulting response accuracy is frequently lower than when the input is provided in English.

We speculate that this behavior is an artifact of GPT being trained mostly on monolingual English texts. Consequently, GPT has developed a representation of knowledge and communication that is strongly biased towards English. We conclude that since GPT is not designed to take advantage of bilingual or multilingual corpora, it is unable to create a single multilingual conceptual representation analogous to

compound multilingualism.

Moreover,

GPT has less training data for non-English languages, compared to its English training data. We postulate that this results in representations for non-English languages that are much weaker than those GPT can create for English. This often leads to lower performance on even translation-equivariant tasks when the task is not presented in English.

## 2.5 Discussion and Future Directions

Our research provides robust support for the notion that LLMs have not achieved the ideal behaviour of compound multilingualism. Even if the quality and quantity of training data in various languages were held constant, we speculate that compound multilingualism would still not be achieved, due to the inherent limitations of current data collection methods and training techniques.

Drawing a parallel to human multi-modal learning offers an intuitive understanding of why this could be the case. Consider how humans acquire concepts related to vision and language: A child grows by consistently pairing visual stimuli with linguistic cues, intertwining the two modalities over time. Consequently, it is rare to observe a mismatch between visual and linguistic perceptions. In this context, humans exhibit a highly integrated understanding of vision and their native languages. However, unless raised in a perfectly bilingual environment, individuals seldom showcase equivalent proficiency in two languages. Indeed, bilingual individuals often demonstrate cognitive variations depending on which language is in use.

A rudimentary multi-modal system can be likened to a crude fusion of a vision model trained on image data and a language model trained on text. These systems possess minimal, if any, shared representations or information overlap. Beneath the facade of a system that seemingly excels at both visual and language tasks, lie two distinct networks. Nevertheless, recent advancements in multi-modal studies, combined

with the availability of extensively captioned image data, have given rise to more sophisticated systems. These systems bridge the gap between the two modalities, moving the field closer to human-like integration.

Acquiring aligned multilingual data is a significant challenge, with the exception of some translation datasets. The majority of online articles and posts are monolingual and cannot be easily paired. Therefore, training on these multilingual corpora results in models that essentially act as an amalgamation of several independent language-specific models, with minimal information interchange, primarily anchored by the translation datasets which comprise a relatively small portion of the corpus. When corpora are disproportionately comprised of some language or set of languages, the models tend to become predominantly subordinate, with minimal coordination arising from monolingual datasets.

Moving forward, our objective is to narrow the divide between languages within a multilingual system and to cultivate language models that lean more towards a compound archetype. This will require both crafting highly parallel paired data across languages and innovating training methodologies that promote the learning of compound representations for universal concepts irrespective of the language used to express them. For the former, we intend to delve into ontology linkages. For the latter, we plan to leverage recent advancements in model training, such as contrastive learning. Our goal is to create multilingual models that are both technically sophisticated and universally adept.

## **2.6 Threats to Validity**

As the OpenAI ChatGPT website application has a limited number of prompts allowed per day and per hour, we can not apply our experiment to the whole dataset. We used GPT3.5 rather than GPT4 as our LLM since access to GPT4 was still restricted at the time the experiments were conducted. We conducted the human evaluation only in English, Spanish, and Chinese, as we did not have access to fluent

speakers of the other languages found in our test sets. At the time the experiments were conducted, the author had no control over the temperature of the generation, thus the results were not always deterministic even with the same prompt. In our experiments, we utilized translations generated by LLMs themselves, and the quality of the translation was not fully verified due to the cost. Therefore, future work could involve more translators, such as Google Translate, and conduct further experiments on these translations to address biases from different translators.

## 2.7 Conclusion

We have proposed a systematic approach to analyzing multilingual abilities of large language models. Our experiments provide new evidence for a subordinate multilingualism in GPT-3.5, with English functioning as the model’s native language. Our experimental results, supplemented by the analysis of specific examples and case studies, demonstrate that such subordinate multilingualism can limit performance even in language-independent tasks. We postulate that explicit inclusion of additional multilingual parallel corpora and multimodal datasets into the training data of LLMs could ameliorate this issue.

## Chapter 3

# The Evaluation of Multi-Modality Capabilities of LLMs

Recent large multimodal models have showcased remarkable capabilities in tasks that require the integration of multiple modalities and sources of information (Huang et al. 2023). Among these, the performance of Vision Large Language Models (VLLMs) (J. Zhang et al. 2023; Z. Yang et al. 2023) stands out, thanks to the vast amounts of image and text data available for training and the rapid progress in both computer vision and language modelling. However, due to the distinct training methodologies employed by these models, such as contrastive learning (Radford et al. 2021) and embodied image-language modeling (Driess et al. 2023), and the varying quality of training data for each modality, these networks often exhibit performance disparities across different modalities.

Previous research has extensively evaluated the performance of individual modalities in multimodal systems. For instance, (Z. Yang et al. 2023) conducted a thorough assessment of GPT-4V’s vision understanding capabilities, and (Chen et al. 2023) analyzed model’s decision-making abilities. However, assessing a model’s performance on each individual modality in isolation does not fully evaluate its true multimodal abilities. It is possible, for example, for a model to excel in numerous vision tasks but still lag significantly behind in language understanding. Moreover, simply testing performance on individual tasks provides no insight into whether and how each modality

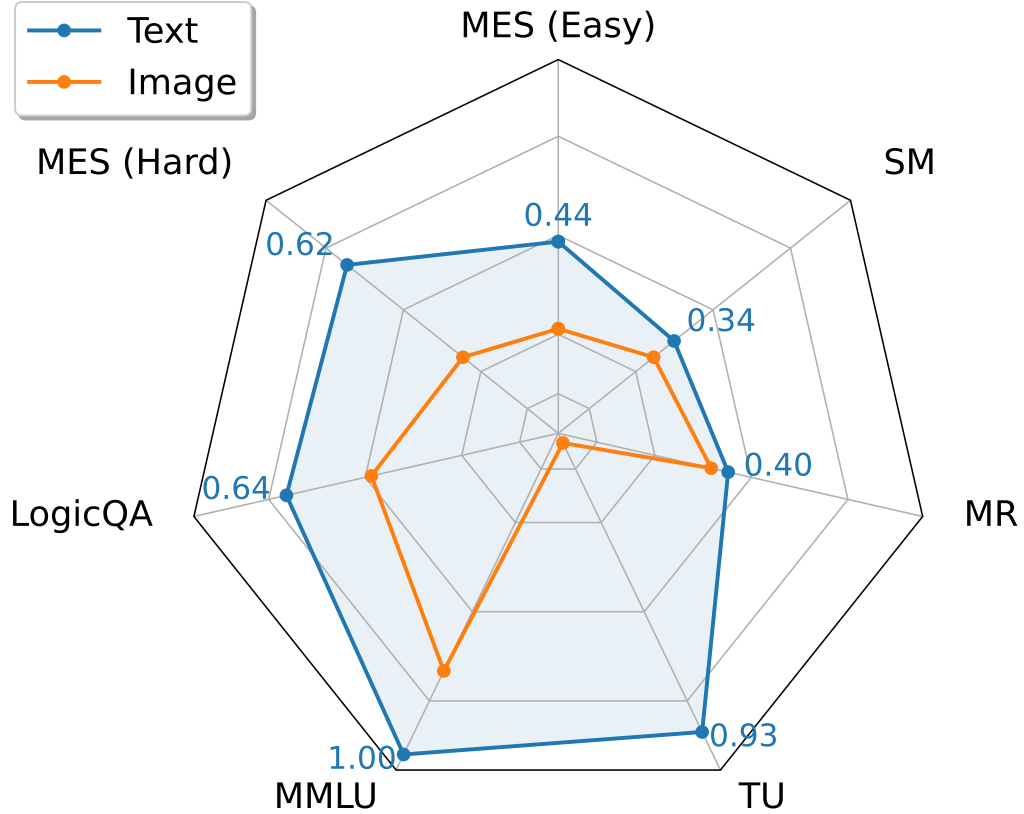


Figure 3.1: Visualization of the performance gap between the modality of text and image in seven different tasks.

of the model influences the others. Unfortunately, the cross-modality relationship is frequently overlooked in the aforementioned research.

In this study (X. Zhang et al. 2023b), we go beyond the traditional approach of simply evaluating multimodal systems through separate downstream tasks and reporting their scores. Our focus is primarily on measuring the inherent *differences* in capabilities between various modalities, with special attention to vision and language, given their prominence among other modalities. To enable a comprehensive analysis, we introduce the concept of *cross-modal consistency*, complete with a formal definition and an evaluation framework. We consider cross-modal consistency to be an essential element in the design of complex multimodal systems with neural components, as it guarantees coherence and reliability in the system’s performance. This is crucial for both interpretability and for fostering user trust.

We subsequently construct a comprehensive vision-language parallel dataset encompassing seven tasks, each designed to highlight different facets of vision and language capabilities. This dataset serves as a tool for evaluating the vision-language consistency of VLLMs. Our experiments with the GPT-4V model on the dataset reveal significant inconsistencies between its vision and language capabilities. The results indicate that its performance varies considerably depending on whether the same task instance is prompted in one modality versus the other.

Our contributions are: (1) We introduce the novel concept of cross-modal consistency, along with a comprehensive evaluation framework. This approach transcends traditional assessment methods for multimodal models, which typically evaluate each modality in isolation. (2) We develop and release seven diverse datasets, carefully designed for vision-language consistency evaluation, opening up opportunities to exploit these datasets in future research. (3) Our experiments on GPT-4V reveal a significant disparity between vision and language abilities within such a system, prompting the introduction of the Vision-Depicting-Prompting (VDP) method as a potential remedy. Our findings offer valuable guidance for more effective future use of such multimodal models.

### 3.1 Related Work

A substantial amount of effort has been dedicated to meticulous evaluation of large multimodal models such as GPT-4V. To assess the capabilities of these models across all their modalities, a wide array of tasks has been tested. E.g., researchers have scrutinized GPT-4V’s aptitude in solving problems within specialized domains, including biomedicine (Z. Liu et al. 2023), medical applications (Wu et al. 2023), and autonomous driving (Wen et al. 2023), employing intricate image inputs. Beyond these domain-specific evaluations, more general skills like chart image understanding (F. Liu et al. 2023) and optical character recognition (Y. Shi et al. 2023) have also been analyzed. However, these evaluations often focus solely on performance

metrics for each test dataset, with little or no exploration of the relative capability gaps between vision and language. In this study, our primary emphasis lies in uncovering the relative *disparities* in the abilities of multimodal models across their various modalities, rather than merely assessing absolute performance within specific tasks.

Despite the lack of cross-modal analysis for multimodal models, previous research has delved into examining cross-lingual abilities in Large Language Models (LLMs). For example, by translating task instances into different languages and analyzing the pairwise results, (X. Zhang et al. 2023a) demonstrated that models like GPT-3.5, primarily trained on English text corpora, exhibit disparities in their performance across various tasks when prompted with different languages. Specifically, these LLMs display a bias toward English. Taking inspiration from such studies, we extend our research to encompass consistency analysis across various modalities, recognizing that different languages can be regarded as distinct modalities as well. Our generalized framework sheds light on the underlying principles governing the **consistency** of multimodal models when confronted with tasks in diverse modalities, thereby contributing to a deeper understanding of their capabilities and limitations.

## 3.2 Preliminaries and Key Concepts

As “consistency” can carry different interpretations within the specific context we are addressing, a formal definition of the concept of cross-modal consistency for multimodal models is warranted. To that end, we establish an instance of task  $t$ , represented as the paired value  $(d_a, q)$ . Here,  $d_a$  represents a data element from the input space  $\mathcal{D}_a$  corresponding to modality  $a$ , while  $q \in \mathcal{Q}$  represents the abstract query, often presented in the form of a question pertinent to the task at hand. A task set within modality  $a$  is then constituted by combining certain data elements from modality  $a$  with the queries  $q$ , which can be denoted as  $S_{t,a} = \{(d_a^{(1)}, q), (d_a^{(2)}, q), (d_a^{(3)}, q), \dots\}$ . When the queries  $q$  are held constant, and elements  $d_b \in \mathcal{D}_b$  in another modality  $b$  are gathered, we obtain the corresponding task set in another modality, denoted as

$S_{q,b}$ . In essence, the task  $t$  embodies the task-specific queries, encompassing, e.g., activities such as solving equations, translation, question answering, etc. Meanwhile, the data elements  $d_m$  may take the form of equation instances or question descriptions within modality  $m$ , which can involve the modalities of image, text, or speech.

We introduce the concept of a 'converter,' a function  $K_{a,b} : \mathcal{D}_a \mapsto \mathcal{D}_b$  which maps data elements from modality  $a$  to  $b$ . While there exist various methods for converting data between modalities (e.g., from language to vision through taking a picture), we are specifically interested in converters that preserve information necessary for solving a given task with query  $q$ , denoted as  $K_{a,b}^q$ . Information-preserving converters are distinctive, as the correct answer for a given task instance  $(d, q)$  depends solely on the information within  $d$  rather than its modality. Therefore, both  $(d_a, q)$  and  $(K_{a,b}^q(d_a), q)$  are guaranteed to share the same gold label. In this chapter, we assume the existence of  $K^q$  for every  $q \in \mathcal{Q}$ , but finding such a converter is beyond the scope of this paper. Inter-modality conversion may be challenging for certain modalities. Some tasks may involve aspects of information, such as emotions in speech or nuanced visual perception in images, that cannot be easily preserved during conversion. We design our experiments with tasks where a  $K^q$  clearly exists.

A multimodal model can be conceptualized as a function, denoted  $M : \mathcal{D} \times \mathcal{Q} \mapsto \mathcal{Y}$ , mapping data elements and queries to an answer. Here,  $\mathcal{D}$  represents the collective space encompassing all the modalities of interest, formally  $\mathcal{D} = \bigcup_m \mathcal{D}_m$ , where  $m$  spans over all relevant modalities. On the other hand, the answer space  $\mathcal{Y}$  refers to a unified and structured representation, which, in the case of GPT-4V, assumes the form of text.

A model  $M$  is said to exhibit consistency between modalities  $a$  and  $b$  provided:

$$M(d_a, q) = M(K_{a,b}^q(d_a), q), \forall d_a \in \mathcal{D}_a, q \in \mathcal{Q}$$

In other words,  $M$  is *consistent* if its output is invariant under any modality transformation  $K^q$  which preserves all essential information necessary for solving the task

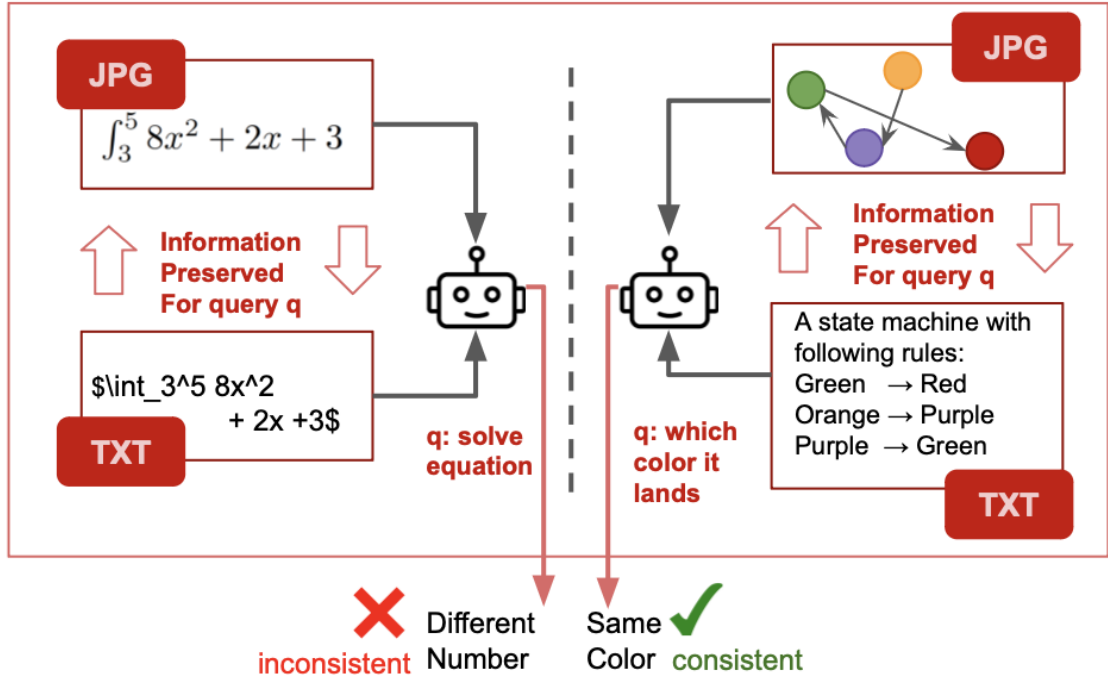


Figure 3.2: Illustration of the concept of cross-modal consistency. A consistent model (right) applies the same internal reasoning to task instances with identical information, regardless of the encoding modality, leading to consistent outcomes. In contrast, an inconsistent model displays significant behavioral changes in response to different input modalities, resulting in varying outcomes as the modality alters.

associated with query  $q$ . E.g., consider solving mathematical equations. A model which solves this task is consistent across the text and image modalities if neither transcribing the equation from image to text, nor imaging an equation presented as text, changes the model’s output.

In short, a consistent model should remain agnostic to the modality of the task instance and yield identical results as long as an equivalent amount of information is provided, reflecting its capacity to handle multimodal data seamlessly.

### 3.3 Method

In this section, we describe our method for testing cross-modal consistency. We establish a quantitative evaluation framework, with a focus on the vision-language cross-modality. We provide a description of our methodology and the specific metrics

we propose for evaluation.

Our second assumption is that the multimodal model  $M$  is a **deterministic** function. Formally, for any given task instance  $(d, q)$ , there exists an output  $y$  within the output space of  $M$  such that  $M(d, q) = y$ . This assumption plays a pivotal role in ensuring cross-modality consistency, as it delineates decoding consistency from modality consistency during result evaluation. In practical terms, we can achieve determinism by setting the decoding temperature to 0 or approximate it by repeating the experimental process and averaging the results.

### 3.3.1 WorkFlow

For an instance set of a given task  $t$  in modality  $a$ , denoted as  $S_{t,a} = \{(d_a^{(1)}, q), (d_a^{(2)}, q), (d_a^{(3)}, q), \dots\}$ , our first step involves constructing a parallel instance set  $S_{t,b}$  in modality  $b$  using an information-preserving converter  $K_{a,b}^q$ . We do so by applying  $K_{a,b}^q$  to each data object  $d_a^{(i)}$  to get the object  $d_b^{(i)} := K_{a,b}^q(d_a^{(i)})$  in modality  $b$ . By doing so, each paired instance  $(d_a^{(i)}, q)$  and  $(d_b^{(i)}, q)$  shares the same gold label since the information in  $d$  is preserved for the task with query  $q$ . In the context of analyzing the vision and language modalities, our converter is comprised of an optical character recognition (OCR) system combined with human verification for converting images to text, and screenshot software for converting text into images. We carefully select tasks where the information required for solving the task can be fully retained through the utilization of this converter, as exemplified by mathematical equation solving.

Next, we independently apply the model  $M$  to each pair of instances  $(d_a^{(i)}, q)$  and  $(d_b^{(i)}, q)$  to obtain pairwise results  $M(d_a^{(i)}, q)$  and  $M(d_b^{(i)}, q)$ .

### 3.3.2 Metrics

We introduce our task consistency score  $C_t$  based on these pairwise instances:

$$C_t = \frac{1}{n} \sum_{i=1}^n c_M^i \quad (3.1)$$

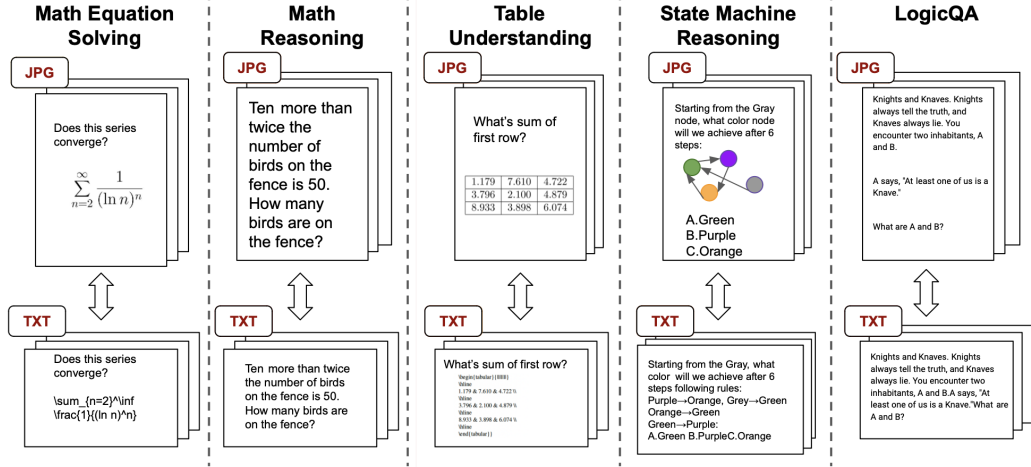


Figure 3.3: An Overview of the Components of Our Vision-Language Consistency Dataset. Data instances are presented in pairs, featuring one in the vision modality and another in the text modality. Notably, Math Equation Solving dataset encompasses two segments, each representing different difficulty levels.

where

$$c_M^i = \begin{cases} 1, & \text{if } M(d_a^{(i)}, q) = M(d_b^{(i)}, q) \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

In essence,  $C_t$  is the proportion of instances for which model  $M$  has consistent performance on the given task, between modalities  $a$  and  $b$ .

## 3.4 Experiments

### 3.4.1 Data Construction

Since there is currently no existing parallel vision-language task dataset, we create our own datasets for both our experiments and also to facilitate future research endeavors. Following the approach outlined in Section 3.3.1, we meticulously selected seven tasks that gauge various facets of Vision-Large Language models. For each of these tasks, we ensure that data instances can be transformed between image and text formats while preserving all task-related information, utilizing a straightforward converter (e.g., OCR). Recognizing that a flawless converter does not exist in practice, we undertake the manual verification of each converted data instance to prevent any

potential errors during the conversion process. We will make our dataset available for use by the research community in the final version of our paper.

### **Task Description.**

**Math Equation Solving.** Mathematical reasoning stands as a cornerstone of multi-modal models’ capabilities. Mathematical problems typically involve equations presented in a visual format, offering a clear depiction of intricate symbols and notations. Given that formulas can be seamlessly converted to text formats like LaTeX without losing any essential information for solving these equations, constructing a parallel dataset for such tasks is a natural fit for analyzing cross-modal consistency. For our dataset, we source math questions with equations from two distinct origins, each representing varying levels of difficulty. For low difficulty levels, we extract 901 high school-level mathematical questions in LaTeX (text) format from MATH dataset (Hendrycks et al. 2021b), rendering each question using a LaTeX compiler to generate corresponding image data. To introduce a greater level of complexity, we gathered 50 college-level calculus questions, along with their corresponding answers, using the same procedure. Consequently, we paired all the image-based math questions with their corresponding text representations to create our comprehensive equation-solving dataset, encompassing both easy and challenging questions. An illustrative example of this dataset can be found in Figure 3.3, and detailed data samples are available in Appendix B.1 and Appendix B.2 .

**Logical Reasoning.** To assess the vision-language consistency in logical reasoning abilities for the VLLMs, we employ two distinct datasets: GSM8K (Cobbe et al. 2021) and LogicQA (J. Liu et al. 2020). GSM8K comprises 8,500 question instances in text format, with each instance representing a problem description in English text paired with a labeled answer. We transform the text into images by capturing screenshots of the rendered text with an appropriate font size and layout. Similarly, LogicQA consists of 8,678 more challenging questions presented in text format and each is con-

verted into an image by us. Subsequently, we pair these resulting images with the original text files, creating a parallel dataset that enables the exploration of this task in both image and text modalities. An illustrative example of this dataset construction can be found in Figure 3.3, and detailed data samples are available in Appendix B.6 and Appendix B.3.

**Table Understanding.** Tables, commonly encountered in everyday life, are often presented as images, and the effective extraction of information from them is vital for various tasks. As well-structured table images can be easily converted into LaTeX text, they serve as an excellent choice for conducting vision-language consistency analysis. To facilitate this analysis, we create 30 distinct tables in LaTeX, each featuring multiple rows and columns, with numerical values in each cell. Our task revolves around accurately summing the numbers within a given row and column. We provide parallel task instances in both LaTeX text and rendered images, as illustrated in Appendix B.5

**State Machine Reasoning.** State machines, which can be effectively visualized as graphs or represented through text with transition rules, serve as an ideal test bed for vision-language consistency in simple computational capabilities of VLLMs. Our approach involves generating images of state machines with varying total numbers of nodes (states). Each node in the state machine is assigned a distinct color and features precisely one outgoing edge, ensuring a unique path and solution. The questions we formulate are of the form, "Starting from the color grey, after  $n$  steps, which color will we end up in?" Here,  $n$  is a variable that we select. Additionally, we generate a text version of these state machines by listing out all the transition rules corresponding to the arrows. To prevent any form of cheating by looking at the last state in the text, we shuffle the order of the rules. We create state machines with different numbers of states and questions with varying numbers of steps, to introduce varying difficulty levels. The data samples can be seen in Appendix B.7.

**Reading Comprehension.** To assess the model’s consistency in comprehending

lengthy English paragraphs across vision and language modalities, we provide the model with the same text content in two different formats: plain text and images of the text. We employ the test part of the Massive Multitask Language Understanding (Hendrycks et al. 2021a), or MMLU dataset as our source, which includes 1,477 extensive text passages, each accompanied by multiple-choice questions designed to evaluate the comprehension of the text content. For this dataset, we convert each text instance into an image by rendering the text into a PDF before converting it to a JPG image. Detailed data samples can be found in the Appendix B.4.

### **3.4.2 Experiment Details.**

We apply our framework and constructed datasets to evaluate the cross-modal consistency of the OpenAI GPT-4V model, known for its proficiency in both vision and language modalities. Given the limited daily access to prompt this model, our experiments were conducted on a randomly selected subset of 50 samples from each dataset. We select the GPT-4V classical mode, which does not include additional plug-ins and employs a relatively low decoding temperature to minimize variance in its output. To ensure a fair comparison of capabilities between the two modalities, we embedded the query questions into the image and exclusively used images for prompting. This avoids the involvement of any text input when testing the vision modality. Additionally, to prevent the model from performing reasoning steps in text and introducing unintended modality conversions, we explicitly instructed the model to output answers without any reasoning steps. Our results are manually collected for pairwise data instances, and we calculate the consistency scores based on the methodology outlined in Section 3.3.1.

| Task          | Modal | Acc    | Consistency |
|---------------|-------|--------|-------------|
| MES(Easy)     | Text  | 0.44   | 0.72        |
|               | Image | 0.24 ↓ |             |
| MES (Hard)    | Text  | 0.62   | 0.62        |
|               | Image | 0.28 ↓ |             |
| LogicQA       | Text  | 0.64   | 0.64        |
|               | Image | 0.44 ↓ |             |
| MMLU          | Text  | 1.00   | 0.74        |
|               | Image | 0.74 ↓ |             |
| TU            | Text  | 0.93   | 0.10        |
|               | Image | 0.03 ↓ |             |
| MR            | Text  | 0.40   | 0.92        |
|               | Image | 0.36   |             |
| State Machine | Text  | 0.34   | 0.67        |
|               | Image | 0.28   |             |

Table 3.1: Test results for vision-language consistency datasets. MES stands for Math Equation Solving, TU stands for Table Understanding and MR stands for math reasoning. The symbol ↓ denotes a sizeable decrease in accuracy (greater than 10%) when input is in the image format.

### 3.4.3 Main Results

The main outcomes of our assessments across seven distinct datasets are outlined in Table 3.1. Notably, even though the input contains an equivalent amount of information necessary for task completion, substantial disparities emerge between image and text input formats. This phenomenon occurs even in tasks where images are conventionally considered to offer a more vivid and intuitive representation from a human perspective.

We note that consistency, being based on response agreement between modalities, can be high or low regardless of per modality accuracy. The highest consistency (0.92) is observed for math reasoning even though both modalities have a relatively low accuracy ( $\leq 0.40$ ). By contrast, the consistency drops to 0.64 on logical reasoning (LogicQA) on which the individual modalities have higher accuracy ( $\geq 0.44$ ).

For tasks that involve intricate reasoning steps, including equation solving, math-/logical reasoning, and state machine reasoning, we observe relatively low accuracy even when the input is presented in pure text format. These tasks align with areas where the model generally struggles. When the input modality shifts to using images, the proficiency in solving such tasks deteriorates further, resulting in a noticeable drop in performance, despite the fact that the images contain an equal amount of information. This emphasizes the substantial inconsistency in task-solving across modalities and highlights the model’s superior ability in one modality (Language) compared to the other (Vision).

On the other hand, for tasks primarily focused on extracting information from provided content and comprehending that information, such as Language Understanding and Table Understanding, we witness near-perfect performance when the model is prompted with text input. However, a more significant drop in accuracy (up to 90%) is observed in such tasks when the input modality shifts to images. This indicates that the change in modality significantly impacts the model’s processing capabilities,

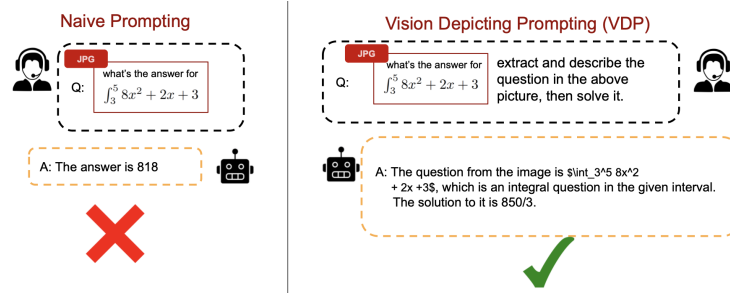


Figure 3.4: Overview of the VDP Method: The left part illustrates the conventional approach to prompting vision tasks, while the right part demonstrates VDP in comparison.

providing strong evidence of the inconsistency of the model.

In conclusion, in multimodal systems like GPT-4V, the language modality demonstrates a **dominant** advantage over vision modality, when tasks are tackled in text format, despite the presence of the same information in image format. This strongly suggests a *non-consistent* cross-modal behavior within the network. While each modality exhibits varying levels of task-solving and reasoning capabilities, the inconsistency across modalities is observed across tasks regardless of the accuracy level of each modality for the task in hand.

### 3.4.4 Ablation Study on Content Extraction from Images

As solving tasks in image format inevitably requires accessing essential information from the images, we conducted additional experiments to investigate whether the performance gap is attributable to the model’s inability to access information. To address this, we conducted one-step Optical Character Recognition (OCR) using the model’s own network on all instances of tasks that exhibited a significant performance gap between image and text. Specifically, for each image input (indicated by the red arrow in Table 3.1), we prompt the model with the instruction ‘extract the exact content in the image’ and compare the results with the original input to determine if they match. This approach allows us to eliminate the possibility that the performance issues in image format are due to the model’s inability to correctly recognize the input.

| <b>DataSet</b> | <b>OCR Accuracy</b> |
|----------------|---------------------|
| MES (Easy)     | 0.68                |
| MES (Hard)     | 0.76                |
| LogicQA        | 0.98                |
| MMLU           | 0.98                |
| TU             | 1.00                |

Table 3.2: Result of performing OCR on the all images of experimented task instances.

As shown in Table 3.2, OCR accuracy approaches nearly 100% for all instances of LogicQA, MMLU, and Table Understanding tasks. This suggests that the model faces no difficulties in accurately extracting information, such as numbers from each row and column in table images. The substantial gap (up to 90%) in accuracy (Table 3.1) between images and text can be attributed solely to the model’s internal reasoning processes for each modality. This underscores the inconsistent internal reasoning employed by the model when presented with the same content in different modalities.

In contrast, we observe lower OCR accuracy for Math equation-solving inputs, as complex math equations pose challenges for accurate recognition and extraction. To isolate and distinguish the source of inconsistency – inaccurate recognition of image data *or* poor actual internal reasoning, we report *conditional consistency scores* for image instances given correct versus incorrect OCR results. From Table 3.3, it becomes evident that there is no direct correlation between consistency scores and direct OCR accuracy. This further bolsters our claim that such models simply exhibit distinct (and inconsistent!) internal behaviors under different modalities.

### 3.5 Vision-Depicting-Prompting (VDP)

As shown in Section 3.4.3, for the same task, VLLMs such as GPT-4V perform much better when questions are presented in text format, even when the information can

| <b>DataSet</b> | <b>YConsistency</b> | <b>NConsistency</b> |
|----------------|---------------------|---------------------|
| MES (Easy)     | 0.70                | 0.75                |
| MES (Hard)     | 0.66                | 0.58                |

Table 3.3: Conditional vision-language consistency score given the OCR results. The term ‘YConsistency’ refers to the consistency given OCR outputs are correct. Conversely, ‘NConsistency’ denotes the consistency score given incorrect OCR outputs.

be completely extracted from the image instances. Inspired by these findings, we propose a novel method of *Vision-depicting-prompting* (VDP) for improving model’s reasoning ability through image context. We now explain VDP.

### 3.5.1 Prompting Details

In the case of a task instance presented in image format, VDP diverges from directly soliciting an answer solely based on the image input, as illustrated in Figure 3.4. Instead, we adopt a two-step process: we first prompt the model to extract and articulate the description of the image task using textual language. The prompt we used in our experiments was: “extract and describe the question in the above picture”. This aims to maximize the transformation of the image signal into a text signal, recognizing the inherently stronger reasoning abilities associated with text information, as demonstrated earlier. Subsequently, we prompt the model to provide an answer, taking into account both the text description of the task and the original image input, as depicted in Figure 3.4. The prompt we used in our experiments was: “then solve it.” The two example prompts were concatenated into a single prompt and then fed to VLLMs with the corresponding image during the experiments.

Unlike previous research that sought to enhance the reasoning abilities of multi-modal models by augmenting input images with supplementary text (Lin et al. 2022; Hu et al. 2023), VDP does not focus on information augmentation. Particularly in the task instances designed for our study, images already contain all the necessary information required to complete the task. Therefore, converting these images into

text format does not provide any additional information that aids in solving the task. Instead, VDP is rooted in the observation that textual signals can significantly stimulate a model’s reasoning capability as model has a bias towards language modality. Instead, VDP is based on the observation that textual signals can significantly stimulate a model’s reasoning capability, given the model’s inherent bias toward the language modality. VDP achieves this by explicitly extracting textual information from the images, thus directly leveraging the model’s language processing capabilities more effectively.

### 3.5.2 Experiment Results for VDP

We apply VDP to five of the tasks previously examined in Section 3.4, where these tasks demonstrate notable performance disparities between image and text inputs. We therefore investigate whether VDP can effectively bridge the performance gap between modalities on such tasks. The outcomes are detailed in Table 3.4.

Remarkably, we observe a substantial improvement in accuracy exceeding 12% when solving problems within the realm of vision modalities using VDP, as compared to naive prompting. In tasks requiring reasoning abilities, we note an average accuracy enhancement of 19%. However, the overall performance still lags behind that of text-based prompting. This discrepancy can likely be attributed to the challenges in accurately depicting and extracting information from objects within images during VDP. In contrast, an impressive average increase of 57% in accuracy is observed in tasks centered around understanding (TU and MMLU). Particularly, in the case of table understanding, we witness a remarkable 90% boost in accuracy, particularly when the table’s content is extracted before any necessary calculations are applied. For these tasks, we find that performance eventually reaches parity with text-based prompting, underscoring the effectiveness of VDP, particularly in tasks that involve a deeper understanding of the information within the input instances.

Furthermore, there is a substantial increase in the consistency score with VDP

| Task       | Modality | Prompt | Acc           | Consistency   |
|------------|----------|--------|---------------|---------------|
| MES (Easy) | text     | naive  | 0.44          | —             |
|            | image    | naive  | 0.24          | 0.72          |
|            |          | VDP    | <b>0.48</b> ↑ | 0.72          |
| MES (Hard) | text     | naive  | 0.62          | —             |
|            | image    | naive  | 0.28          | 0.62          |
|            |          | VDP    | <b>0.50</b> ↑ | <b>0.76</b> ↑ |
| LogicQA    | text     | naive  | 0.64          | —             |
|            | image    | naive  | 0.44          | 0.64          |
|            |          | VDP    | <b>0.56</b> ↑ | <b>0.80</b> ↑ |
| MMLU       | text     | naive  | 1.00          | —             |
|            | image    | naive  | 0.74          | 0.74          |
|            |          | VDP    | <b>0.98</b> ↑ | <b>0.98</b> ↑ |
| TU         | text     | naive  | 0.93          | —             |
|            | image    | naive  | 0.03          | 0.10          |
|            |          | VDP    | <b>0.93</b> ↑ | <b>0.90</b> ↑ |

Table 3.4: Result of VDP prompting. MES stands for Math Equation Solving and TU stands for Table Understanding. ↑ represents an improvement of more than 10% using VDP.

compared to prompting with plain images (naive prompting), e.g., from 0.64 to 0.80 on LogicQA and from 0.10 to 0.90 on TU. These results reinforce our hypothesis that models such as GPT-4V exhibit varied and often inconsistent reasoning capabilities across different modalities and underscore the effectiveness of our VDP approach for enhancing consistency. Properly addressing such disparities between modalities as done by our VDP approach can also help to improve the performance in solving the tasks.

### **3.6 Threats to Validity**

In this study, due to cost concerns, data samples in each dataset in the modality of image were standardized with regard to a single setting, such as font, font size, table borders, image size, and other parameters. Testing data instances with different settings would be reasonable work in the future. At the time of conducting the experiments up to now, the OpenAI website for GPT-4V has daily and hourly limits on the number of prompts. Consequently, we did not test our experiments on the entire dataset.

### **3.7 Conclusion**

In this study, we performed a systematic analysis of the consistency across modalities in multimodal systems. Our results demonstrate that models such as GPT-4V maintain a relatively independent internal representation of reasoning between visual and textual signals, as evidenced by results we obtained on our datasets which we specially designed for the tasks. Notably, GPT-4V exhibits superior performance in language modelling compared to reasoning within a visual context. These findings offer valuable insights into the potential applications of such multimodal systems and highlight the need for more integrated system designs. Furthermore, we introduce a Vision-depicting-Prompting solution to effectively address this inconsistency.

## Chapter 4

# Lexicalization Generation and Lexical Gap Detection

In computational semantics, the term *lexicon* refers to the vocabulary which speakers of the language use to express concepts. A language  $L$  *lexicalizes* a concept  $s$  if it has a lexeme (or a word) that can express  $s$ ; otherwise  $s$  corresponds to a *lexical gap* in  $L$  (Murphy et al. 2010). For example, the Polish word *pojutrze* is a *lexicalization* of the concept “the day after tomorrow”, which corresponds to a lexical gap in English. In particular, kinship terms describe familial relations such as “grandparent” and “female cousin”. The clear definitions (*glosses*) and regular hierarchical structure of kinship concepts make them well-suited for investigations into lexicons and lexical gaps.

The task of identifying lexicalizations for a given concept underlies automatic construction of *multilingual wordnets* Martelli et al. 2023, lexical knowledge bases modeled after the Princeton WordNet Miller et al. 1990. Wordnets are comprised of synonym sets (*synsets*), each of which corresponds to a single concept, and contains the set of words which can express that concept Bond et al. 2013. Wordnets constructed by expanding the synsets of the Princeton WordNet may misrepresent terms and concepts that correspond to English lexical gaps Kwong 2018. For example, in BabelNet 5.2, the Spanish words *prima* “female cousin” and *primo* “male cousin”, which are clearly *not* synonymous, are both included in the same synset as English

---

**Zh:** 我有一个 堂哥 , 但是我没 堂姐 。

---

**En:** I have a cousin , but I have no cousin .

---

Table 4.1: An incorrect translation from Chinese to English from Google Translate. 堂哥 and 堂姐 mean “elder son of father’s brother” and “elder daughter of father’s brother”, respectively.

*cousin*.<sup>1</sup> Translation models often fail to correctly translate words which correspond lexical gaps in the target language, such as in the example in Table 4.1.

Prior NLP work on semi-automatic identification of concept lexicalizations and lexical gaps has leveraged bilingual dictionaries, wordnets and linguistic typologies. Bentivogli et al. (2000) apply a decision tree approach based on information from a machine-readable bilingual dictionary, but their experiments are restricted to English and Italian. Gregori et al. (2017) focus on a subset of bilingual action verbs in the context of video-based translation, but establish no mapping to wordnet-type concepts. Khishigsuren et al. (2022) compile a dataset of 1911 kinship terms and a list of lexical gaps in 176 languages by combining native speaker expertise in 10 languages, lexicalization information from Wiktionary, and a set of typological patterns of Murdock (1970). We leverage their resource to develop a method which is fully automated, language-independent, and not specific to kinship terms.

The principal idea behind our approach is to reduce the task of lexical gap detection to the task of lexicalization generation. The latter can be viewed as populating a concept-language matrix (Figure 4.1) in which each cell contains a lexicalization of the concept in the corresponding language (if one exists). Given a correctly populated lexicalization matrix, empty cells correspond to lexical gaps. The task of lexicalization generation is similar to populating wordnet synsets, but requires returning a single lexicalization, rather than all concept lexicalizations.

---

<sup>1</sup>babelnet.org, version 5.3, synset ID bn:00023333n

We propose a translation-based method for lexicalization generation, which we also apply to lexical gap detection. Our method is grounded in theoretical propositions based on the hypernym/hyponym relationships between concepts, which provide a basis for detecting incorrect lexicalizations. In particular, our model predicts that languages tend to avoid ambiguity between lexicalizations within disjunctive triples of concepts such as *parent/mother/father*. We develop a method for lexical gap detection via filtering concept lexicalizations produced by translating unambiguous “seed” words in the context of the corresponding concept glosses. We leverage existing lexical knowledge bases and machine translation systems, as well as the tree structure of kinship concepts, to decide if a given translation is literal or indicative of a lexical gap. While we focus on kinship terms in this work, our approach is also applicable to other sets of concepts.

Our evaluation on kinship terms across 10 diverse languages demonstrates substantial improvements over BabelNet and ChatGPT. ChatGPT exhibits a tendency to generate overly specific or irrelevant responses, while BabelNet fails to represent many of the concepts that are not lexicalized in English. We identify three main causes of errors made by our algorithm: inaccurate translations, non-standard terms, as well as errors and omissions in the benchmark dataset itself. We release our code and BabelNet concept mapping on GitHub.

## 4.1 Theoretical Framework

We start this section by discussing the linguistic background related to the issue of lexical gaps. We then formally define the tasks addressed in this paper, as well as related theoretical concepts, such as literal translations, seed words, and disjunctive concept triples. This is followed by propositions and proofs that form the theoretical basis of our method.

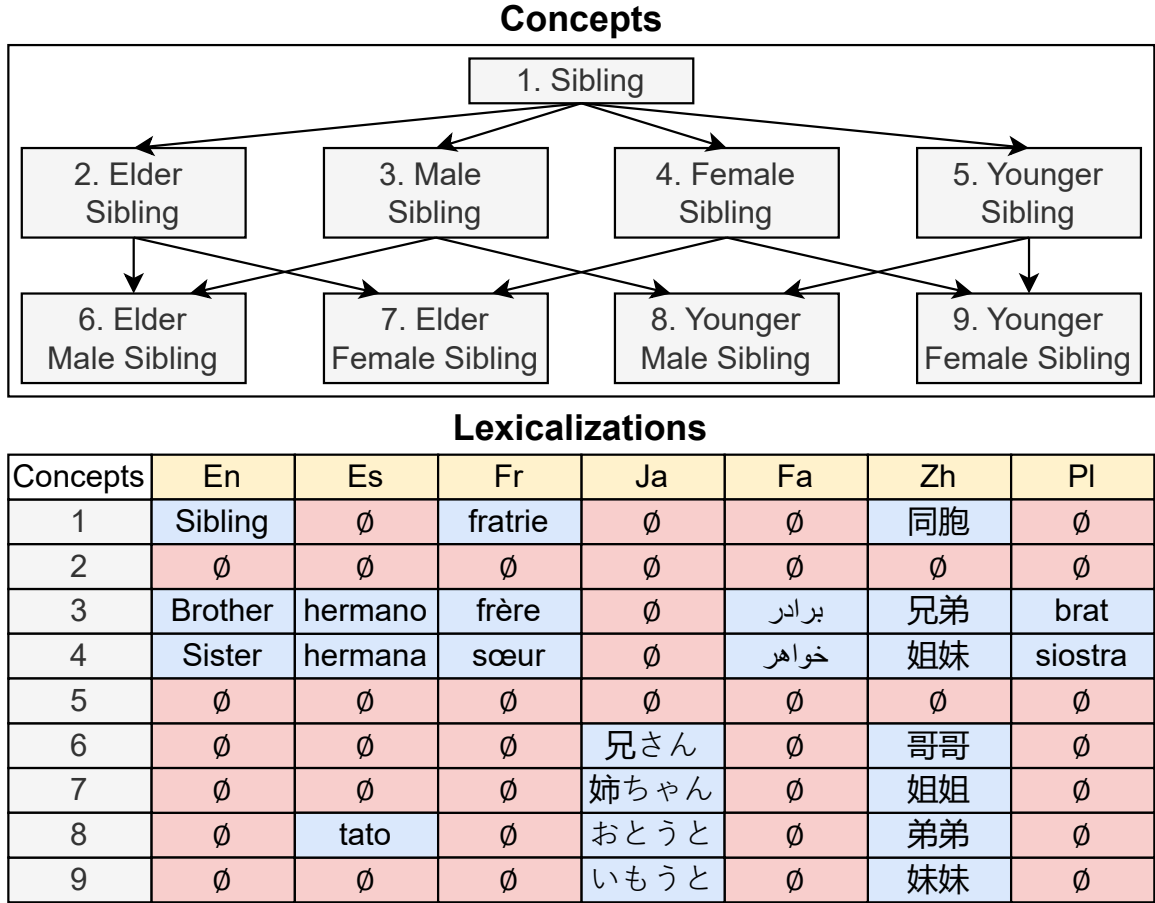


Figure 4.1: An illustration of a hierarchical structure of kinship concepts (top), and a concept-language matrix of lexicalizations (bottom).

#### 4.1.1 Linguistic Background

Chomsky et al. (1965) introduce the distinction between *accidental gaps* (words that could theoretically exist) and *systematic gaps* (words that would contravene phonological constraints). Lehrer (1974) discusses several types of gaps: phonological, morphological, syntactic, paradigmatical, derivational, functional, and semantic. Ivir (1977) questions the utility of systemic gaps, and focuses instead on lexical gaps (concepts that are specific to individual cultures) and conceptual gaps (missing lexicalizations of “universal” concepts). The latter type, which includes kinship terms, is considered more important, being an inter-language rather than intra-language phenomenon.

In the context of translation, Cvilikaite (2006) defines lexical gaps as instances of

lack of lexicalization for a given concept, and emphasizes the difficulty of identifying them prior to translation. Janssen (2004) observes that lexical gaps correspond to words for which there is no single-word translation in a target language. For example, the concept expressed by the Russian word *goluboj* is “light blue”, which is considered a lexical gap in English, even though it can be approximately translated with a single word *blue*. According to Gouws (2002) a translation dictionary entry for a lexical gap needs to include a “brief paraphrase of meaning” (*gloss*) and/or a “loan word” (source language term); e.g., “*bobotie, South African curried mincemeat.*”

### 4.1.2 Definitions

A *wordnet* is a semantic knowledge base composed of synonym sets, or *synsets*. Each synset corresponds to a unique concept, and to a different sense of each word that it contains. Each synset is associated with a part of speech, and a gloss that defines the meaning of the concept. Each word in a synset can express (i.e., *lexicalizes*) the corresponding concept.

Hauer et al. (2023a) define a theoretical binary problem  $\text{Sense}(w, s)$  for deciding whether the word  $w$  can express the concept  $s$ . A word lexicalizes a concept if it can express the meaning conveyed by the concept’s gloss. For example, unlike the English compositional phrase *female cousin* or the Spanish word *prima*, the English word *cousin* on its own cannot express the concept of “female cousin”, which is defined as “the daughter of your aunt or uncle”. A method that solves the Sense problem could theoretically be used to populate any wordnet synset, by testing each word in the lexicon on whether it can express the concept corresponding to that synset.

We define the task of *lexicalization generation* (LexGen) as follows: given a language  $L$  and a concept  $s$ , a method must return either a word in  $L$  which lexicalizes  $s$ , or a special GAP token indicating that no such word exists. For example, the word *prima* is a possible return value of  $\text{LexGen}(\text{SPA}, \text{“female cousin”})$ . The LexGen task is reducible to the Sense problem by returning any word in  $L$  for which  $\text{Sense}(w, s)$

is TRUE, or GAP if no such word exists.

We define the binary task of *lexical gap detection* (LexGap) as follows: given a language  $L$  and a concept  $s$ ,  $\text{LexGap}(L, s)$  returns TRUE if  $L$  has no word that lexicalizes  $s$ , or FALSE otherwise. For example,  $\text{LexGap}(\text{ENG}, \text{“female cousin”})$  returns TRUE, as there is no word in the English lexicon to express the concept. LexGap is reducible to LexGen in a straightforward manner by returning TRUE if and only if LexGen returns GAP. LexGap can also be reduced directly to Sense:

$$\text{LexGap}(L, s) \Leftrightarrow \forall w \in L : \neg \text{Sense}(w, s)$$

A *literal translation* is an expression in the target language that preserves the meaning of the expression in the source language in a given context. In the case of a *literal lexical translation*, the target word expresses the same concept as the source word. We assume that a translator, which can be either a human or a machine, is guided by the following priorities: (1) fidelity (meaning preservation), (2) brevity (conciseness), and (3) fluency Hauer et al. 2023b. Therefore, a translator prefers literal to non-literal translations, as well as single-word translations to multi-word phrases. In the case of a lexical gap, a literal lexical translation is not an option. Both non-literal and phrase translations can therefore be considered indications of lexical gaps in the target language. For example, Spanish *prima* can be translated into *female cousin* (phrasal translation) or just *cousin* (non-literal translation). A heuristic for detecting non-literal lexical translations is the *back-translation test*: a source word  $w$  in context  $C$  is first translated into a target word  $w'$ , which is then translated back in the same context into a source language word  $w''$ ; the test succeeds if and only if  $w'' = w$ . For example, *cousin* as a translation of *prima* may fail the back-translation test.<sup>2</sup>

We introduce a notion of *seed words*, defined as words that lexicalize exactly one concept within a set of concepts. For example, the Spanish word *prima* is considered

---

<sup>2</sup>Google translates *Amo a mi prima* into *I love my cousin* and then back into *Amo a mi primo* (accessed April 18, 2024).

a seed word for the concept of “female cousin” within the set of kinship terms. We use seed words in Section 4.3 as unambiguous source words to generate target concept lexicalizations via translation.

### 4.1.3 Disjunctive Triples

Simple natural language statements can often be mapped to symbolic logic, and vice versa, with the logical operators represented by conjunctions such as *and*, *or*, and *not*. In particular, an apparent colloquial or textual contradiction can often be expressed as a logical proposition that is false for all values of its variables. For example, “*Robin is brave and not brave*” intuitively corresponds to  $\text{brave}(x) \wedge \neg \text{brave}(x)$ , where the variable  $x$  represents *Robin*. We refer to such natural language expressions as *colloquial contradictions*.

The typological phenomena used to construct the hypernymy graph of kinship terms are binary. Therefore, kinship concepts can often be arranged into triples, wherein a concept  $s_0$  is an *exclusive disjunction* of its hyponym concepts  $s_1$  and  $s_2$ . Among the kinship terms, the principal type of exclusive disjunction is *gender*; for example, a *sibling* is either a *sister* or *brother*. The gender distinction can be indirect; for example, an *uncle* can be referred to as either *maternal* or *paternal*. Another type of disjunction is *relative age*; for example, a *cousin* can be either younger or older. Other distinctions are possible, such as the speaker’s gender, or consanguinity vs. affinity.

Because hyponymy is the **IS-A** relation, any instance of  $s_0$  must be either an instance of  $s_1$  or  $s_2$  (but not both). If a single word  $w$  could express both  $s_1$  and  $s_2$ , then  $w$  would also necessarily express the hypernym  $s_0$ . To avoid confusion, if a speaker specifically wishes to refer to concept  $s_1$ , as opposed to its hypernym  $s_0$ , it is logical to choose a word (or phrase) which excludes  $s_2$ . For example, since the Spanish word *padre* can lexicalize both concepts of “father” and “parent” (especially in its plural form), speakers may instead use the word *progenitor* to express the latter

concept.

In symbolic logic, an exclusive disjunction is expressed by the XOR (exclusive OR) operator:  $\oplus$ . In plain English, an exclusive disjunction can be expressed as “*either ... or ...*”; if a concept  $s_0$  is an exclusive disjunction of its hyponyms  $s_1$  and  $s_2$ , the phrase that combines the glosses of the hyponyms as “*either  $C_1$  or  $C_2$* ” is a possible gloss for  $s_0$ . For example, since “parent” is the exclusive disjunction of its hyponyms “father” and “mother”, it can be defined as “father or mother.”

#### 4.1.4 Propositions

In the remainder of this section, we present two propositions formulated on exclusive disjunctive triples, which form the basis of our methods in Section 4.3 for removing spurious lexicalizations.

**Proposition 1** *If a concept  $s_0$  is an exclusive disjunction of its hyponym concepts  $s_1$  and  $s_2$ , then expressing both  $s_0$  and  $s_1$  with the same word can result in a colloquial contradiction.*

**Proof.** Suppose that there exists a word  $w$  that lexicalizes both concept  $s_0$  and its hyponym  $s_1$ . Since  $s_1$  and  $s_2$  are disjunctive hyponyms of  $s_0$ , the meaning of  $s_2$  could be expressed by a phrase “ $w$  but not  $w$ ”, in which  $w$  is used in two different senses of  $s_0$  and  $s_1$ . This phrase intuitively corresponds to a logical contradiction:  $w(x) \wedge \neg w(x)$ . ■

Intuitively, the use of the same word to lexicalize both members of a hypernym/hyponym pair can lead to highly ambiguous expressions, which is undesirable in any natural language. For example, since Spanish *padre* can mean both “parent” and “father”, Google Translate translates the English sentence “*Robin is my parent but not my father*” into “*Robin es mi padre pero no mi padre.*” Contextual disambiguation of such apparently contradictory statements is particularly difficult if the two concepts are closely related by hyponymy.

**Proposition 2** *If a concept  $s_0$  is an exclusive disjunction of its hyponym concepts  $s_1$  and  $s_2$ , then expressing both  $s_1$  and  $s_2$  with the same word can result in a colloquial contradiction.*

**Proof.** Suppose that there exists a word  $w$  that lexicalizes both  $s_1$  and  $s_2$ . Since  $s_1$  and  $s_2$  are disjunctive hyponyms of  $s_0$ , the meaning of  $s_0$  could be expressed by a phrase “either  $w$  or  $w$ ”, in which  $w$  is used in two different senses of  $s_1$  and  $s_2$ . This phrase intuitively corresponds to a logical contradiction:  $w(x) \oplus w(x)$ . ■

For example, the concepts of “female cousin” and “male cousin” which are lexicalized in Spanish by *prima* and *primo*, respectively, correspond to lexical gaps in English. Given the Spanish sentence “*Tengo una prima pero no tengo ningún primo.*”, Google Translate<sup>3</sup> produces a translation which is at best ambiguous, at worst non-sensical: “*I have a cousin but I have no cousin*”.

Taken together, Propositions 1 and 2 yield the following corollary, which applies to disjunctive triples of concepts, based on the intuition that colloquial contradictions should be rare.

**Corollary 1** *If a concept  $s_0$  is an exclusive disjunction of its hyponyms  $s_1$  and  $s_2$  then all their lexicalizations should be different.*

Figure 4.2 shows 10 possible types of exclusive disjunctive triples, of which 4 types (numbered 7-10) are excluded by Corollary 1 because two or more concepts are lexicalized by the same word. Types 7, 8, and 10 fall under Proposition 1, while types 9 and 10 fall under Proposition 2.

## 4.2 Taxonomy

The kinship taxonomy is composed of six directed acyclic graphs (DAGs), in which nodes correspond to concepts, and edges represent the IS-A relationship between

---

<sup>3</sup>translate.google.com, February 15, 2024.

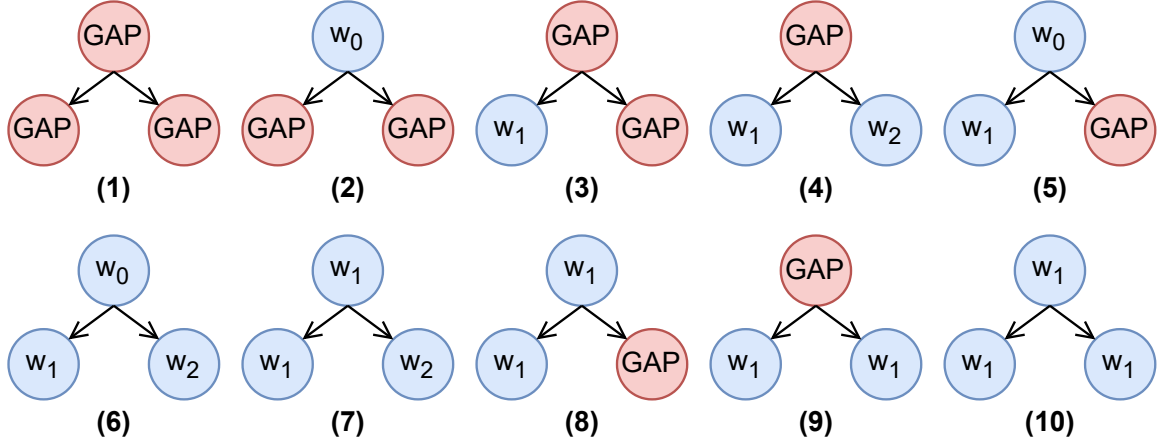


Figure 4.2: Types of concept triples. Distinct lexicalization are represented by different variables.

| Root        | Compound Gloss             | Size |
|-------------|----------------------------|------|
| sibling     | sibling                    | 9    |
| grandchild  | child of child             | 9    |
| grandparent | parent of parent           | 9    |
| auncle      | sibling of parent          | 27   |
| nibling     | child of sibling           | 27   |
| cousin      | child of sibling of parent | 81   |

Table 4.2: The glosses and the number of generated nodes for each root concept.

hypernyms and hyponyms. Each concept is represented by an ordered list of *atomic* kinship relations (or attributes): child, parent, and sibling. The list of relations is specific to each DAG, as shown in Table 4.2. For example, the list of relations for the root concept “cousin” is [child, sibling, parent] which translates into a compound gloss “child of sibling of parent.”

Each relation in the ordered list can have a value. The three atomic relations admit the gender distinction (i.e. male vs. female). The age distinction (i.e. younger vs. elder) is relative either to the speaker (when referring to siblings or cousins) or to “sibling of parent”. For example, the concept “younger, male child of sibling of female parent” (that is, “son of mother’s sibling, younger than the speaker”) is represented

```

G.create(); Q.create()
sr ← concept(root)
G.addNode(sr); Q.enqueue(sr)
while not Q.isEmpty() do
  s0 = Q.dequeue()
  for each undefined relation in s0 do
    for each possible value of relation do
      s1 ← concept(s0)
      s1.relation ← value
      if s1 ∉ G then
        G.addNode(s1); Q.enqueue(s1)
      G.addEdge(s0, s1)

```

Figure 4.3: The algorithm for generating a concept graph.

by [child = male, sibling = undefined, parent = female, age = younger]. Concepts that have the same representation are considered identical, so there is at most one node in any graph with a given representation.

We generate our kinship taxonomy automatically. The algorithm in Figure 4.3 generates a complete DAG given one of the root concepts from Table 4.2. The graph  $G$  is initialized with the root concept, in which every relation set to *undefined*. The algorithm maintains a queue  $Q$  which contains the nodes to be expanded by setting each of the available relations to either of its possible values. Each iteration of the innermost *foreach* loop results in a creation of a directed edge between the current node  $s_0$  and one of its hyponyms  $s_1$ . If the hyponym node  $s_1$  has not yet been created, it is added to the graph and the queue.

## 4.3 Methodology

In this section, we describe our approach to lexicalization generation. The essence of the method is to generate a candidate lexicalization for each concept by translating a seed word into the target language in the context of the concept gloss, and then apply a series of filters to remove incorrect candidates. For each concept, we output the corresponding lexicalization if it has not been filtered out, or GAP otherwise.

### 4.3.1 Candidate Generation

Given the seed word for a concept, we translate the seed words in the context of the concept gloss using the template “[*seed word*]: [*concept gloss*]” which we refer to as a *gloss context*. Including the gloss yields better results than providing only the seed word, by providing the translation system with additional context. Ideally, the translation of the seed word should be a lexicalization of the input concept.

After translating the gloss context into the target language, we extract the candidate lexicalization by retrieving the part of the translation before the colon. For example, to identify a French lexicalization for the concept of “aunt”, we translate the gloss context “*aunt: a parent’s sister*” into “*tante: la sœur d’un parent.*” We then extract the lexicalization candidate *tante*.<sup>4</sup>

### 4.3.2 Candidate Filtering

Translation errors and lexical gaps may lead to inaccurate, non-literal, or non-lexical translations, which are not appropriate as lexicalizations. We therefore apply a sequence of filters to remove incorrect candidate lexicalizations. The pseudo-code of the algorithm is shown in Figure 4.4.

**Multi-Word Filter (#1)** The multi-word filter rejects any candidates which are composed of multiple word tokens. This effectively enforces a strict definition of a lexicalization as a single orthographic word. We found that multi-word expressions, such as *female cousin* are usually compositional, and therefore not suitable as lexicalizations. Some linguists adopt an even more strict definition that a lexicalization must be a mono-morphemic word Khishigsuren et al. 2022, however, we do not enforce this constraint in our method. Since the Chinese language does not separate words orthographically, we detect multi-word expressions by identifying characters which are indicative of word boundaries: 的 and 或.

---

<sup>4</sup>Example translations in this section were obtained from Google Translate on February 14, 2024.

**Horizontal Filter (#2)** In addition to considering individual concepts separately, we also leverage lexicalization patterns to filter out non-literal translations. If both hyponyms in a disjunctive triple are found to have the same candidate lexicalization, Proposition 2 from Section 4.1.4, implies that this may be a non-literal, hypernym translation, indicating the existence of two lexical gaps. For example, if the Spanish terms *primo* “male cousin” and *prima* “female cousin” are both translated into English as *cousin*, the horizontal filter replaces both instances of *cousin* with GAP indicators.

**Back-Translation Filter (#3)** If the candidate lexicalization can indeed express the same concept as the seed word in the context of its gloss, it should be possible to recover the seed word by *back-translating* the candidate in the context of the translated gloss. The back-translation filter is designed to detect non-literal translations by applying this test, removing any lexicalizations which do not pass. If the original seed word is not recovered, the candidate is discarded, and the output for that concept is a lexical gap. For example, if the Chinese seed word 弟弟 “younger brother” is translated into English as *brother*, and then back-translated into Chinese as 兄弟 “brother” then the filter removes this lexicalization and labels this concept as a lexical gap in English.

**Vertical Filter (#4)** Our final filter is based on Proposition 1 from Section 4.1.4, which implies that a concept and its hyponym within a disjunctive triple are unlikely to share the same lexicalization. If such a case is detected, the vertical filter removes one of the two instances of the lexicalization. Intuitively, we expect languages to be consistent in their lexicalization patterns; for example, if a language has no word for “elder brother” it is less likely to lexicalize “elder sister”. Therefore, we label both hyponyms as gaps if the lexicalization of the co-hyponym in the triple has already been removed by an earlier filter; otherwise, we remove the hypernym candidate word

```

for each concept  $s$  do
   $L_0(s) \leftarrow \text{Translate}(\text{seed}(s), \text{gloss}(s))$ 
for each concept  $s$  do ▷ Filter #1
   $L_1(s) \leftarrow \text{GAP}$  if  $L_0(s)$  is not a word
for each triple  $(s_0, s_1, s_2)$  do ▷ Filter #2
   $L_2(s_1) \leftarrow \text{GAP}$ ;  $L_2(s_2) \leftarrow \text{GAP}$  if  $L_1(s_1) = L_1(s_2)$ 
for each concept  $s$  do ▷ Filter #3
   $L_3(s) \leftarrow \text{GAP}$  if  $\text{BackTrans}(L_2(s), \text{gloss}(s)) \neq \text{seed}(s)$ 
for each triple  $(s_0, s_1, s_2)$  do ▷ Filter #4
  if  $L_3(s_0) = L_3(s_1)$  then
    if  $L_3(s_2) = \text{GAP}$  then  $L_4(s_1) \leftarrow \text{GAP}$  else  $L_4(s_0) \leftarrow \text{GAP}$ 

```

Figure 4.4: Pseudo-code of the algorithm for lexicalization generation and lexical gap detection. The lexicalizations in  $L_i$  are copied to  $L_{i+1}$  after each loop.

instead.

## 4.4 Experiments

This section describes the empirical evaluation of our method. We provide information on our datasets, resources, metrics, and baselines.

### 4.4.1 Data

Our principal dataset is the *Database of Lexical Diversity in Kinship Domain* released by Khishigsuren et al. (2022), henceforth referred to as the kinship database. It compiles data from 699 languages pertaining to 198 kinship concepts, divided into six subdomains: *cousins*, *grandchildren*, *grandparents*, *nephews/nieces*, *siblings*, and *uncles/aunts*. It explicitly lists over 37k lexical gaps, based on various resources and inference methods, as well as 1911 lexicalizations, from 168 of the 699 languages.

For each of the six concept categories, more specific concepts are derived by the application of mutually exclusive distinctions. Each distinction induces two hyponyms of a given concept, which together form an exclusive disjunctive triple (Section 4.1.3). For example, the application of the *relative age* distinction to the concept “sibling” yields the concepts “elder sibling” and “younger sibling”. This property is crucial, as

it admits the application of Propositions 1 and 2 in our method.

The 198 concepts in the kinship database do not include all possible concepts that could be derived by the application of the *gender* and *age* distinctions, because the creators of the kinship database excluded concepts that were not attested in their sources. Furthermore, 74 terms are distinguished only by the gender of the speaker; we do not consider such terms to denote a separate concept.

When applied to the six root concepts in Table 4.2, our algorithm in Figure 4.3 generates DAGs which include all 124 distinct concepts in the kinship database, as well as 38 additional concepts.<sup>5</sup> An example of an additional concept is “child of younger sibling.” It is an open question whether such concepts are lexicalized in any natural language, but gender-independent concepts are expected to be introduced in the future.<sup>6</sup>

#### 4.4.2 Seed Words and Glosses

In order to generate a candidate lexicalization for a given concept, we construct a gloss context (Section 4.3.1) by concatenating a seed word with a concept gloss in the same language. We then extract the candidate lexicalization from the translation of the gloss context obtained with Googletrans API.

We select the seed words from the set of words that lexicalize exactly one concept in the kinship database. If there is more than one such word, we prefer highly-resourced languages, according to their coverage in BabelNet v5.3, which are likely to yield more accurate translations.<sup>7</sup> The majority of the seed words, 53 out of 71, are from the three languages used for the method development: English (eng), Chinese (zho), and Persian (pes).

For most concepts, we use the glosses provided in the kinship database, such as

---

<sup>5</sup>Specifically, 28 cousin, 6 nibling, and 4 auncler concepts. The full list of concepts is available in our mapping resource.

<sup>6</sup>For example, the concept “sibling of parent” (aunt or uncle) is lexicalized in the kinship database only in constructed languages Esperanto, Ido, and Volapük.

<sup>7</sup><https://babelnet.org/statistics>

| LexGap (F1)   |              |             |             |             |             |             |             |             |             |             |             |             |             |             |
|---------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Method        | eng          | zho         | pes         | spa         | rus         | fra         | deu         | pol         | ara         | ita         | mon         | hun         | hin         | Test Avg.   |
| All-Gaps      | 81.6         | 62.4        | 79.5        | 81.7        | 83.8        | 75.0        | 82.2        | 81.9        | 84.4        | 71.4        | <b>92.5</b> | 61.3        | <b>70.5</b> | 78.5        |
| BabelNet      | 98.4         | 59.5        | 76.7        | 85.7        | 91.2        | 87.7        | 93.1        | 80.6        | 85.7        | <b>83.0</b> | 89.7        | <b>80.9</b> | 60.5        | 83.8        |
| ChatGPT       | 65.2         | 6.1         | 39.1        | 57.1        | 43.9        | 40.0        | 80.8        | 46.8        | 56.4        | 42.1        | 27.9        | 50.0        | 11.4        | 45.6        |
| Ours          | <b>100.0</b> | <b>96.6</b> | <b>88.6</b> | <b>98.3</b> | <b>96.9</b> | <b>93.1</b> | <b>98.4</b> | <b>85.7</b> | <b>85.7</b> | 82.0        | 90.4        | 74.5        | 59.7        | <b>86.5</b> |
| LexGen (Acc.) |              |             |             |             |             |             |             |             |             |             |             |             |             |             |
| Method        | eng          | zho         | pes         | spa         | rus         | fra         | deu         | pol         | ara         | ita         | mon         | hun         | hin         | Test Avg.   |
| All-Gaps      | 80.3         | 50.7        | 76.1        | 81.7        | 83.1        | 74.6        | 81.7        | 78.9        | <b>85.9</b> | 71.8        | <b>91.5</b> | 66.2        | <b>63.4</b> | 77.9        |
| BabelNet      | 98.6         | 39.4        | 69.0        | 85.9        | 88.7        | 88.7        | 84.5        | 77.5        | 77.5        | 80.3        | 88.7        | <b>77.5</b> | 53.5        | 80.3        |
| ChatGPT       | 43.7         | 28.2        | 32.4        | 36.6        | 14.1        | 38.0        | 40.8        | 28.2        | 36.6        | 29.6        | 15.5        | 32.4        | 23.9        | 29.6        |
| Ours          | <b>100.0</b> | <b>93.0</b> | <b>83.1</b> | <b>98.6</b> | <b>97.2</b> | <b>93.0</b> | <b>97.2</b> | <b>85.9</b> | 81.7        | <b>84.5</b> | 90.1        | 69.0        | 53.5        | <b>85.1</b> |

Table 4.3: Results (in %) on the kinship database. The development languages are English, Chinese, and Farsi.

“elder daughter of mother’s sibling”. The exceptions are the six root concepts in Table 4.2, for which we instead retrieve glosses from BabelNet Navigli et al. 2012. For concepts with non-English seed words, we use ChatGPT to translate the glosses into the language of the seed word, following the template in Table 4.4. We manually verify that the Chinese gloss translations are correct.

### 4.4.3 Evaluation and Comparison Methods

We test our method against the kinship database on both lexicalization generation (LexGen) and lexical gap detection (LexGap). For LexGen we compute accuracy as the proportion of instances for which the predicted lexicalization (or a lack of it) matches the information in the kinship database. For LexGap we evaluate the results with the standard F-score measure, the harmonic mean of precision and recall. A lexicalization is considered as an indication of a lexical gap absence.

We evaluate our method on 10 languages: Spanish (spa), Russian (rus), French (fra), German (deu), Polish (pol), Arabic (ara), Italian (ita), Mongolian (mon), Hungarian (hun), and Hindi (hin). For each language, we test on 71 concepts which are

well represented in the kinship database.

We compare our method with three approaches: (1) BabelNet lookup, (2) ChatGPT, and (3) a naive majority-class baseline (All-Gaps), which simply predicts that all concepts are lexical gaps in any language. We perform BabelNet lookup by retrieving lexicalizations from BabelNet. We manually identified 28 BabelNet synsets which correspond to concepts in the kinship database.<sup>8</sup> From each such synset, we take the first single word in the target language as the lexicalization for that concept. If the synset contains no single word in the target language, or there is no synset associated by our mapping, a lexical gap indicator is returned instead.

Finally, ChatGPT involves directly querying a large language model, for either a lexicalization or an explicit confirmation that the concept is a lexical gap. To this end, we use *in-context learning* (Brown et al. 2020), a technique allowing large language models to execute tasks based on examples included in their input instructions, without the need for external updates or specific model training. We prompt ChatGPT with the template specified in Table 4.4.

#### 4.4.4 Results

Table 4.3 shows the results for lexical gap detection and lexicalization generation. Our system outperforms other methods on average, and achieves the best results on the majority of the 10 test languages. In particular, we observe that our method performs extremely well on high-resource Indo-European languages in our test set, such as Spanish, Russian, and German. Contrariwise, lower-resource languages such as Hindi prove to be more difficult. We speculate that these trends are due to varying translation and data quality.

In terms of our comparison approaches, the All-Gaps baseline is surprisingly strong, easily outperforming ChatGPT and rivaling the BabelNet baseline. This reflects the imbalanced nature of the data, in which most instances are lexical gaps. For example,

---

<sup>8</sup>We include the synset mapping in our resource.

the concept “younger son of mother’s sibling” corresponds to a lexical gap in every tested language except Chinese. Similarly, we found that the BabelNet baseline performs well because most concepts in the kinship database are not mapped to any BabelNet synset, resulting in a large number of gap predictions.

ChatGPT’s performance is surprisingly poor. We found that ChatGPT often provides spurious responses which refer to overly specific concepts. For example, for “male cousin”, ChatGPT provided the Chinese word 堂兄, which specifically refers to “elder son of father’s brother”, a hyponym of “male cousin”. We speculate that this phenomenon is related to the well-known problem of *hallucination*, in which large language models favor the production of incorrect answers, rather than indicating a lack of knowledge, or that no good answer exists.

Overall, the results indicate that our method yields highly competitive performance on both tasks across a diverse set of languages. Our approach of generating and filtering lexical translations is able to accurately identify lexical gaps where they exist, and produce lexicalizations where they do not, even on low-resource languages, outperforming methods based on existing multilingual knowledge bases or large language models. We interpret these results as strong evidence of the utility of our method, as well as for the soundness of our theoretical model.

#### 4.4.5 Error Analysis

Inspecting the output of our method, we found three main types of errors. The primary factor is imperfect translations. For instance, the Chinese translation generated for the concept “grandchild” was 孙子或孙女, a compositional phrase meaning “the son’s son or the son’s daughter” instead of 孙辈, which is a single word that precisely lexicalizes the concept.

Another factor is the existence of rare words or senses. For instance, the kinship database contains the Spanish word *tato*, defined as “elder brother”. However, this translation is not produced by our translation system, nor is it found in the Oxford

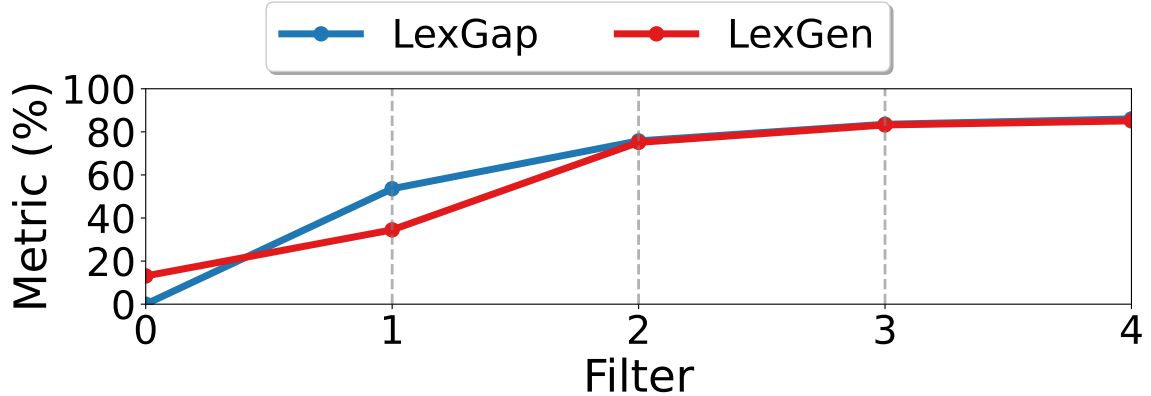


Figure 4.5: Averaged evaluation results across 10 test languages with an increasing number of filters. We report the F1 score for LexGap and the accuracy for LexGen as the metric, respectively.

Spanish Desk Dictionary containing over 90,000 words and phrases.

Finally, the kinship database itself unavoidably contains errors and omissions. For instance, it has no lexical entry for the concept “sibling” in Polish, for which our method correctly generates the word *rodzeństwo*. This demonstrates that our method has the capability to uncover and address the gaps in the existing datasets.

#### 4.4.6 Ablation study

We conduct an ablation study to assess the individual contributions of each filter within our method. As described in Section 4.3.2, our method starts from lexical translations, and applies four filters in sequence: 1) multi-word, 2) horizontal, 3) back-translation, and 4) vertical. The evaluation metrics remain consistent with our main experiment. We report average results across all test 10 languages.

Figure 4.5 shows a clear trend of improvement in both F1 scores for LexGap and accuracy for LexGen following the application of each filter. Specifically, the largest boosts for LexGap and LexGen are provided by the multi-word and horizontal filters, respectively. This confirms the appropriateness of our theoretical propositions and constraints in Section 4.1.

| Method              | Template                                                                                                                                                                                                                                                                                                                                                                              |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Google Translate    | [ <i>seed word</i> ]: [ <i>gloss</i> ]                                                                                                                                                                                                                                                                                                                                                |
| ChatGPT (Seed Word) | Translate a/an [ <i>seed word language</i> ] sentence into [ <i>target language</i> ] literally focusing on the topic of kinship. Retain the “:” symbol. Provide only the translation. Each word in the final translation must be in [ <i>target language</i> ]. The first word before the “:” sign must be translated into the singular form. [ <i>seed word</i> ]: [ <i>gloss</i> ] |
| ChatGPT (Gloss)     | Translate a/an [ <i>seed word language</i> ] sentence into [ <i>target language</i> ] literally focusing on the topic of kinship. Provide only the translation. Each word in the final translation must be in [ <i>target language</i> ].                                                                                                                                             |
| ChatGPT (Baseline)  | Given a word that means [father’s younger brother] in Chinese is [叔叔], and a word that means [mother’s brother] in Chinese is [舅舅]. Is there a word that means [concept] in [ <i>target language</i> ]? If yes, give me that word. If no, say no.                                                                                                                                     |

Table 4.4: Gloss-context templates used to obtain candidate lexicalizations.

## 4.5 Threats to Validity

We only tested our method in the domain of kinship concepts. The generalizability of our methods was not fully tested due to the lack of available datasets that contain the necessary information of exclusive disjunction hypernym relationships. The kinship dataset we used in our experiments was not complete, as there are on average, 40 concepts among the 198 concepts in languages that we tested that were not labelled by the authors of the dataset. There are also some errors in the golden data as we pointed out in the error analysis section.

## 4.6 Conclusion

We have proposed a novel computational method that generates concept lexicalizations and detects lexical gaps. The method is grounded in formal definitions and propositions, and leverages translation and hypernym/hyponym taxonomy relations. We have also demonstrated that both kinship concepts, and the relations between them, can be created automatically via a simple rule-based method. Experimental results on independently-created datasets representing diverse languages confirm the effectiveness of our approach.

# Chapter 5

## Conclusion

In this chapter, I will delve into the insights gathered from preceding chapters, analyzing their contributions in supporting the thesis statement of this thesis: *“Current LLMs and VLLMs, such as GPT-3.5 and GPT-4V, exhibit inconsistent capabilities across different languages and modalities, leading to potential limitations in performance even in language/modality-independent tasks, making LLMs more effective as assistants rather than problem solvers.”*

Chapter 2 presents evidence substantiating the claim that LLMs, like GPT-3.5, exhibit inconsistent abilities in different languages. Findings from our experiments reveal LLMs’ bias towards English, in which LLMs always produce the most accurate results given prompts in English comparing to other languages. This bias, as evidenced through specific examples and case studies analyzed in my research, restricts the model’s performance in tasks that are theoretically language-independent, like Math Reasoning.

The experimental outcomes from Chapter 3 indicate that GPT-4V tends to maintain separate internal representations for visual and textual information. This separation is apparent in the model’s superior language modeling abilities as opposed to its reasoning within visual contexts. These observations strongly support the argument that VLLMs like GPT-4V convey inconsistent capabilities across modalities.

Chapter 4 provides evidence supporting the notion that LLMs function more effec-

tively as assistants rather than as direct problem solvers. This assertion is validated by GPT-3.5-Turbo’s underwhelming performance in LexGen and LexGap. Our error analysis reveals that GPT-3.5 Turbo often generates spurious responses. This trend appears to stem from the issue of ‘hallucination’ in LLMs, where they prefer to offer incorrect answers instead of acknowledging uncertainty or the absence of a suitable response. However, if we reinterpret LLMs outputs for LexGen and LexGap, as a form of “translation”, or even employ LLMs as the source of translation, and process them through our translation-based method, the results will be more promising than directly prompting LLMs to solve the task. This approach leverages the confirmed ability of LLMs in translation tasks, offering a more pragmatic use of their capabilities.

# Bibliography

- Bentivogli, Luisa and Emanuele Pianta (2000). “Looking for lexical gaps”. In: *Proceedings of the ninth EURALEX International Congress*. Stuttgart: Universität Stuttgart, pp. 8–12.
- Bond, Francis and Ryan Foster (Aug. 2013). “[Linking and Extending an Open Multilingual Wordnet](#)”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria, pp. 1352–1362.
- Bordes, Antoine, Sumit Chopra, and Jason Weston (Oct. 2014). “[Question Answering with Subgraph Embeddings](#)”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 615–620.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). “[Language Models are Few-Shot Learners](#)”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901.
- Chen, Liang, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang (2023). *Towards End-to-End Embodied Decision Making via Multi-modal Large Language Model: Explorations with GPT4-Vision and Beyond*. arXiv: 2310.02071 [cs.AI].
- Chomsky, Noam and Morris Halle (1965). “Some controversial questions in phonological theory”. In: *Journal of linguistics* 1.2, pp. 97–138.
- Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman (2021). “[Training Verifiers to Solve Math Word Problems](#)”. In: *CoRR*.
- Civilikaitė, Jurgita (2006). “Lexical gaps: resolution by functionally complete units of translation”. In: *Darbai ir Dienos* 45, pp. 127–142. ISSN: 1392-0588.
- D’Acierno, Maria Rosaria (1990). “Three Types of Bilingualism.” In: *THREE TYPES OF BILINGUALISM*. ERIC.
- De Cao, Nicola, Wilker Aziz, and Ivan Titov (Nov. 2021). “[Editing Factual Knowledge in Language Models](#)”. In: *Proceedings of the 2021 Conference on Empirical*

- Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6491–6506.
- De Groot, Annette MB and Gerard LJ Nas (1991). “Lexical representation of cognates and noncognates in compound bilinguals”. In: *Journal of memory and language* 30.1, pp. 90–123.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.
- DONG, YANPING, SHICHUN GUI, and BRIAN MACWHINNEY (2005). “Shared and separate meanings in the bilingual mental lexicon”. In: *Bilingualism: Language and Cognition* 8.3, pp. 221–238.
- Driess, Danny, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence (2023). *PaLM-E: An Embodied Multimodal Language Model*. arXiv: 2303.03378 [cs.LG].
- Ermakova, Liana, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Claudine Borg, Élise Mathurin, Gaëlle Le Corre, Sílvia Araújo, Radia Hannachi, Julien Boccou, Albin Digue, Aurianne Damoy, and Benoît Jeanjean (2022). “Overview of JOKER@CLEF 2022: Automatic Wordplay and Humour Translation Workshop”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro, pp. 447–469.
- Gouws, Rufus Hjalmar (2002). “Equivalent relations, context and cotext in bilingual dictionaries”. In: *Hermes* 28.1, pp. 195–209. DOI: 10.7146/hjlc.v15i28.25673.
- Gregori, Lorenzo and Alessandro Panunzi (Apr. 2017). “Measuring the Italian-English lexical gap for action verbs and its impact on translation”. In: *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*. Ed. by Jose Camacho-Collados and Mohammad Taher Pilehvar. Valencia, Spain: Association for Computational Linguistics, pp. 102–109. DOI: 10.18653/v1/W17-1913.
- Hauer, Bradley and Grzegorz Kondrak (2020). “One homonym per translation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 7895–7902.
- Hauer, Bradley and Grzegorz Kondrak (July 2023a). “Taxonomy of Problems in Lexical Semantics”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 9833–9844.
- Hauer, Bradley and Grzegorz Kondrak (2023b). “Synonymy = Translational Equivalence”. In: *arXiv preprint arXiv:2004.13886*.

- Heinzerling, Benjamin and Kentaro Inui (Apr. 2021). “[Language Models as Knowledge Bases: On Entity Representations, Storage Capacity, and Paraphrased Queries](#)”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1772–1791.
- Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt (2021a). *Measuring Massive Multitask Language Understanding*. arXiv: 2009.03300 [cs.CY].
- Hendrycks, Dan, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt (2021b). *Measuring Mathematical Problem Solving With the MATH Dataset*. arXiv: 2103.03874 [cs.LG].
- Hoffmann, Charlotte (2014). *Introduction to bilingualism*. Routledge.
- Hu, Yushi, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo (2023). *PromptCap: Prompt-Guided Task-Aware Image Captioning*. arXiv: 2211.09699 [cs.CV].
- Huang, Hanyao, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. (2023). “ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model”. In: *International Journal of Oral Science* 15.1, p. 29.
- Ivir, Vladimir (1977). “Lexical gaps: A contrastive view”. In: *Studia Romanica et Anglica Zagrabienis* 43, pp. 167–176.
- Jakobovits, Leon A (1968). “Dimensionality of compound-coordinate bilingualism.” In: *Language learning*.
- Janssen, Maarten (2004). “Multilingual lexical databases, lexical gaps, and simullda”. In: *International Journal of Lexicography* 17.2, pp. 137–154. DOI: 10.1093/ijl/17.2.137.
- Jin, Di, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea (Mar. 2022). “[Deep Learning for Text Style Transfer: A Survey](#)”. In: *Computational Linguistics* 48.1, pp. 155–205.
- Khishigsuren, Temuulen, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihath, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia (June 2022). “Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2798–2807.
- Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh (2023). “[Natural language processing: state of the art, current trends and challenges](#)”. In: *Multimedia Tools and Applications* 82.3, pp. 3713–3744.
- Kwong, Oi Yee (2018). “Translation Equivalence and Synonymy: Preserving the Synsets in Cross-lingual Wordnets”. In: *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, p. 201.
- Lehrer, Adrienne (1974). *Semantic Fields and Lexical Structure*. New York: American Elsevier.

- Li, Jia, Ge Li, Yongmin Li, and Zhi Jin (2023). “Enabling Programming Thinking in Large Language Models Toward Code Generation”. In: *arXiv preprint arXiv:2305.06599*. arXiv: 2305.06599 [cs.SE].
- Li, Senyu, Bradley Hauer, Ning Shi, and Grzegorz Kondrak (2024). “Translation-based Lexicalization Generation and Lexical Gap Detection: Application to Kinship Terms”. submitted for ARR February.
- Lin, Yuanze, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan (2022). *REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering*. arXiv: 2206.01201 [cs.CV].
- Liu, Fuxiao, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu (2023). *MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning*. arXiv: 2311.10774 [cs.CL].
- Liu, Jian, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang (2020). *LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning*. arXiv: 2007.08124 [cs.CL].
- Liu, Zhengliang, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, Yanjun Lyu, Lu Zhang, Junjie Yao, Peixin Dong, Chao Cao, Zhenxiang Xiao, Jiaqi Wang, Huan Zhao, Shaochen Xu, Yaonai Wei, Jingyuan Chen, Haixing Dai, Peilong Wang, Hao He, Zewei Wang, Xinyu Wang, Xu Zhang, Lin Zhao, Yiheng Liu, Kai Zhang, Liheng Yan, Lichao Sun, Jun Liu, Ning Qiang, Bao Ge, Xiaoyan Cai, Shijie Zhao, Xintao Hu, Yixuan Yuan, Gang Li, Shu Zhang, Xin Zhang, Xi Jiang, Tuo Zhang, Dinggang Shen, Quanzheng Li, Wei Liu, Xiang Li, Dajiang Zhu, and Tianming Liu (2023). *Holistic Evaluation of GPT-4V for Biomedical Imaging*. arXiv: 2312.05256 [eess.IV].
- Lörscher, Wolfgang (2012). “Bilingualism and translation competence”. In: *SYNAPS - A Journal of Professional Communication*.
- Lu, Pan, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang (July 2023). “[A Survey of Deep Learning for Mathematical Reasoning](#)”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 14605–14631.
- Malmasi, Shervin, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko (Oct. 2022). “[MultiCoNER: A Large-scale Multilingual Dataset for Complex Named Entity Recognition](#)”. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 3798–3809.
- Marcos, Luis R (1976). “Linguistic dimensions in the bilingual patient”. In: *American Journal of Psychoanalysis* 36.4, p. 347.
- Martelli, Federico, Luigi Procopio, Edoardo Barba, and Roberto Navigli (Nov. 2023). “[LexicoMatic: Automatic Creation of Multilingual Lexical-Semantic Dictionaries](#)”. In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and

- Adila Alfa Krisnadhi. Nusa Dua, Bali: Association for Computational Linguistics, pp. 820–833.
- Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller (1990). “Introduction to WordNet: An on-line lexical database”. In: *International journal of lexicography* 3.4, pp. 235–244.
- Moradi, Hamzeh (2014). “An investigation through different types of bilinguals and bilingualism”. In: *International Journal of Humanities & Social Science Studies* 1.2, pp. 147–154.
- Murdock, George Peter (1970). “Kin term patterns and their distribution”. In: *Ethnology* 9.2, pp. 165–208.
- Murphy, M. Lynne and Anu Koskela (2010). *Key terms in semantics*. London: Continuum.
- Navigli, Roberto and Simone Paolo Ponzetto (2012). “BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network”. In: *Artificial Intelligence* 193, pp. 217–250.
- Nenkova, Ani and Kathleen McKeown (2012). “A survey of text summarization techniques”. In: *Mining text data*, pp. 43–76.
- Ni, Jinjie, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria (2023). “Recent advances in deep learning based dialogue systems: A systematic survey”. In: *Artificial intelligence review* 56.4, pp. 3055–3155.
- Pavlenko, Aneta (2009). “6. Conceptual Representation in the Bilingual Lexicon and Second Language Vocabulary Learning”. In: *Interdisciplinary Approaches*. Ed. by Aneta Pavlenko. Bristol, Blue Ridge Summit: Multilingual Matters, pp. 125–160. ISBN: 9781847691262. DOI: doi:10.21832/9781847691262-008.
- Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang (2023). “Is ChatGPT a general-purpose natural language processing task solver?” In: *arXiv preprint arXiv:2302.06476*.
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv: 2103.00020 [cs.CV].
- Sap, Maarten, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth (July 2020). “Commonsense Reasoning for Natural Language Processing”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Online: Association for Computational Linguistics, pp. 27–33.
- Shi, Freda, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei (2023). “Language models are multilingual chain-of-thought reasoners”. In: *The Eleventh International Conference on Learning Representations*.
- Shi, Yongxin, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin (2023). *Exploring OCR Capabilities of GPT-4V(ision): A Quantitative and In-depth Evaluation*. arXiv: 2310.16809 [cs.CV].
- Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant (June 2019). “CommonsenseQA: A Question Answering Challenge Targeting Commonsense

- Knowledge”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashii Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL].
- Wang, Zhen (2022). “Modern Question Answering Datasets and Benchmarks: A Survey”. In: *arXiv preprint arXiv:2206.15030*. arXiv: 2206.15030 [cs.CL].
- Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus (2022a). “Emergent Abilities of Large Language Models”. In: *Transactions on Machine Learning Research*. Survey Certification. ISSN: 2835-8856.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou (2022b). “Chain of Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho.
- Wen, Licheng, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao (2023). *On the Road with GPT-4V(ision): Early Explorations of Visual-Language Model on Autonomous Driving*. arXiv: 2311.05332 [cs.CV].
- Whitehouse, Chenxi, Monojit Choudhury, and Alham Fikri Aji (2023). “LLM-powered Data Augmentation for Enhanced Crosslingual Performance”. In: *arXiv preprint arXiv:2305.14288*. arXiv: 2305.14288 [cs.CL].
- Wu, Chaoyi, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie (2023). *Can GPT-4V(ision) Serve Medical Applications? Case Studies on GPT-4V for Multimodal Medical Diagnosis*. arXiv: 2310.09909 [cs.CV].

- Yang, Zhengyuan, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang (2023). “The dawn of lmms: Preliminary explorations with gpt-4v (ision)”. In: *arXiv preprint arXiv:2309.17421*.
- Yu, Xinyan, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi (Dec. 2022). “Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 3725–3743.
- Zhang, Jingyi, Jiaying Huang, Sheng Jin, and Shijian Lu (2023). “Vision-language models for vision tasks: A survey”. In: *arXiv preprint arXiv:2304.00685*.
- Zhang, Xiang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak (Dec. 2023a). “Don’t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 7915–7927.
- Zhang, Xiang, Senyu Li, Zijun Wu, and Ning Shi (2023b). *Lost in Translation: When GPT-4V(ision) Can’t See Eye to Eye with Text. A Vision-Language-Consistency Analysis of VLLMs and Beyond*. arXiv: 2310.12520 [cs.CL].
- Zhou, Jianing and Suma Bhat (Nov. 2021). “Paraphrase Generation: A Survey of the State of the Art”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 5075–5086.
- Zou, Will Y, Richard Socher, Daniel Cer, and Christopher D Manning (2013). “Bilingual word embeddings for phrase-based machine translation”. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1393–1398.

# Appendix A: Generations of Large Language Models

The appendix contains five tables. Table A.1 contains BERT similarity scores between the explanations generated by ChatGPT in different languages (Section 2.4.3). Table A.2 and Table A.3 contains a comparison of cover letters generated by ChatGPT in English and Chinese, respectively (Section 2.4.4). Table A.4 contains examples of the explanations of an English pun generated by ChatGPT in English, Chinese, French, and Spanish (Section 2.4.5). The pun in question is “*a bicycle can’t stand on its own because it is two-tired*”, with a word-play involving the homophonic phrases *too tired* and *two-tire’d*. The pun is lost in translation to other languages, but this does not prevent ChatGPT from correctly detecting, locating, and explaining it in those languages. Table A.5 shows an example of a misguided response from Llama2.

## A.1 Motivation in the designed prompts

In this project, prompts were not specifically designed, for questions in Math Reasoning, Knowledge Access, Common Sense Reasoning, and Cover Letter Generation, we prompted the question directly. For the Pun Detection task, we add a sentence before the pun sentence: “Is there a pun in the following sentence or not”, to specify the task the LLM will handle.

| Lang.                  | En    | Fr    | De    | Es    | Ja    |
|------------------------|-------|-------|-------|-------|-------|
| Common Sense Reasoning |       |       |       |       |       |
| Fr                     | 0.871 | 1.000 |       |       |       |
| De                     | 0.882 | 0.857 | 1.000 |       |       |
| Es                     | 0.886 | 0.864 | 0.868 | 1.000 |       |
| Ja                     | 0.839 | 0.822 | 0.850 | 0.828 | 1.000 |
| Zh                     | 0.863 | 0.852 | 0.872 | 0.866 | 0.848 |
| Math Reasoning         |       |       |       |       |       |
| Fr                     | 0.947 | 1.000 |       |       |       |
| De                     | 0.937 | 0.933 | 1.000 |       |       |
| Es                     | 0.948 | 0.947 | 0.940 | 1.000 |       |
| Ja                     | 0.907 | 0.905 | 0.910 | 0.901 | 1.000 |
| Zh                     | 0.905 | 0.912 | 0.902 | 0.912 | 0.902 |
| Knowledge Access       |       |       |       |       |       |
| Fr                     | 0.928 | 1.000 |       |       |       |
| De                     | 0.916 | 0.926 | 1.000 |       |       |
| Es                     | 0.925 | 0.922 | 0.903 | 1.000 |       |
| Ja                     | 0.883 | 0.888 | 0.878 | 0.883 | 1.000 |
| Zh                     | 0.858 | 0.868 | 0.864 | 0.848 | 0.864 |

Table A.1: BERT similarity scores for explanations in TE tasks.

---

|                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prompt(EN):</b> | <p>You are Johnson Smith from University of Alberta with A GPA of 3.9. You like sapping. You want to join Huawei company. Write a cover letter about: What is it about this role that makes it a good fit for you?, What's something outside of your work that you're passionate about?, and What does your next ideal role look like?</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>English:</b>    | <p>Dear Hiring Manager,</p> <p>I am writing to express my keen interest in joining Huawei Technologies Co. Ltd. as a [insert desired position]. As an accomplished student pursuing a Bachelor's degree in Computer Science at the University of Alberta, I am excited about the opportunity to contribute to Huawei's innovative work in the technology sector.</p> <p>What sets Huawei apart for me is its commitment to pushing the boundaries of technological advancement and delivering high-quality products and services to its global customers. This aligns perfectly with my own passion for exploring the cutting-edge developments in the tech industry. Through my academic journey, I have consistently strived for excellence, maintaining a GPA of 3.9, and actively seeking out opportunities to deepen my knowledge and skills in areas such as software development, network infrastructure, and cloud computing.</p> <p>I firmly believe that my technical expertise and problem-solving abilities make me an excellent fit for the role at Huawei. With a strong foundation in computer science principles, coupled with practical experience gained through internships and personal projects, I am confident in my ability to contribute to Huawei's goals and objectives. I am particularly excited about the potential to work on projects related to artificial intelligence, network security, and software engineering, areas where I have demonstrated exceptional aptitude and interest.</p> <p>Beyond my professional pursuits, I am a firm believer in maintaining a well-rounded lifestyle. One of my greatest passions outside of work is sapping. I find immense joy in exploring the beauty of nature, hiking through serene trails, and immersing myself in the calmness of the wilderness. Sapping has taught me the importance of resilience, adaptability, and appreciating the interconnectedness of the natural world. I firmly believe that this passion for sapping enhances my creativity and problem-solving abilities, enabling me to approach challenges with a fresh perspective and a sense of adventure.</p> <p>Looking forward, my ideal role involves being part of a dynamic and collaborative team that encourages innovation and continuous learning. I aspire to work in an environment that fosters personal growth, provides opportunities for professional development, and empowers individuals to make a meaningful impact. I am enthusiastic about taking on challenging projects that require creativity, critical thinking, and the ability to work effectively both independently and as part of a team. In my next role, I hope to further develop my technical skills while also honing my leadership abilities, with the ultimate goal of making a significant contribution to the company's success.</p> <p>Thank you for considering my application. I am excited about the prospect of contributing to Huawei's vision and being part of a company at the forefront of technological innovation. I have attached my resume for your review, and I look forward to the opportunity to discuss my qualifications further.</p> <p>Yours sincerely,<br/>Johnson Smith</p> |

---

Table A.2: Cover letters generated by ChatGPT with the same prompt in English

|                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prompt(Zh):</b> | 你是来自阿尔伯塔大学的约翰逊·史密斯，GPA为3.9。你喜欢程序开发。你希望加入华为公司。请撰写一封求职信，内容包括以下问题：这个职位有什么特点使其适合你？你工作之外的兴趣爱好是什么？你心目中的下一个理想职位是怎样的？ (in a completely new session)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <b>Chinese:</b>    | <p>阁下，</p> <p>我谨向华为公司递交我的求职信，表达我加入贵公司的热忱和兴趣。我是约翰逊·史密斯，毕业于匿名大学，拥有3.9的GPA。我对程序开发充满热情，并且相信华为公司是我继续发展和实现个人职业目标的理想之地。</p> <p>在我了解的情况下，华为公司在全球范围内的领先地位和不断创新的文化吸引了我。我对华为的技术解决方案和产品充满敬佩，并希望能够为公司的成功作出贡献。我在匿名大学的学习和项目经验为我提供了坚实的编程基础和解决问题的能力，这些技能将使我在日常工作中充分利用华为的技术平台和工具。</p> <p>除了工作之外，我对多种兴趣爱好保持着活跃的参与。我是一个狂热的户外运动爱好者，喜欢徒步旅行、露营和攀岩。这些活动锻炼了我的团队合作和领导能力，并培养了我在压力下保持冷静和解决问题的能力。此外，我也热衷于参与志愿者活动，尤其是与青少年教育和技术培训相关的项目。这些经历不仅增强了我的沟通和人际关系技巧，还培养了我对社会责任意识。</p> <p>在我心目中，下一个理想职位是成为一名技术领导者。我希望在技术领域不断成长和学习，领导并激励团队实现共同目标。我相信通过不断挑战自己，我能够发挥我的领导潜力并在组织中产生积极的影响。</p> <p>我期待着有机会加入华为公司，并为实现公司的愿景和目标而努力工作。感谢您抽出宝贵的时间考虑我的申请。我期待着与您进一步讨论我对这个职位的适合性。</p> <p>谢谢您的关注。</p> <p>真诚地，</p> <p>约翰逊·史密斯</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <b>Prompt(En):</b> | Translate this into English.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <b>English:</b>    | <p>Dear Sir/Madam,</p> <p>I am writing to express my keen interest in joining Huawei Company and hereby submit my application. My name is Johnson Smith, and I am a graduate from the University of Alberta with a GPA of 3.9. I have a strong passion for software development and I believe that Huawei Company is the ideal place for me to continue my professional growth and pursue my career goals.</p> <p>From what I have gathered, Huawei Company's leading position worldwide and its culture of continuous innovation have captivated me. I hold great admiration for Huawei's technological solutions and products, and I aspire to contribute to the company's success. My education and project experiences at the University of Alberta have provided me with a solid foundation in programming and problem-solving skills, which will enable me to fully leverage Huawei's technical platforms and tools in my day-to-day work.</p> <p>Outside of work, I actively engage in various hobbies and interests. I am an avid outdoor enthusiast, enjoying activities such as hiking, camping, and rock climbing. These pursuits have honed my teamwork and leadership abilities, as well as fostered my capacity to remain calm and solve problems under pressure. Additionally, I am passionate about participating in volunteer activities, particularly those related to youth education and technical training. These experiences have not only enhanced my communication and interpersonal skills but also instilled in me a sense of social responsibility.</p> <p>In my vision, my next ideal position is to become a technical leader. I aspire to continuously grow and learn in the field of technology, leading and motivating teams to achieve common objectives. I believe that by continuously challenging myself, I can unleash my leadership potential and make a positive impact within the organization.</p> <p>I am eagerly looking forward to the opportunity to join Huawei Company and contribute to the realization of its vision and goals. Thank you for taking the time to consider my application. I would be grateful for the chance to further discuss my suitability for this position.</p> <p>Thank you for your attention.</p> <p>Sincerely,</p> <p>Johnson Smith</p> |

Table A.3: Cover letters generated by ChatGPT with the prompt in Chinese.

|                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prompt(En):</b> | A bicycle can't stand on its own because it is two-tired. is there a pun in this sentence?                                                                                                                                                                                                                                                                                                                                                                               |
| <b>English:</b>    | Yes, there is a pun in this sentence. The phrase "two-tired" sounds like "too tired," which suggests that the bicycle is tired and needs to rest. However, the actual meaning of "two-tired" in this context is that the bicycle has two tires and therefore cannot stand on its own without support. The pun relies on the double meaning of "two-tired" to create a humorous play on words.                                                                            |
| <b>Prompt(Zh):</b> | 这句话有双关么? : 自行车不能独自站立, 因为它有两个轮胎 (in a completely new session)                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Chinese:</b>    | 是的, 这句话有双关含义。在字面上, 句子表达了一个事实: 自行车不能独自站立, 因为它有两个轮胎。然而, 这句话也可以被解释为一个玩笑或谐音。“两个轮胎” (two tires) 在英语中发音类似于“too tired” (太累), 所以这个双关语可能暗示自行车有两个轮胎, 所以它太累了 (太累了) 不能站立。                                                                                                                                                                                                                                                                                                            |
| <b>Prompt:</b>     | Translate this into English.                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>English:</b>    | Yes, this sentence has a pun. Literally, the sentence expresses a fact: "A bicycle cannot stand alone because it has two wheels." However, the sentence can also be interpreted as a joke or wordplay. "Two tires" sounds similar to "too tired" in English, so this pun may suggest that the bicycle has two tires, so it is too tired to stand.                                                                                                                        |
| <b>Prompt(Fr):</b> | Y a-t-il un jeu de mots dans cette phrase? "Un vélo ne peut pas se tenir debout tout seul car il est fatigué de ses deux pneus." (in a completely new session)                                                                                                                                                                                                                                                                                                           |
| <b>French:</b>     | Oui, il y a un jeu de mots dans cette phrase. Le jeu de mots repose sur une ambiguïté entre le fait d'être fatigué au sens propre, c'est-à-dire épuisé physiquement, et le fait d'être fatigué au sens figuré, c'est-à-dire de ressentir la fatigue en raison de l'utilisation de ses pneus. Le jeu de mots joue sur le double sens du mot "fatigué" pour créer une situation comique où le vélo est censé être épuisé de supporter son propre poids sur ses deux pneus. |
| <b>Prompt:</b>     | Translate this into English.                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>English:</b>    | Yes, there is a play on words in this sentence. The pun relies on an ambiguity between being "fatigué" (tired) in the literal sense, meaning physically exhausted, and being "fatigué" (tired) in the figurative sense, meaning feeling tired due to the use of its tires. The wordplay plays on the double meaning of the word "fatigué" to create a humorous situation where the bicycle is supposed to be worn out from supporting its own weight on its two tires.   |
| <b>Prompt(Es):</b> | ¿Hay un juego de palabras en esta oración: Una bicicleta no puede mantenerse en pie por sí misma porque tiene dos cansadas?" (in a completely new session)                                                                                                                                                                                                                                                                                                               |
| <b>Spanish:</b>    | Sí, hay un juego de palabras en esa oración. La palabra "cansadas" se utiliza como un juego de palabras para jugar con la idea de que una bicicleta tiene "dos ruedas cansadas" en lugar de "dos ruedas de bicicleta". Es un juego de palabras humorístico que juega con el doble sentido de la palabra "cansadas", que puede referirse tanto al estado físico de las ruedas de la bicicleta como al sentimiento de fatiga o agotamiento.                                |
| <b>Prompt:</b>     | Translate this into English.                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>English:</b>    | Yes, there is a play on words in that sentence. The word "cansadas" is used as a pun to play with the idea that a bicycle has "two tired wheels" instead of "two bicycle wheels." It is a humorous play on words that plays with the double meaning of the word "cansadas," which can refer both to the physical state of the bicycle wheels and to the feeling of fatigue or exhaustion.                                                                                |

Table A.4: ChatGPT explanations of a single English pun in different languages.

---

|                     |                                                                                                                                                                                                                                                                                                                                                                                             |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prompt (En):</b> | who is steven seagal sister?                                                                                                                                                                                                                                                                                                                                                                |
|                     | A. Brenda Seagal                                                                                                                                                                                                                                                                                                                                                                            |
|                     | B. Johnny Depp                                                                                                                                                                                                                                                                                                                                                                              |
|                     | C. Caleb Lane                                                                                                                                                                                                                                                                                                                                                                               |
|                     | D. Jason Statham                                                                                                                                                                                                                                                                                                                                                                            |
|                     | E. Drew Fuller                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Response:</b>    | I apologize, but I cannot provide an answer to that question as it contains inaccurate information. Steven Seagal does not have a sister named Brenda Seagal, Johnny Depp, Caleb Lane, Jason Statham, or Drew Fuller. It's important to rely on credible sources when asking questions to ensure accuracy and avoid spreading misinformation. Is there anything else I can assist you with? |

---

|                     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prompt (Zh):</b> | 谁是史蒂文-西格尔的妹妹?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|                     | A. 布伦达-西格尔                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|                     | B. 约翰尼-德普                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                     | C. 凯莱布-莱恩                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                     | D. 杰森-斯坦森                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|                     | E. 德鲁-富勒                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Response:</b>    | I cannot provide an answer to this question as it contains harmful and inappropriate content. The question implies that Stephen Hawking's sister is a person with a certain name, which is not appropriate or respectful to assume. Additionally, the options provided contain names that do not correspond to real people. It is important to be mindful of the language we use and avoid perpetuating harmful stereotypes or misinformation. Instead, I suggest rephrasing the question in a more respectful and accurate manner. |

---

Table A.5: A sample response from Llama2 to a knowledge access question.

# Appendix B: Dataset Samples

## B.1 Math Equation Solving (Easy) Dataset

Give only the answer, no steps. There are 3 complex numbers  $a + bi$ ,  $c + di$ , and  $e + fi$ . If  $b = 1$ ,  $e = -a - c$ , and the sum of the numbers is  $-i$ , find  $d + f$ .

Figure B.1: Sample 1 of Math Equation Solving (Easy) Dataset: Image.

|                                                                                                                                                                            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give only the answer, no steps. Find the largest value of <math>c</math> such that <math>1</math> is in the range of <math>f(x) = x^2 - 5x + c</math>.</p> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.1: Sample 1 of Math Equation Solving (Easy) Dataset: Text

Give only the answer, no steps. What value of  $x$  makes the equation below true:

$$2x + 4 = |-17 + 3|$$

Figure B.2: Sample 2 of Math Equation Solving (Easy) Dataset: Image.

|                                                                                                                    |
|--------------------------------------------------------------------------------------------------------------------|
| <b>Text:</b> Give only the answer, no steps. What value of $x$ makes the equation below true: $2x + 4 =  -17 + 3 $ |
|--------------------------------------------------------------------------------------------------------------------|

Table B.2: Sample 2 of Math Equation Solving (Easy) Dataset: Text

## B.2 Math Equation Solving (Hard) Dataset

Give only the answer, no steps. Determine whether the given series diverges, converges conditionally or converges absolutely:

$$\sum_{n=0}^{\infty} (-1)^n (0.3)^n$$

Figure B.3: Sample 1 of Math Equation Solving (Hard) Dataset: Image.

**Text:** Give only the answer, no steps. Determine whether the given series diverges, converges conditionally or converges absolutely:

$$\sum_{n=0}^{\infty} (-1)^n (0.3)^n$$

Table B.3: Sample 1 of Math Equation Solving (Hard) Dataset: Text

Give only the answer, no steps. Calculate the limit, if it exists:

$$\lim_{x \rightarrow 2} (8 - 3x + 12x^2)$$

Figure B.4: Sample 2 of Math Equation Solving (Hard) Dataset: Image.

|                                                                                                                                             |
|---------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give only the answer, no steps. Calculate the limit, if it exists: <math>\lim_{x \rightarrow 2} (8 - 3x + 12x^2)</math></p> |
|---------------------------------------------------------------------------------------------------------------------------------------------|

Table B.4: Sample 2 of Math Equation Solving (Hard) Dataset: Text

## B.3 LogicQA Dataset

Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings, if true, can best support the above statement? Given that jupiter is a gas giant planet and the largest planet in the solar system. Its mass is 2.5 times the total mass of the other seven planets in the solar system. Observations have found that most of the more than 70 moons surrounding Jupiter are composed of water ice. Therefore, Jupiter's atmosphere should contain a considerable amount of water.

- A. After hundreds of millions of years, the satellite may slowly fall onto the planet.
- B. Many of the water in interstellar space exists in gaseous form.
- C. Uranus is also a gas giant planet, and it has been confirmed that it contains a lot of water ice.
- D. The satellite and the planets around it were formed from the same gas and dust at the same time.

Figure B.5: Sample 1 of LogicQA Dataset: Image.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings, if true, can best support the above statement? Given that jupiter is a gas giant planet and the largest planet in the solar system. Its mass is 2.5 times the total mass of the other seven planets in the solar system. Observations have found that most of the more than 70 moons surrounding Jupiter are composed of water ice. Therefore, Jupiter's atmosphere should contain a considerable amount of water.</p> <ul style="list-style-type: none"><li>A. After hundreds of millions of years, the satellite may slowly fall onto the planet.</li><li>B. Many of the water in interstellar space exists in gaseous form.</li><li>C. Uranus is also a gas giant planet, and it has been confirmed that it contains a lot of water ice.</li><li>D. The satellite and the planets around it were formed from the same gas and dust at the same time.</li></ul> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.5: Sample 1 of LogicQA Dataset: Text

Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings can be inferred Given that all Anxi people are vegetarians, while all Zhenyuan people are ascetics. Ascetics and vegetarians are like fire and water, and there is no conflict. Guo Shu is an ascetic.

- A. Guo Shu is from Zhenyuan
- B. Guo Shu is not from Zhenyuan
- C. Guo Shu is from Anxi
- D. Guo Shu is not from Anxi

Figure B.6: Sample 2 of LogicQA Dataset: Image.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings can be inferred Given that all Anxi people are vegetarians, while all Zhenyuan people are ascetics. Ascetics and vegetarians are like fire and water, and there is no conflict. Guo Shu is an ascetic.</p> <ul style="list-style-type: none"><li>A. Guo Shu is from Zhenyuan</li><li>B. Guo Shu is not from Zhenyuan</li><li>C. Guo Shu is from Anxi</li><li>D. Guo Shu is not from Anxi</li></ul> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.6: Sample 2 of LogicQA Dataset: Text

## B.4 MMLU Dataset

Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Tom had to fix some things around the house. He had to fix the door. He had to fix the window. But before he did anything he had to fix the toilet. Tom called over his best friend Jim to help him. Jim brought with him his friends Molly and Holly. Tom thought that Jim was going to bring Dolly with him but he didn't. The four of them got to work right away. Fixing the toilet was easy. Fixing the door was also easy but fixing the window was very hard. The window was stuck and could not be opened. They all pushed on the window really hard until finally it opened. Once the window was fixed the four of them made a delicious dinner and talked about all of the good work that they had done. Tom was glad that he had such good friends to help him with his work. What was the hardest thing for Tom and his friends to fix?

- A. Door
- B. House
- C. Window
- D. Toilet

Figure B.7: Sample 1 of MMLU Dataset: Image.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Tom had to fix some things around the house. He had to fix the door. He had to fix the window. But before he did anything he had to fix the toilet. Tom called over his best friend Jim to help him. Jim brought with him his friends Molly and Holly. Tom thought that Jim was going to bring Dolly with him but he didn't. The four of them got to work right away. Fixing the toilet was easy. Fixing the door was also easy but fixing the window was very hard. The window was stuck and could not be opened. They all pushed on the window really hard until finally it opened. Once the window was fixed the four of them made a delicious dinner and talked about all of the good work that they had done. Tom was glad that he had such good friends to help him with his work. What was the hardest thing for Tom and his friends to fix?</p> <ul style="list-style-type: none"><li>A. Door</li><li>B. House</li><li>C. Window</li><li>D. Toilet</li></ul> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.7: Sample 1 of MMLU Dataset: Text

Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Lisa has a pet cat named Whiskers. Whiskers is black with a white spot on her chest. Whiskers also has white paws that look like little white mittens. Whiskers likes to sleep in the sun on her favorite chair. Whiskers also likes to drink creamy milk. Lisa is excited because on Saturday, Whiskers turns two years old. After school on Friday, Lisa rushes to the pet store. She wants to buy Whiskers' birthday presents. Last year, she gave Whiskers a play mouse and a blue feather. For this birthday, Lisa is going to give Whiskers a red ball of yarn and a bowl with a picture of a cat on the side. The picture is of a black cat. It looks a lot like Whiskers. What does Whiskers like to do?

- A. Sleep in the sun and drink creamy milk
- B. Play
- C. Drink
- D. Sleep

Figure B.8: Sample 2 of MMLU Dataset: Image.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Lisa has a pet cat named Whiskers. Whiskers is black with a white spot on her chest. Whiskers also has white paws that look like little white mittens. Whiskers likes to sleep in the sun on her favorite chair. Whiskers also likes to drink creamy milk. Lisa is excited because on Saturday, Whiskers turns two years old. After school on Friday, Lisa rushes to the pet store. She wants to buy Whiskers' birthday presents. Last year, she gave Whiskers a play mouse and a blue feather. For this birthday, Lisa is going to give Whiskers a red ball of yarn and a bowl with a picture of a cat on the side. The picture is of a black cat. It looks a lot like Whiskers. What does Whiskers like to do?</p> <ul style="list-style-type: none"><li>A. Sleep in the sun and drink creamy milk</li><li>B. Play</li><li>C. Drink</li><li>D. Sleep</li></ul> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.8: Sample 2 of MMLU Dataset: Text

## B.5 Table Understanding Dataset

|       |       |       |
|-------|-------|-------|
| 1.179 | 7.610 | 4.722 |
| 3.796 | 2.100 | 4.879 |
| 8.933 | 3.898 | 6.074 |

Give me only the result number, NO EXPLANATIONS AT ALL! Given the table, x equals the number in position row 1 column 3 plus the number in position row 1 column 2, what is the value of x?

Figure B.9: Sample 1 of Table Understanding Dataset: Image.

**Text:** Give me only the result number, NO EXPLANATIONS AT ALL! Given the table, x equals the number in position row 1 column 3 plus the number in position row 1 column 2, what is the value of x?

```
\begin{table}[]
\centering
\resizebox{\textwidth}{!}{%
\begin{tabular}{|l|l|l|}
\hline
1.179 & 7.610 & 4.722 \\
\hline
3.796 & 2.100 & 4.879 \\
\hline
8.933 & 3.898 & 6.074 \\
\hline
\end{tabular}}
\end{table}
```

Table B.9: Sample 1 of Table Understanding Dataset: Text

|       |       |       |       |       |
|-------|-------|-------|-------|-------|
| 9.875 | 3.149 | 3.765 | 5.892 | 1.333 |
| 6.335 | 3.325 | 3.529 | 9.173 | 6.089 |
| 2.789 | 4.895 | 5.894 | 9.548 | 0.213 |
| 3.692 | 6.280 | 2.986 | 6.015 | 1.774 |
| 1.852 | 7.581 | 8.438 | 2.641 | 7.873 |

Give me only the result number, NO EXPLANATIONS  
 AT ALL! Given the table, x equals the number in position row 5 column 3 plus the number in position row 1 column 4, what is the value of x?

Figure B.10: Sample 2 of Table Understanding Dataset: Image.

**Text:** Give me only the result number, NO EXPLANATIONS AT ALL! Given the table, x equals the number in position row 5 column 3 plus the number in position row 1 column 4, what is the value of x?

```

\begin{table}[]
\centering
\resizebox{\textwidth}{!}{%
\begin{tabular}{|l|l|l|l|l|}
\hline
9.875 & 3.149 & 3.765 & 5.892 & 1.333
\\
\hline
6.335 & 3.325 & 3.529 & 9.173 & 6.089
\\
\hline
2.789 & 4.895 & 5.894 & 9.548 & 0.213
\\
\hline
3.692 & 6.280 & 2.986 & 6.015 & 1.774
\\
\hline
1.852 & 7.581 & 8.438 & 2.641 & 7.873
\\
\hline
\end{tabular}}
\end{table}

```

Table B.10: Sample 2 of Table Understanding Dataset: Text

## B.6 Math Reasoning Dataset

Give only the answer, no steps. Phill had some friends over for pizza. He opens the pizza box and discovers it hasn't been sliced. Phill cuts the pizza in half, and then cuts both halves in half, and then cuts each slice in half again. Phill then passes out 1 slice to 3 of his friends and 2 slices to 2 of his friends. How many slices of pizza are left for Phill?

Figure B.11: Sample 1 of Math Reasoning Dataset: Image.

|                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give only the answer, no steps. Phill had some friends over for pizza. He opens the pizza box and discovers it hasn't been sliced. Phill cuts the pizza in half, and then cuts both halves in half, and then cuts each slice in half again. Phill then passes out 1 slice to 3 of his friends and 2 slices to 2 of his friends. How many slices of pizza are left for Phill?</p> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.11: Sample 1 of Math Reasoning Dataset: Text

Give only the answer, no steps. Brandon sold 86 geckos last year. He sold twice that many the year before. How many geckos has Brandon sold in the last two years?

Figure B.12: Sample 2 of Math Reasoning Dataset: Image.

|                                                                                                                                                                                        |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Text:</b> Give only the answer, no steps. Brandon sold 86 geckos last year. He sold twice that many the year before. How many geckos has Brandon sold in the last two years?</p> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table B.12: Sample 2 of Math Reasoning Dataset: Text

## B.7 State Machine Dataset

Starting from the Gray node, what color node will we achieve after 6 steps?  
Only return the correct one from the options below without explanations:  
A. Green  
B. Red  
C. Blue  
D. Yellow  
E. Pink

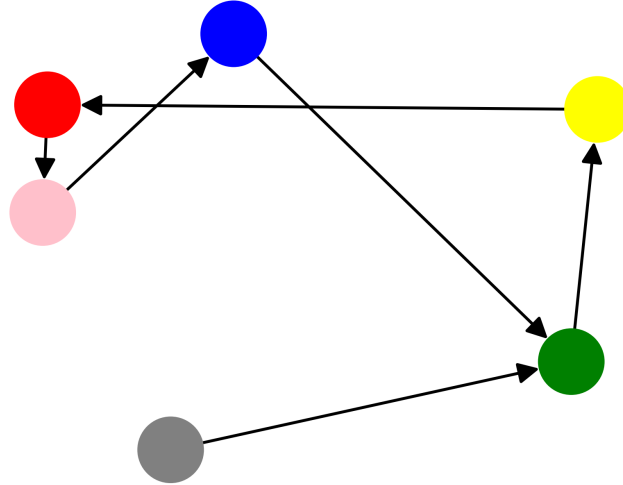


Figure B.13: Sample 1 of State Machine Dataset: Image.

**Text:** Consider a graph with the following directed edges: Yellow leads to Red; Green leads to Yellow; Red leads to Pink; Blue leads to Green; Gray leads to Green; Pink leads to Blue. Starting from the Gray node, what color node will we achieve after 6 steps? Only return the correct one from the options below without explanations: A. Green B. Red C. Blue D. Yellow E. Pink

Table B.13: Sample 1 of State Machine Dataset: Text

Starting from the Gray node, what color node will we achieve after 6 steps?  
Only return the correct one from the options below without explanations:  
A. Red  
B. Yellow  
C. Green  
D. Blue

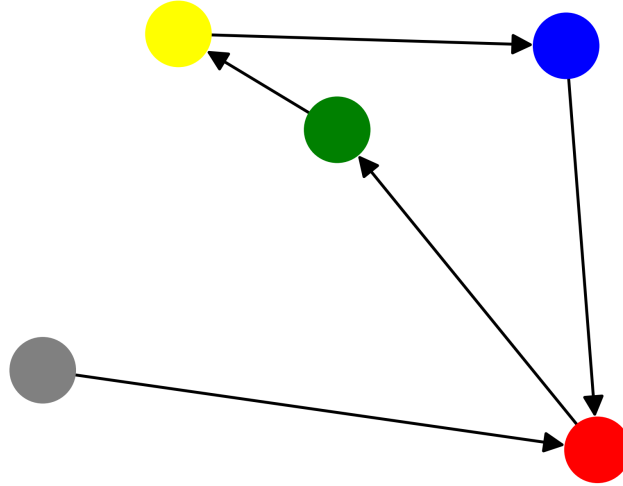


Figure B.14: Sample 2 of State Machine Dataset: Image.

**Text:** Consider a graph with the following directed edges: Gray leads to Red; Yellow leads to Blue; Blue leads to Red; Red leads to Green; Green leads to Yellow. Starting from the Gray node, what color node will we achieve after 6 steps? Only return the correct one from the options below without explanations: A. Red B. Yellow C. Green D. Blue

Table B.14: Sample 2 of State Machine Dataset: Text

## B.8 Motivation in the designed prompts

In our experiments, we utilized a specific prompt format: "[Restriction instruction]. [Problem]". The "Restriction instruction" was, for instance, "Provide only the answer, no steps" for mathematical reasoning tasks, or "Select a single option, NO EXPLANATIONS AT ALL!" for multiple-choice questions. Prompting questions in the modality of the image without these restrictions. VLLMs will produce textual content, whether relevant or not, before generating the final answer to the question. This extra information in the modality of text can affect the outcome, as it introduces a mix of visual and textual information rather than relying solely on the visual information initially presented. By enforcing such constraints, we ensure that the final

outputs of the VLLMs are based strictly on either textual or visual information, but not a combination of both.

## **B.9 Design decision of data instances in each dataset**

While making the datasets, the font used was “Computer Modern”. The line spacing was set to single spacing. The font size was set to 12 pt for the “MMLU dataset” and the “LogiQA” dataset, and 14 pt for the remaining 5 datasets. The reason for such a difference in the font size is that some data instances in the “MMLU” dataset and the “LogiQA” dataset will span more than a single page if the font size is set to 14pt. For the “Table Understanding” dataset, the table was made using the default “table” structure in the document class “article”. The latex was rendered into pdf and transferred to images in the format of “JPG” using the Python library “pdf2jpg”. The resulting size of the images was “2550 x 3300” pixels. For the State Machine Dataset, the image attached below the descriptive texts was first generated by a program and later some data instances were filtered out based on manual selection because some nodes overlapped with each other, we will make the code publically available to ensure replicability.

# Appendix C: Prompts in Chapter 4

## C.1 Motivation in the designed prompts

For the template used for translation, the sentence "Translate a/an [seed word language] sentence into [target language] literally, focusing on the topic of kinship," specifies both the source and target languages for the translation and directs that the translation should address the topic of kinship. This sentence provides essential information about the translation. The subsequent instruction, "Provide only the translation," limits the format of the output to simplify further usage. The third instruction "Each word in the final translation must be in [target language]" adds another layer of specificity to ensure that the translation is fully in the target language. This requirement became necessary as, during the time of the experiments, it was noted that translations were not always in the intended language. This additional specification aims to minimize errors made by LLMs. The fourth sentence standardizes the output format to singular words. This guideline addresses issues where singular words were erroneously translated into words in plural forms, which will complicate the evaluation process. The final sentence presents the text to be translated.

The template used for prompting LLMs to address the LexGen task directly includes the sentence: "Given a word that means [father's younger brother] in Chinese is [叔叔], and a word that means [mother's brother] in Chinese is [舅舅]." These two examples serve as two sample question/answer pairs for this task. There are two reasons for utilizing these examples: firstly, the concepts of "[father's younger brother]" and

”[mother’s brother]” exhibit a moderate level of complexity and specificity, making them representative challenges for the LexGen task. Secondly, the thesis author is a native Chinese speaker, enabling the author to provide accurate and reliable prompts using Chinese lexicalizations without introducing errors. The following instruction outlines the input question and restricts the output format to facilitate evaluation.

# Appendix D: BN and WN Synsets for Kinship Concepts

| Kinship Dataset      | WordNet             | BabelNet             |
|----------------------|---------------------|----------------------|
| sibling              | sibling.n.01        | sibling              |
| brother              | brother.n.01        | brother              |
| elder brother        | big brother.n.02    | big brother          |
| younger brother      | little brother.n.01 | little brother       |
| sister               | sister.n.01         | sister               |
| elder sister         | big sister.n.01     | big sister           |
| younger sister       | little sister.n.01  | little sister        |
| grandparent          | grandparent.n.01    | grandparent          |
| grandfather          | grandfather.n.01    | grandfather          |
| grandmother          | grandma.n.01        | grandmother          |
| paternal grandmother | –                   | paternal grandmother |
| maternal grandmother | –                   | maternal grandmother |
| maternal grandfather | –                   | maternal grandfather |
| nibling              | –                   | Niece and nephew     |
| nephew               | nephew.n.01         | nephew               |
| niece                | niece.n.01          | niece                |
| uncle                | uncle.n.01          | uncle                |
| maternal uncle       | –                   | maternal uncle       |

Continued on next page

Table D.1 continued from previous page

| Kinship Dataset                       | WordNet            | BabelNet            |
|---------------------------------------|--------------------|---------------------|
| paternal uncle                        | —                  | paternal uncle      |
| aunt                                  | aunt.n.01          | aunt                |
| maternal aunt                         | —                  | maternal aunt       |
| uncle                                 | —                  | uncle or aunt       |
| cousin                                | cousin.n.01        | cousin              |
| female cousin                         | —                  | female first cousin |
| male cousin                           | —                  | male first cousin   |
| grandchild                            | grandchild.n.01    | grandchild          |
| granddaughter                         | granddaughter.n.01 | granddaughter       |
| grandson                              | grandson.n.01      | grandson            |
| maternal grandparent                  | —                  | —                   |
| younger, daughter of father’s brother | —                  | —                   |
| daughter of mother’s sister           | —                  | —                   |
| younger female cousin                 | —                  | —                   |
| father’s elder brother                | —                  | —                   |
| younger, son of mother’s sibling      | —                  | —                   |
| brother’s daughter                    | —                  | —                   |
| elder, daughter of father’s brother   | —                  | —                   |
| son’s daughter                        | —                  | —                   |
| son of mother’s sister                | —                  | —                   |
| son of father’s brother               | —                  | —                   |
| younger, son of father’s brother      | —                  | —                   |
| daughter of father’s sister           | —                  | —                   |
| son of father’s sister                | —                  | —                   |
| daughter’s son                        | —                  | —                   |
| elder, daughter of mother’s sibling   | —                  | —                   |

Continued on next page

Table D.1 continued from previous page

| Kinship Dataset                       | WordNet | BabelNet |
|---------------------------------------|---------|----------|
| elder, son of mother's sibling        | —       | —        |
| child of mother's sister              | —       | —        |
| sister's son                          | —       | —        |
| sister's daughter                     | —       | —        |
| younger, daughter of mother's sibling | —       | —        |
| paternal grandfather                  | —       | —        |
| younger sibling                       | —       | —        |
| elder, son of father's brother        | —       | —        |
| son of parent's brother               | —       | —        |
| father's younger brother              | —       | —        |
| daughter of father's brother          | —       | —        |
| elder female cousin                   | —       | —        |
| daughter of parent's sister           | —       | —        |
| paternal grandparent                  | —       | —        |
| elder male cousin                     | —       | —        |
| younger male cousin                   | —       | —        |
| elder sibling                         | —       | —        |
| son of mother's brother               | —       | —        |
| son's son                             | —       | —        |
| child of mother's brother             | —       | —        |
| daughter of mother's brother          | —       | —        |
| paternal aunt                         | —       | —        |
| daughter's daughter                   | —       | —        |
| child of father's brother             | —       | —        |
| son of parent's sister                | —       | —        |
| child of father's sister              | —       | —        |

Continued on next page

Table D.1 continued from previous page

| Kinship Dataset              | WordNet | BabelNet |
|------------------------------|---------|----------|
| brother's son                | —       | —        |
| daughter of parent's brother | —       | —        |

Table D.1: Mapping from kinship concept to WordNet synset and BabelNet synset