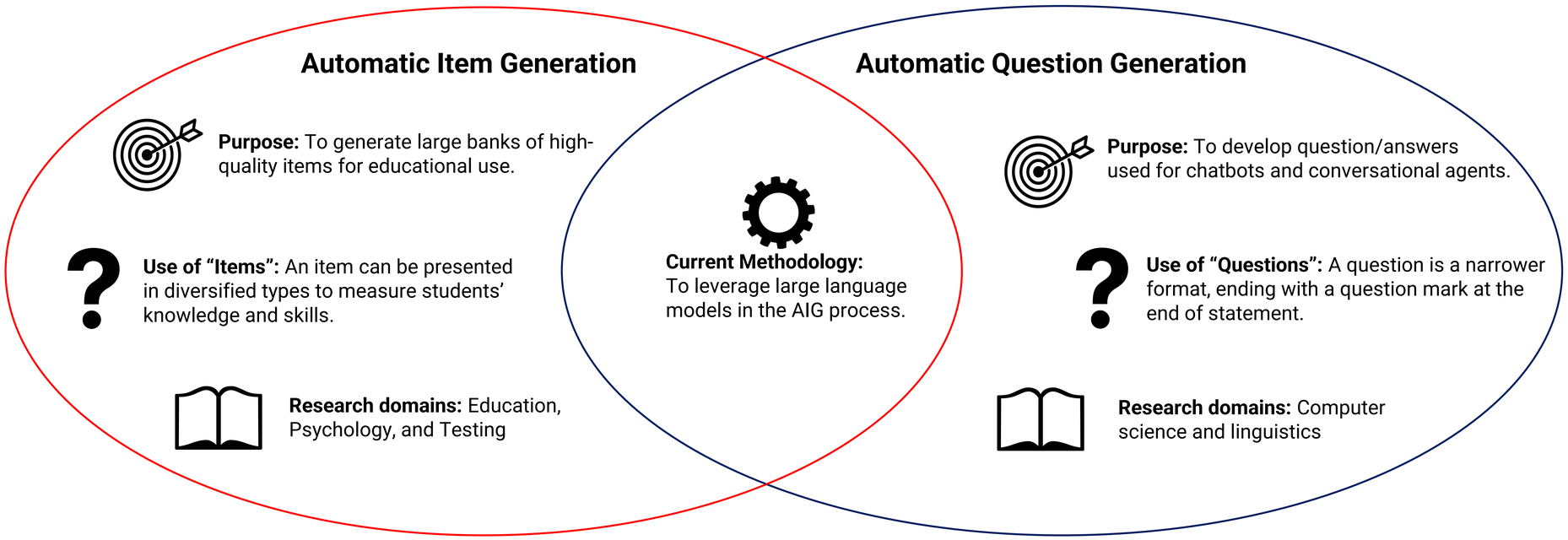
A Review of Automatic Item Generation in the Large Language Model Era

Bin Tan, Nour Armoush, Elisabetta Mazzullo, Okan Bulut, Mark Gierl

UNIVERSITY OF ALBERTA

MEASUREMENT, EVALUATION, AND DATA SCIENCE

Background: Automatic Item Generation VS. Automatic Question Generation



Purpose of this study

It is now crucial to raise awareness of both Automatic Item Generation (AIG) and Automatic Question Generation (AQG) research. By integrating experiences from both fields, we can gain a thorough understanding of the existing literature and evaluate the potential and usefulness of large language models (LLMs) for AIG.

Study objectives:

- To identify what has been achieved in the literature regarding the use of LLMs for AIG
- To explore the knowledge gaps between the two research domains.

Current State of Research

- The most used types of LLMs are T5, BERT, GPT, and their variants.
- The specific use of LLMs for AIG can be broken down into three stages:
  - **Pre-generation stage:** This involves cleaning, structuring, and understanding the original input text data.
  - **Item-generation stage:** Items are created either directly by LLMs or through traditional methods.
  - **Post-generation stage:** Generated items are selected based on certain criteria such as answerability.
- There were distinct usage patterns of BERT, GPT, and T5 across the three AIG stages.
- LLMs can be an effective and flexible solution to generating a large number of items, with few constraints on item type, language, subject domain, or the data source used for training LLMs to create items.

Current Gap in the Literature

- We searched for some keywords in the 60 reviewed studies and here are the results:

Keyword searched	Validity	Reliability	Fairness	Item difficulty	Item discrimination
# of studies	10	8	3	0	1
Keyword searched	Pedagogical	Bloom's taxonomy	Cognitive level	Content specialists	Subject matter experts
# of studies	8	0	3	1	3

- Many studies **lack an educational foundation**, as they did not incorporate learning and measurement theories in the AIG process.
- A thorough **item evaluation** after item generation is missing in the current literature.
- AIG does not end with generating a large number of items, but with ensuring that the items generated are of high quality and can fulfill their demands for educational purposes and contexts.

Suggestions for Future Research

- Our first suggestion is to **clarify the assessment context and measurement goals** for AIG.
  - As the assessment purposes (e.g., formative vs. summative assessments) differ, the desired characteristics of the generated items also differ.
  - Measurement goals clarify what to measure, why it is being measured, and what the measurement process aims to accomplish.
- The second suggestion is to **evaluate item quality** to ensure their usability in educational contexts.
  - Measurement properties: reliability, validity, and fairness
  - Pedagogical soundness

References: https://shorturl.at/pzV68