Generating Educational Questions using Large Language Models: An Evaluation of Quality and Alignment with Pedagogical Principles

Malte Grube

April 15, 2025

1 Introduction and Research Motivation

Assessment plays a crucial role in education, with practice questions being essential tools for students to evaluate their understanding of instructional material. Large Language Models (LLMs) present a comfortable opportunity to automate the generation of such educational questions. This thesis explores the capabilities of State-of-the-Art LLMs in generating diverse, effective, and pedagogically valuable questions from computer science instructional texts.

The research focuses on how these generated questions align with established educational frameworks, particularly Bloom's Taxonomy of cognitive levels. While current research has explored LLMs for content generation, there remains a significant gap in understanding their effectiveness for educational question generation that adheres to specific pedagogical principles and maintains fidelity to source materials. The aim is to provide practical insights for educators considering LLM-assisted question creation, with implications for both educational practice and the advancement of AI applications in education.

2 Research Questions

This thesis addresses the following key research questions:

1. To what extent do LLMs adhere to the content of diverse provided instructional texts when generating questions?

2. How does the relationship between diverse question formats and Bloom's Taxonomy levels influence the pedagogical effectiveness of LLM-generated questions?

3 Theoretical Framework

3.1 Bloom's Taxonomy in Educational Assessment

Bloom's Taxonomy provides a hierarchical framework for categorizing educational objectives according to cognitive complexity. Using the revised taxonomy (Anderson & Krathwohl, 2001) with its six levels (remembering, understanding, applying, analyzing, evaluating, and creating), this thesis examines how LLM-generated questions can be aligned with these cognitive levels to create comprehensive assessments.

3.2 Question Types

Different question formats serve distinct pedagogical purposes. This research evaluates the LLMs' ability to generate questions across various question formats.

4 Methodology and Research Design

This thesis will employ a mixed-methods empirical study with the following components:

4.1 Experiment Design Overview

The research will be structured around two complementary experiments:

4.1.1 Input Variation Impact

This experiment will systematically examine how different types of input materials affect the quality of LLM-generated questions. These will be paired with both a common and a complex prompt to assess how input format influences question quality, relevance, and faithfulness to source material. Additionally, the input materials will be modified once to contain contradictory information, allowing for an evaluation of how LLMs handle inconsistencies in the source material.

4.1.2 Bloom's Taxonomy and Question Type Guidance

This experiment will investigate how explicitly specifying Bloom's Taxonomy levels and / or question types in prompts affects the quality and cognitive level of generated questions. The experiment will use either the highest-performing input material identified from Experiment 1 or LLM system knowledge if the results from Experiment 1 are inconclusive. The experimental conditions will include:

- Question type specification only
- Bloom's level specification only
- Combined specification of both Bloom's level and question type

4.2 LLM Selection and Technical Implementation

Based on preliminary assessment of capabilities and accessibility, four state-of-the-art LLMs have been selected: OpenAI o3-mini, Google Gemini 2.0 Flash **or** the recently published Gemini 2.5 Pro, Anthropic Claude 3.7 Sonnet, DeepSeek R1.

4.3 Test Corpus Selection

A carefully curated collection of computer science instructional texts will be assembled, focusing particularly the ISO-OSI reference model. This ensures content validity while maintaining sufficient complexity for meaningful assessment.

4.4 Evaluation Methodology

The evaluation will combine quantitative and qualitative approaches:

4.4.1 Quantitative Metrics

- **Semantic similarity:** Measuring cosine similarity between source text and generated questions using embedding models.
- **Bloom's level adherence ratio:** The proportion of questions correctly matching the specified cognitive level, done with an LLM providing feedback to each question's level.

4.4.2 Qualitative Assessment

- **Depth evaluation:** Sample-based rating of questions on a 1-5 scale for cognitive complexity.
- **Content fidelity:** Analysis of how faithfully questions represent source material concepts.
- **Error propagation:** Assessment of how LLMs handle inconsistencies in source materials.

5 Expected Results and Contributions

This research is expected to yield the following contributions:

- A comprehensive evaluation for assessing LLM-generated educational questions
- Empirical data on the comparative performance of leading LLMs in educational question generation
- Practices for prompt engineering to achieve questions at targeted cognitive levels
- Guidelines for educators on effective LLM-assisted question generation
- Insights into the relationship between input material quality, its correctness and output question effectiveness.

6 Timeline and Work Plan

The thesis will be completed over a 20-week period with the following phases:

- **Weeks 1-4:** Literature review and test corpus preparation
- **Week 5:** Implementation of Experiment 1 – First Run
- **Weeks 6-7:** Implementation of Experiment 2 – First Run
- **Week 8:** Data analysis and preliminary results compilation
- **Week 9:** Experiment 1 – Second Run with probable adjustments
- **Weeks 10-11:** Experiment 2 – Second Run with probable adjustments
- **Week 12:** Data analysis and preliminary results compilation
- **Week 13:** Intermediate defense
- **Weeks 14-15:** Results interpretation and discussion
- **Weeks 16-17:** Concentrating on the thesis document
- **Week 18:** Review and feedback incorporation
- **Weeks 19-20:** Final revisions and submission