Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Can We Learn Together? How LLMs Adapt to Educational Levels

Antun DROBNJAK[a*] & Ivica BOTICKI[a]
[a]*Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia*
*antun.drobnjak@fer.hr

Abstract: This study investigates the adaptability of Large Language Models (LLMs) in educational contexts by designing tailored prompts for various educational levels and evaluating the models' performance using readability scores. Generated responses were analyzed to assess their alignment with the intended educational levels. Findings reveal differences in the models' ability to adjust content complexity, highlighting the importance of prompt design in achieving appropriate readability. This research underscores the potential of LLMs to enhance educational materials through prompt engineering, contributing to more personalized and accessible learning experiences.

Keywords: Large Language Models, Educational Adaptability, AI in Education, Prompt Engineering for Education

1. Introduction

Large language models (LLMs) have revolutionized natural language understanding and generation, enabling nuanced dialogues, content creation, and problem-solving support in educational environments (Li et al., 2024).

The capacity of an Artificial Intelligence (AI) tutor in tailoring language complexity, instructional strategy, and feedback depth to a learner's level is critical to effective pedagogy (Kinder et al., 2025; Rooein et al., 2023). Primary students need simplified vocabulary and scaffolding, whereas advanced undergraduates benefit from open-ended challenges and higher-order prompts (Weijers et al., 2024). Interventions based on LLMs should mitigate the risk of under- or over-challenging students to maintain engagement and learning gains.

Bouchra et al. (2025) compared ChatGPT and DeepSeek on scientific queries across educational levels, showing that ChatGPT excels at sparking discussion but can lack depth, while DeepSeek better organizes technical resources. Likewise, Huang et al. (2024) demonstrated that few-shot prompting lets GPT-3.5, LLaMA-2 70B, and Mixtral 8×7 B adapt materials to specific readability levels without sacrificing meaning.

This research aims to explore how adaptable LLMs are across user groups. A range of LLMs were instructed to generate responses tailored to four distinct educational levels: Elementary School (ages 5-10), Middle School (ages 11-13), High School (ages 14-18), and Higher Education (ages 18 and above) (A Guide to U.S. Education Levels, n.d.).

Accordingly, this study is guided by the following key research questions:

1. Can LLMs adopt a tone and linguistic complexity suitable for younger audiences while also generating sophisticated, nuanced content for advanced learners
2. How do the adaptations of tone and linguistic complexity influence the overall effectiveness of the answer, in terms of clarity, engagement, and educational value.

2. Methodology

To answer this study research questions, outputs to questions from six different LLMs were analyzed: ChatGPT 4omini, Claude 3.7 Sonnet, DeepSeek-V3, Duck.ai Llama 3.3 70B, Microsoft Copilot – FastResponse, and Gemini 2.0Flash. A range of linguistic and readability metrics was applied to evaluate these responses to a set of robotics related questions – *Explain what a robot is for someone new to robotics.*; *Describe different types of robots, such as industrial robots, humanoid robots, and autonomous vehicles.*; *Describe the key*

components of a robot, including sensors, actuators, controllers, and power sources.; *Explain how robots move using different mechanisms such as wheels, legs, and tracks.*
Describe common sensors used in robotics and how they help robots perceive their environment.; Discuss different ways robots are powered, such as batteries, solar energy, and fuel cells.; Explain how robotic arms work and their applications in industries like manufacturing and healthcare.; Explain how robots are programmed and the different methods used, such as manual programming and AI-based learning.; Describe how artificial intelligence helps robots make decisions and adapt to their environment.; Explain how self-driving cars work and the role of sensors and AI in their operation.; Describe how robots use cameras and image processing to recognize objects and navigate spaces.; Discuss the future of robotics and how advancements in AI and technology might shape new robotic systems.

Coherence measures how logically and consistently ideas are presented, while semantic similarity assesses the alignment of each model's output with the intended meaning. Readability was analyzed using several grade-level metrics: the Flesch-Kincaid, Gunning Fog, and Dale-Chall scores, which provided insights into sentence complexity and word familiarity, which is critical for tailoring content to different age groups. Additional indices like the Automated Readability Index (ARI), Coleman-Liau, and Spache offer alternative perspectives based on character count, sentence structure, and vocabulary difficulty, each relevant for distinguishing between elementary, middle, and higher education levels. Finally, the Flesch Reading Ease offered a holistic view of how accessible the responses were, with higher scores indicating easier-to-read content.

A zero-shot Hugging Face classification model[1] labeled each answer as "relevant" or "irrelevant" using the hypothesis *"The {answer} to the {prompt} is relevant/irrelevant."*. Every model's responses were classified as relevant, confirming they addressed the prompts across all educational levels. When interacting with AI, especially in educational settings, it is essential to specify the target audience's level so that the response can be tailored appropriately. In this study, a prompt was structured as follows: *Please* answer *the following query in a way that is suitable for the specified level of education. Level of education: "Elementary school/middle school/High school/Higher education" Query: "{Question from table}"*

3. Results

The graphs presented in Figure 1 reveal distinct strengths of specific language models for the given educational levels. For elementary school content, Claude, Duck.ai, Gemini, and ChatGPT stood out for their readability, with Claude and Duck.ai excelling in simplicity and coherence, whereas Gemini's simplicity was slightly offset by lower coherence. Other models, such as DeepSeek and Microsoft Copilot, provided good results but with more lexical difficulty.

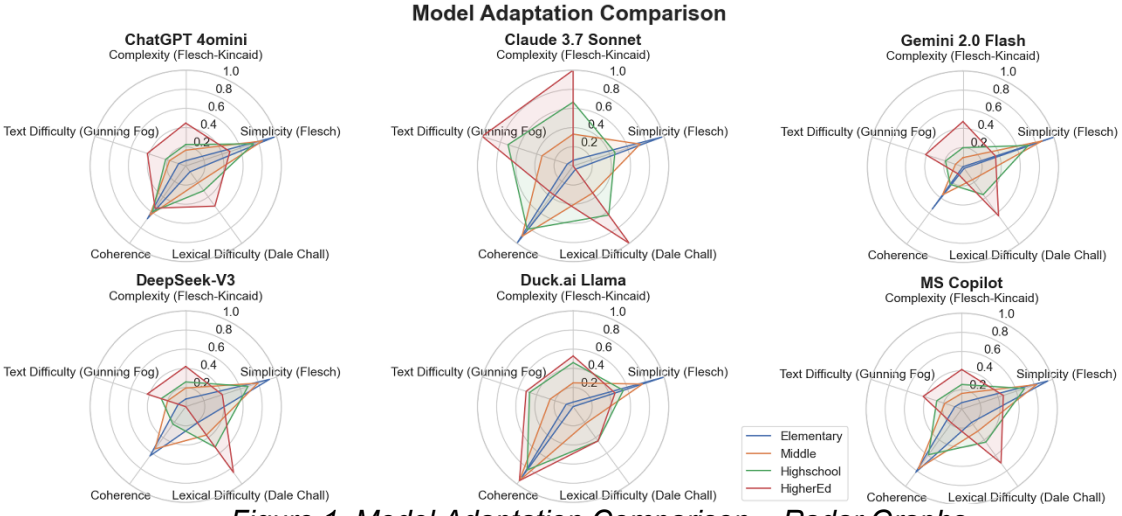

Figure 1. Model Adaptation Comparison – Radar Graphs

[1] https://huggingface.co/tasks/zero-shot-classification

For middle school, ChatGPT and Microsoft Copilot struck a balance of simplicity and sophistication, offering moderately complex texts. Claude and Duck.ai shifted towards higher complexity, which might be more challenging for this educational level. High school content requires more mature and clear language, where ChatGPT maintained a good balance between complexity and simplicity, while Claude and Duck.ai produced denser texts, potentially too advanced for some high school readers. Gemini and Microsoft Copilot offered more moderate outputs, with Gemini's coherence slightly lagging. At the higher education level, models like ChatGPT and Duck.ai presented more complex outputs with advanced vocabulary, while Claude produced particularly dense and challenging text. DeepSeek-V3 registered the lowest coherence scores, substantially limiting its suitability for higher education; Gemini 2.0 Flash likewise showed suboptimal cohesion, whereas Microsoft Copilot balanced intermediate complexity and lexical difficulty with higher textual cohesion.

4. Conclusions

Defining the educational level for the desired output when interacting with LLMs and AI is essential for several reasons. When the intended audience is specified, the AI can tailor its language accordingly. This involves adjusting vocabulary, sentence structure, and the overall complexity to ensure the final text is both accessible and engaging for the target reader.

The metrics chosen in this study highlight each model's context-specific strengths: some (e.g., Gemini, ChatGPT) are naturally inclined towards more accessible language suited for younger audiences, while others (e.g., Claude, DeepSeek) were tuned for the complex and detailed prose expected at higher academic levels. The choice of a model should be driven by aligning its inherent output characteristics with the educational objectives, ensuring language complexity, coherence, and readability without compromising clarity. Clearly defining educational levels helps in maintaining the appropriate tone and context.

It is important to mention that in education, the trustworthiness of LLMs, such as GPT-4, Claude, and LLaMA, can vary significantly across dimensions such as factual accuracy, bias, robustness, and safety (Huang et al., 2023; Liang et al., 2023), potentially influencing students' understanding and decision-making. As LLMs expand their role in educational contexts, ensuring trustworthiness remains crucial for the integrity of the learning process.

References

A guide to U.S. education levels. (n.d.). USAHello. Retrieved April 18, 2025, from https://usahello.org/education/children/grade-levels/

Bouchra, A., Hanane, K., & Mohamed, L. (2025). Evaluating ChatGPT and DeepSeek for Science Education: A Comparative Analysis of AI-Powered Learning Assistants.

Huang, Y., Zhang, Q., Y, P. S., & Sun, L. (2023). TrustGPT: A Benchmark for Trustworthy and Responsible Large Language Models (No. arXiv:2306.11507). arXiv. https://doi.org/10.48550/arXiv.2306.11507

Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. Computers and Education: Artificial Intelligence, 8, 100349. https://doi.org/10.1016/j.caeai.2024.100349

Li, Q., Fu, L., Zhang, W., Chen, X., Yu, J., Xia, W., Zhang, W., Tang, R., & Yu, Y. (2024). Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges (No. arXiv:2401.08664). arXiv. https://doi.org/10.48550/arXiv.2401.08664

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). Holistic Evaluation of Language Models (No. arXiv:2211.09110). arXiv. https://doi.org/10.48550/arXiv.2211.09110

Rooein, D., Curry, A. C., & Hovy, D. (2023). Know Your Audience: Do LLMs Adapt to Different Age and Education Levels? (No. arXiv:2312.02065). arXiv. https://doi.org/10.48550/arXiv.2312.02065

Weijers, R., De Castilho, G. F., Godbout, J.-F., Rabbany, R., & Pelrine, K. (2024). Quantifying learning-style adaptation in effectiveness of LLM teaching. Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), 112–118. https://aclanthology.org/2024.personalize-1.10/