Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 5 n.4 (2025)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Pre-service Teachers Comparative Evaluation of Large Language Models: A survey among Learning Sciences -

Emiliana Murgia [1], Flippo Bruni[2,*]

[1] Università di Genova; emiliana.murgia@edu.unige.it
[2] Università del Molise; filippo.bruni@unimol.it
* Correspondence: emiliana.murgia@edu.unige.it

Abstract: Abstract: The rise of generative Artificial Intelligence (AI) necessitates moving beyond digital competencies to AI literacy for educators. This study examines preservice primary teachers' knowledge, perceptions, and preferences regarding generative AI tools, specifically comparing ChatGPT and Gemini for instructional design tasks. Understanding user perceptions can help tailoring effective teacher training pathways. A sample of 172 preservice teachers completed three comparative tasks using both ChatGPT and Gemini: generating lesson designs from given prompts, creating prompts following guidelines, and generating mathematical problems. Participants rated models on six performance dimensions and provided open-ended justifications. Data were analyzed using qualitative content analysis of 721 responses and segmentation analysis across experience levels, gender, and task types. ChatGPT received majority preference (67% overall), with consistent superiority across all tasks and experience levels. Qualitative analysis revealed three primary evaluation criteria: completeness, coherence, and organization. Experienced users showed stronger ChatGPT preference (72%) compared to novices (61%). Task complexity moderated preferences, with ChatGPT demonstrating strongest advantages in creative and analytical tasks. In conclusion, User preferences are driven by functional attributes—particularly response organization and coherence—rather than advanced features. Effective teacher training requires differentiated pathways based on prior experience, explicit instruction in prompt engineering, and continuous assessment of evolving AI tools.

Keywords: LLMs Comparative Evaluation; Preservice teacher training; AI Teachers' Perception; AI Literacy; ChatGPT and Gemini.

1. Introduction

Teachers' digital competencies have been widely debated for decades. The rise of generative AI and its implications in teaching and learning require a further investigation, both in how to present AI in general terms, and in which tools to select in teaching activities. From digital competences defined as 'the confident, critical and responsible use of, and engagement with, digital technologies for learning, at work, and for participation in society' (European Commission, 2019, p. 10), we need to move on adding AI Literacy (AIL), i.e. to 'a set of competencies that enables individuals to evaluate AI technologies critically; communicate and collaborate effective-

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

ly with AI; and use AI as a tool online, at home, and in the workplace' (Long & Magerko, 2020; Biagini, 2025).

Within this context, there is a need to find appropriate ways for initial training of primary school teachers. Whilst certain objectives must necessarily be established a priori regarding the courses themselves, the specification of particular learning outcomes, the selection of pedagogical instruments, and the sequencing of instructional activities should be conceived with due consideration for individual contexts and, consequently, for the competencies, perceptions, and requirements of trainee teachers (De La Higuera, 2019; Sperling, et al., 2024). In this work, we aim to answer two research questions: RQ1, What are preservice teachers' knowledge and the perceptions of generative AI? RQ2, Comparing two generative AI Apps, Gemini and ChatGPT, in instructional design activities, which are preservice teachers' preferences and perceptions?'

The common side of these research questions is the importance of perception. Although perception is a subjective data point, therefore quite separate from the objective parameters used to compare different generative AI tools, it is a crucial element for understanding learners' predispositions, interests, and motivation levels toward the adoption of a technology (Cope & Ward, 2002, Murgia & Bruni 2024). Furthermore, it provides valuable insights for tailoring training pathways to different initial conditions.

The context for the research was the Educational Technology workshops held at the University of Molise (Learning Sciences) during the 2024-25 academic year. One of the eight meetings was dedicated to AI. The guiding principle underpinning the design and facilitation of this session was that, through recognising students' tacit and often unconscious applications, teachers can primarily utilise AI for instructional design activities. Amongst the activities proposed, this study focuses on three of them: in the first, students were asked to employ the models to generate an instructional design from a given prompt; in the second, to compose a prompt following standard guidelines and subsequently generate an instructional design; and in the third, a mathematics problem. For all activities, students used the same prompt (either provided or self-generated) on two different artificial intelligence tools: ChatGPT and Gemini. The aim for this approach is linked to two objectives: firstly, to ascertain, through students' perceptions, potential differences in effectiveness or characteristics amongst the various generative AI tools; and secondly, to cultivate—given the risk of hasty usage leading to frequently uncontrolled outcomes—a conscious and critical approach.

2. Materials and Methods

2.1. Participants

A convenience sample of 172 participants was recruited at the Educational Technology workshop at the faculty of Learning Sciences The sample was predominantly female (n = 167, 97.1%) with limited male representation (n = 5, 2.9%). Educational attainment varied, with participants not having any university's degree, while others holding bachelor's degrees (laurea triennale) and master's degrees (laurea magistrale). Whilst convenience sampling limits generalisability, it is commonly

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

employed in exploratory technology evaluation research where rapid feedback on emerging systems is prioritised over population representativeness (Hargittai, 2020). Furthermore, within the specific context of primary preservice teachers, the sample is representative of the Italian reality, where aspiring primary teachers are predominantly female and possess no prior degree, as the qualification pathway consists of a five-year integrated master's degree programme with no intermediate bachelor's degree.

AI experience levels were distributed as follows: no experience (n = 133, 77.3%), a few months of experience (n = 11, 6.4%), less than 3 years (n = 6, 3.5%), and more than 1 year (n = 4, 2.3%). The majority reported daily usage of AI tools ("quotidianamente", on a daily basis) across multiple contexts. This distribution reflects the current state of AI adoption, where most users remain in early stages of engagement with large language models (Dwivedi et al., 2023).

2.2. Materials: the questionnaire

The questionnaire consisted of five sections administered in Italian via Google Forms. This tool was selected for its accessibility, ease of use, and ability to collect both structured and unstructured data in a single instrument (Vasantha Raju & Harinarayana, 2016). In Section 1 were collected data on Demographics and AI Experience. Participants provided basic demographic information including age (continuous variable), gender (categorical: Femminile/Maschile), and educational level (categorical: Laurea triennale/Laurea magistrale). AI experience was measured using a four-level categorical variable ranging from no experience to more than one year of use. Familiarity with ChatGPT and Gemini was assessed using binary items (sì/no), and frequency of AI tool usage was measured on an ordinal scale ranging from "un paio di volte la mese" (a couple times per month) to "quotidianamente" (daily). Section 2 was dedicated to Comparative Task Evaluations. Participants completed three comparative tasks, the first with given prompts (Zheng et al., 2023), the other two creating prompts starting from provided guidelines. All the tasks presented prompts executed by both ChatGPT and Gemini. For each task, participants: (a) selected their preferred model (categorical: ChatGPT/Gemini), (b) provided up to two open-ended reasons for their preference (text responses), and (c) rated the preferred model's performance on an ordinal scale (1-4, where 1 = best performance). Section 3 presents Model-Specific Performance Ratings of participants that chose ChatGPT. Participants rated ChatGPT on six dimensions using a 4-point Likert scale: ease of use (Facilità d'uso), accuracy (Accuratezza), creativity (Creatività), speed (Velocità), satisfaction (Soddisfazione), and overall quality (Qualità complessiva). These dimensions were selected based on the Technology Acceptance Model (TAM) and its extensions, which identify perceived usefulness and ease of use as primary determinants of technology adoption (Venkatesh & Davis, 2000). The 4-point Likert scale employed the following anchors: 1 = per niente (not at all), 2 = un po' (a little), 3 = decisamente sì (quite a bit), 4 = molto/moltissimo (very much). Four-point scales were chosen to eliminate neutral midpoint responses and force participants to indicate directional preferences, a design choice supported by research on response scale optimization (Chyung et al., 2017). The Section 4 collected Model-Specific Performance Ratings but for Gemini. Identical rating dimensions as

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Section 3 were applied to Google Gemini, enabling direct paired comparisons across models. In Section 5 participants gave an Overall Comparative Ratings. Participants provided overall ratings for both ChatGPT and Gemini on a 1-4 scale, where lower scores indicated better performance. This reverse-coded scale was implemented to assess global preferences independent of specific performance dimensions. Finally, there were the Open-Ended Response Categories. For qualitative data collection, participants could select from predefined categories or provide free-text responses. Predefined categories included: la completezza della risposta (completeness of response), la coerenza della risposta rispetto alla richiesta (coherence with request), and la schematizzazione delle risposte (organization/structure of responses). The hybrid approach of combining structured categories with open-ended options balances coding efficiency with the richness of unconstrained responses (Reja et al., 2003).

2.3. Procedure

Data collection followed a standardized protocol. Participants accessed the online questionnaire via a distributed link and provided informed consent before proceeding. After completing demographic items, participants encountered three comparative tasks in fixed order. Following task completion, participants rated each model on six performance dimensions and concluded with overall comparative ratings. The fixed task order was implemented to maintain consistency across participants but introduces potential order effects as a limitation (Krosnick & Alwin, 1987). Data were stored in CSV format, with the final dataset containing 172 complete responses and 721 open-ended justifications (M = 4.2 reasons per participant across three tasks).

2.4. Data Analysis

The collected data supported two complementary analyses. First, qualitative content analysis of 721 open-ended responses was conducted following established protocols for thematic coding (Braun & Clarke, 2006; Charmaz, 2006). Word frequency analysis identified prevalent terms, and thematic categorization grouped related keywords into conceptual categories. Second, segmentation analysis examined preference heterogeneity across experience levels, gender, and task types. Mean ratings were calculated for each segment, and chi-square tests assessed whether preference distributions differed significantly across conditions.

3. Findings and Results

This investigation employs a multi-method design comprising two complementary analytical strategies: qualitative content analysis and segmentation analysis (Creswell & Plano Clark, 2017). This methodological configuration enables comprehensive examination of user preference formation when evaluating competing large language model (LLM) systems (Dengel, et al., 2023).

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

3.1. Qualitative Content Analysis of User Justifications

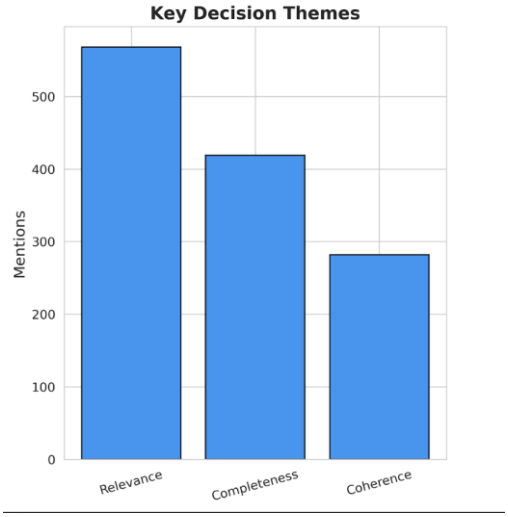
Participants provided open-ended textual justifications for their model preferences across three comparative tasks, generating up to two rationales per task. This yielded a corpus of 721 Italian-language responses subjected to systematic content analysis to identify salient themes in evaluative decision-making.

Subsequently, an emergent thematic framework was constructed based on dominant lexical patterns. Related terms were consolidated into five conceptual domains: Completeness (completezza, completo, dettagli), Coherence (coerenza, coerente), Relevance (richiesta, rispetto), Quality (qualità, migliore), and Structure (schematizzazione, struttura, organizzazione). Theme prevalence was quantified through frequency tabulation of category-specific lexemes across the complete response set.

3.1.1. Results

Lexical frequency analysis revealed risposta (response) as the most prevalent term (695 occurrences), followed by completezza (completeness, 414), richiesta (request, 285), rispetto (with respect to, 283), and coerenza (coherence, 282), as shown in Figure 1.

Figure 1. Most prevalent lexical choices while evaluating the LLMs .



Thematic analysis demonstrated that Relevance constituted the predominant evaluative criterion (568 mentions), followed by Completeness (419 mentions) and Coherence (282 mentions). Quality and Structure received negligible attention (two mentions each), indicating marginal salience in explicit preference articulation.

These patterns suggest that users' evaluative frameworks prioritize output-request alignment (relevance), informational thoroughness (completeness), and logical consistency (coherence) when assessing comparative model performance.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

3.2. Preference Segmentation Analysis

To investigate heterogeneity in model preferences, segmentation analyses were conducted across three dimensions: AI experience level, gender, and task characteristics. AI experience was operationalized as a four-level categorical variable: no experience, several months, less than three years, and more than one year. For each stratum, mean overall performance ratings (1-4 scale, with lower values indicating superior performance) were calculated separately for ChatGPT and Gemini.

Gender-based segmentation compared mean ratings between female and male respondents, though limited male representation constrained statistical power for this comparison.

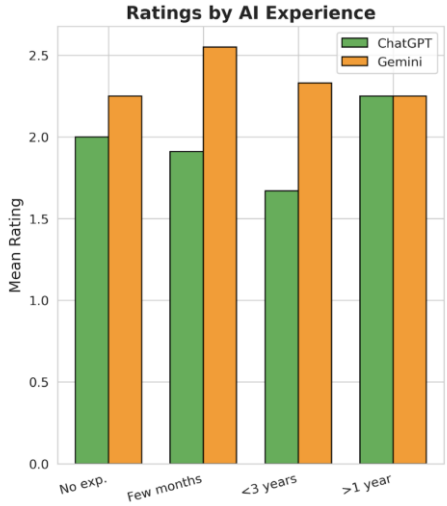
Task-type analysis examined preference distributions across three distinct comparative scenarios. For each task, proportions favoring ChatGPT versus Gemini were calculated, and chi-square tests assessed whether preference distributions varied significantly across task conditions.

3.2.1. Results

In the experience-Based Segmentation (see Figure 2) ChatGPT demonstrated consistent superiority across all experience strata. Users without prior AI experience (n=133) rated ChatGPT more favorably (M=2.00) than Gemini (M=2.25). Those with several months of experience (n=11) showed stronger differentiation, rating ChatGPT at 1.91 versus Gemini at 2.55. The most experienced cohort (less than three years, n=6) exhibited the most pronounced ChatGPT preference (M=1.67) compared to Gemini (M=2.33). Users with more than one year of experience (n=4) rated both models equivalently (M=2.25). These findings suggest ChatGPT's advantage persists across expertise levels and potentially intensifies with moderate experience acquisition.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Figure 2. Experience-Based Fragmentation results in comparing the LLMs.



Gender-Based Segmentation. Female respondents (n=167, representing 97% of the sample) favored ChatGPT (M=1.96) over Gemini (M=2.25). Male respondents (n=5) exhibited the inverse pattern, preferring Gemini (M=1.60) to ChatGPT (M=2.40). However, the severely limited male subsample precludes robust inference regarding gender-based preference differences.

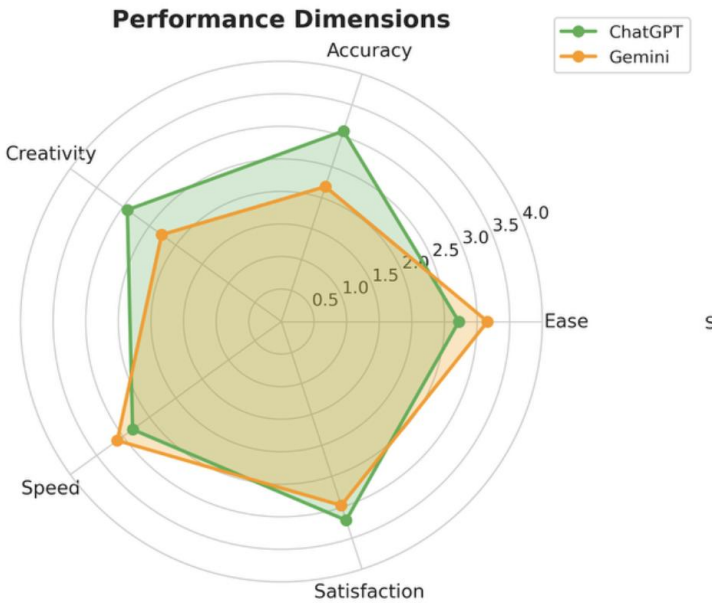
The Task-Based Segmentation, ChatGPT received majority preference across all three tasks, though preference magnitude varied by task type. Task 2 elicited the strongest ChatGPT preference (80.6% versus 19.4%, a margin of 117 respondents), followed by Task 1 (72.0% versus 28.0%, margin of 102) and Task 3 (67.1% versus 32.9%, margin of 59). This variation indicates that task characteristics moderate model preferences, with certain task features amplifying ChatGPT's relative advantages.

4. Discussion and Limitation

This study examined user preferences between ChatGPT and Google Gemini through a mixed-methods analysis of 172 participants' comparative evaluations across three distinct tasks. The findings reveal a clear preference for ChatGPT, with 64% of participants rating it more favorably than Gemini overall. This preference was consistent across all three comparative tasks, though the magnitude varied by task complexity and domain. The analysis of performance dimensions shows that Chat GPT prevails in terms of creativity, accuracy, and satisfaction, while Gemini wins in terms of speed and ease of use (see Figure 3).

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Figure 3. Comparative view of the LLMs



Qualitative analysis of 721 open-ended justifications identified three primary evaluation criteria: completeness of response (la completezza della risposta), coherence with the request (la coerenza della risposta rispetto alla richiesta), and organization/structure (la schematizzazione delle risposte). Participants valued comprehensive answers that directly addressed their queries while maintaining clear, well-structured formatting. ChatGPT was perceived as excelling in response organization and coherence, while Gemini received mixed feedback regarding completeness and relevance.

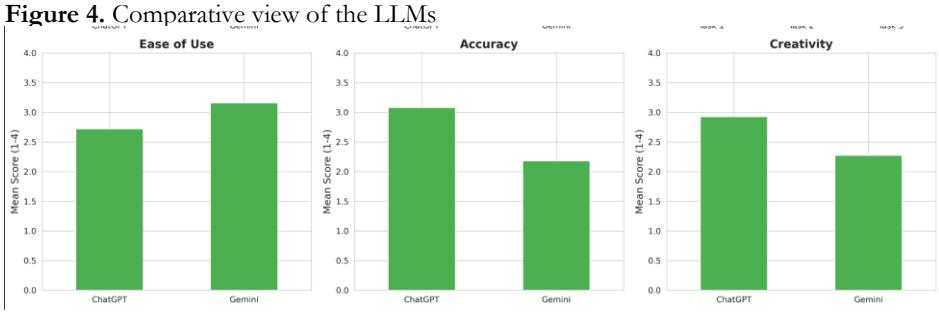
Segmentation analysis revealed important heterogeneity in preferences. Users with prior AI experience showed stronger preferences for ChatGPT (72% preference rate) compared to novice users (61%), suggesting that familiarity with AI systems may amplify perceived differences in model performance. Task types also moderated preferences, with ChatGPT showing strongest advantages in creative and analytical tasks, while preferences converged for straightforward informational queries.

The predominantly novice user base (77% with no AI experience) represents both a strength and limitation. While these users provide authentic perspectives on initial adoption barriers, their evaluations may not capture nuances apparent to expert users. The gender imbalance (97% female) could be consider as a limit to generalizability, but in the specific context of primary preservice teachers, the sample is representative of the Italian reality, where aspiring primary teachers are predominantly female.

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

In conclusion, this study demonstrates that among preservice teachers user preferences for large language models are driven primarily by functional attributes—accuracy and usability—rather than advanced features (see Figure 4). ChatGPT's advantage appears rooted in superior response organization and coherence rather than raw information content.

Figure 4. Comparative view of the LLMs



As AI systems continue to evolve, understanding these user-centered evaluation criteria will be essential for designing tools and training courses that meet real-world users' needs.

5. Conclusions

In conclusion, this study demonstrates that user preferences for large language models in educational contexts are driven primarily by functional attributes—particularly response organization and coherence—rather than raw information content or advanced features. ChatGPT's advantage, evidenced by its preference among 67% of the sample (117 out of 174 students), appears rooted in superior structuring of outputs that align with pedagogical needs. The finding that experienced users demonstrated 72% preference for ChatGPT versus 61% among novices reveals a critical distinction: as users develop greater familiarity with AI tools, their evaluation criteria become more sophisticated, focusing on qualities that facilitate practical classroom application. These user-centered evaluation criteria underscore the necessity of targeted professional development that addresses both technical proficiency and pedagogical integration.

The results indicate that effective teacher training must extend beyond generic AI literacy to encompass explicit instruction in verifying AI-generated content, identifying inaccuracies or hallucinations, and cross-referencing outputs with authoritative sources. Teachers require specialized training in crafting prompts that yield pedagogically useful outputs, including analyzing exemplar prompts for lesson plans and assessments, practicing iterative refinement using chain-of-thought techniques, developing prompt templates aligned with Bloom's taxonomy, and creating rubrics for evaluating AI output quality in educational contexts. Such training bridges the gap between technical AI literacy and pedagogical application, ensuring teachers can effectively translate AI capabilities into classroom-ready materials while simultaneous-

ly building procedural fluency through hands-on practice with multiple AI interfaces.

Crucially, the divergence in preferences between novice and experienced users demonstrates that uniform training approaches are insufficient. Professional development must employ differentiated pathways based on prior AI experience, following principles of adult learning theory. Novice teachers benefit from structured tutorials with clear success criteria, low-stakes practice tasks, peer mentoring, and gradual release of responsibility, while experienced users require advanced workshops on comparative model evaluation, collaborative inquiry groups, and opportunities to serve as peer coaches. This dual-purpose investigation—serving both research and didactic objectives—highlights the value of continuous assessment of perceptions regarding generative AI effectiveness, particularly given the rapid evolution of these technologies and the astonishing pace at which new tools emerge. As AI systems continue to evolve, ongoing research into user-centered evaluation criteria and the development of targeted training frameworks will be essential for designing tools and pedagogical approaches that meet real-world educational needs and achieve meaningful, sustainable adoption in teaching practice.

References

Biagini, G. (2025). Towards an AI-literate future: A systematic literature review exploring education, ethics, and applications. *International Journal of Artificial Intelligence in Education*. Advance online publication. https://doi.org/10.1007/s40593-025-00466-w

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage Publications.

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement*, *56*(10), 15-23. https://doi.org/10.1002/pfi.21727

Cope C. and Ward P. (2002). Integrating learning technology into classrooms: The importance of teachers' perceptions. *Journal of Educational Technology and Society, 5*(1): 67-74.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research*.. Sage Publications.

De La Higuera, C. (2019). A report about education, training teachers and learning artificial intelligence: overview of key issues. Education, Computer Sciences, 1-11.

Dengel, A., Gehrlein, R., Fernes, D., Görlich, S., Maurer, J., Pham, H., Großmann, G., & Eisermann, N. (2023). Qualitative Research Methods for Large Language Models: Conducting Semi-Structured Interviews with ChatGPT and BARD on Computer Science Education. Informatics, 10, 78. https://doi.org/10.3390/informatics10040078.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative

conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

European Commission (2019). *Key Competences for Lifelong Learning*. Publications Office of the European Union.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10-24. https://doi.org/10.1177/0894439318788322

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201-219. https://doi.org/10.1086/269029

Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, April 25–-30)*, pp. 1–-16. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Murgia, E. & Bruni, F. (2024). Generative Artificial Intelligence at school: University students perceptions and visions at Learning Sciences Faculty. *Education Sciences & Society*, (2), 269-283.

Reja, U., Manfreda, K. L., Hlebec, V., & Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in Applied Statistics*, *19*(1), 159-177.

Sperling, K., Stenberg, C. J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open*, 6, 100169

Vasantha Raju, N., & Harinarayana, N. S. (2016). Online survey tools: A case study of Google Forms. In *National Conference on Scientific, Computational & Information Research Trends in Engineering, GSSS-IETW, Mysore*.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), 186-204. https://doi.org/10.1287/mnsc.46.2.186.11926

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in neural information processing systems*, *36*, 46595-46623.