EVALUATING GPT-4's PROFICIENCY ON NORWEGIAN EXAMS AND TESTS—AND EXPLORING THE BROADER IMPLICATIONS FOR EDUCATIONAL PRACTICE

Rune Krumsvik and Lise Jones
University of Bergen
NORWAY

Abstract

A Meta-analyses show Intelligent Tutoring Systems excel one-to-one, yet none support Norwegian. This position paper treats GPT-4—though not designed as an ITS—as a candidate tutor, testing it on Norwegian exams in medicine, nursing, psychology, dentistry, military theory, driving, university entrance, citizenship and maths teaching, plus IQ, social and multimodal medical-image tasks. GPT-4 averaged 94.3% accuracy, handled descriptive-procedural questions, adapted to Sami, and surpassed conventional ITSs in linguistic and cultural flexibility. Findings reveal broad multilinguistic, cognitive and multimodal strengths with significant implications for formative and summative assessment across education.

Keywords: AI; GPT-4; Multilingualism; Exams, Tests, Norwegian; Performance
Introduction

Introduction

With the launch of language models such as XLNet, BERT, ChatGPT, GPT-4, Gemini Advanced, Claude, we are facing a technological paradigm shift that may also influence how we perceive Intelligent Tutoring Systems (ITS) and related areas in the future. Since the 1970s, research has explored ITS and the potential of AI to provide personalized tutoring, inspired by Bloom's (1984) well-known 2-sigma finding. Numerous meta-analyses comparing traditional teaching methods with ITS have found that, under certain conditions, ITS can effectively provide one-on-one tutoring. However, existing ITSs do not support the Norwegian language. While large language models like GPT-4, Gemini Advanced, and Claude are not specifically designed as ITS, they share significant similarities. The knowledge base shows that these models have the potential to address certain educational challenges in both the education and healthcare sectors, opening up a broader discussion about their role in the future of education. GPT-4, an advanced language model developed by OpenAI, has proven capable in various English-speaking academic fields, exams, and tests (Ray, 2023; Agarwal et al., 2023; Brin et al., 2023; Brodeur et al., 2023; Deng et al., 2025; Goh et al., 2024; Hirunyasiri et al.,

2023; Jin et al., 2024; Karthikesalingam and Natarajan, 2024; Kim et al. 2020; Liu et al, 2024; McDuff et al., 2025; Nori et al., 2023; Rajpurkar et al., 2020; Phung et al., 2023). However, the current state of knowledge lacks studies on how it performs in Norwegian in Norwegian-language exam and test contexts. This position paper, based on a case study, aims to evaluate how GPT-4 manages multilingual challenges with Norwegian as an exam/test language, focusing on descriptive-procedural questions and various exam and test contexts both within and outside academia. These contexts include exams in medicine, nursing, psychology, dentistry, a military theory test, the Norwegian driving test, the Swedish university entrance exam (SweSAT), the Norwegian citizenship test, and a national teacher exam in mathematics. A primary objective is to assess the reliability and generalizability of this type of AI in academic settings and in the Norwegian language. More specifically it focuses on whether GPT-4 is capable of answering various exams and tests that are primarily given in Norwegian in Norwegian educational and societal contexts, how reliable it is, and what implications this might have for both multilingualism summative and formative assessment elements within and outside academia.

This abovementioned literature review and our former studies (Krumsvik, 2024, 2025a, 2025b, 2025c) generated a number of explorative questions and reflections around this topic: How effectively can GPT-4 handle multilingual challenges, particularly in Norwegian, across both academic exam tasks and general IQ and social tests? What is its precision rate, and how reliable is its performance in these contexts? Does GPT-4's ability to manage descriptive and procedural questions align with international findings, and how well does it adapt to Norwegian exam and test contexts? Can it demonstrate linguistic and cultural adaptability that conventional ITSs lack? Furthermore, what are the limitations of GPT-4 in these contexts, and how does a case study contribute to our understanding of its multilingual and cognitive capabilities in summative assessments? Can GPT-4 serve as an effective tool in formative assessment contexts, and how well does the case study design perform in research environments characterized by rapid development? These preliminary questions and reflections can be summarized in one main research question we will examine in this position paper:

How capable is GPT-4 of answering selected Norwegian exams and tests, and what potential implications might this have for formative and summative assessment in educational sciences and healthcare education?

Methodology

This position paper is based on a case study which is exploratory and intrinsic (Stake, 1995, 2006). I conducted a cumulative data collection and analysis process (Creswell & Guetterman, 2021), based on performance on nine exams and tests

inside and outside academia in Norway where I applied chain-of-thought prompting with the exact same wording as in the exam and test text in Norwegian (Figure 1).

Figure 1

The Research Process of the Intrinsic Case Study

Chain-of-thought prompting is an approach in which a user explicitly asks a language model to reveal its step-by-step reasoning process before giving the final answer, improving transparency and often boosting solution quality for complex tasks. Furthermore, in the supplementary data collection (blue arrows in second and third position in the figure), I integrated the research questions into the dialogue (A) of the results from 1, and interacted with GPT-4 around the preliminary findings. Finally, further digital fieldwork (B) was conducted to check for possible biases and misinterpretations, ongoing fine-tuning of GPT-4, as well as in light of the aforementioned current state of knowledge on this topic.

The main test period was carried out from March 25, 2023, to August 5, 2023, and the exams and tests were from different areas both inside and outside academia to check GPT-4s ability to handle different contexts. Four of the exams were fullscale exams, while the five other exams and tests were based on random selection (two sub-tasks in one test had to be omitted due to task drawings that GPT-4 could not perceive and "see"). All the exams and tests were in the Norwegian language (except the Swedish SweSAT) and consisted mainly of text questions. Scoring of all the exams was based on the grading guidelines (sensorveiledning) derived from different sources. Interaction with GPT-4 was conducted based on the questions in nine exams and tests, which were posed to GPT-4 using chain-of-thought prompting, and responses were recorded (each response was considered final).

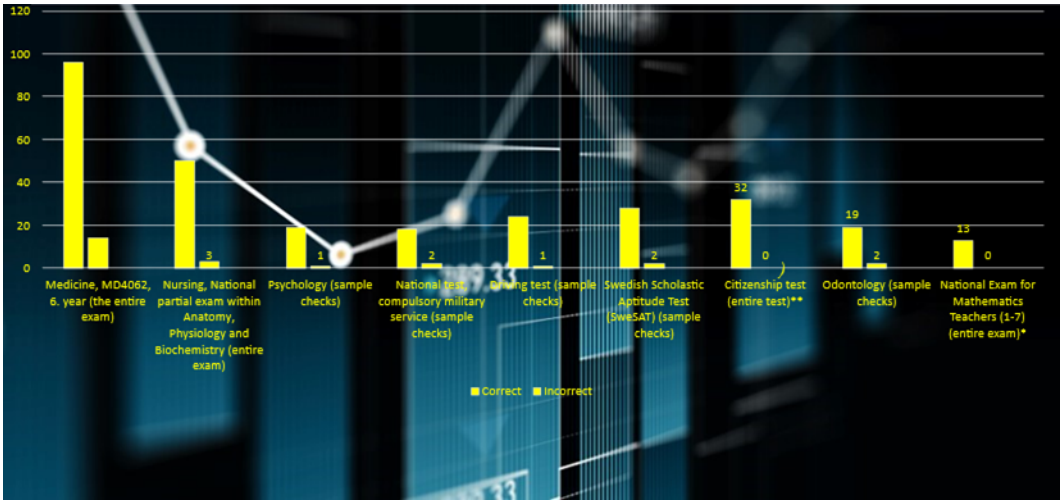
Data analysis, step 1, was based on GPT-4's performance on the nine exams and tests selected randomly from previous exam sets. The supplemental data was collected from August 2023 to April 2024 and consisted of comprehensive interactions with GPT-4 and digital fieldwork (described above).

Results

The tests were conducted from March 20 to August 10, 2023. Figure 2 illustrates the number of questions in each of the nine exams and tests.

Figure 2

Number of Questions in Each of the Nine Exams and Tests



Note. Sample checks: When only sample checks of the exam/test were performed. Entire exam: When a test of the entire exam/test was conducted.

* Two sub-tasks had to be omitted due to a task drawing that GPT-4 cannot see (thus, 13 out of 15 sub-tasks were completed).

** This test currently consists of 36 questions, but the version publicly available and tested consisted of 32 questions.

Table 1 shows that the average precision rate of 94.26% indicates that GPT-4 performs very well across various fields within and outside academia, spanning a relatively broad range of exams, tasks, tests, and domains. It demonstrates multilingual and cognitive skills a high level and GPT-4 generally has capabilities comparable to the human level in such exam and test contexts. While all nine exams/tests have an element of descriptive knowledge (knowing that), the medical exam includes a number of exam tasks that lean towards procedural knowledge (knowing how) (Anderson, 2005) as they are formulated as patient cases (and not factual knowledge per se). Additionally, about 10 percent of the 110 exam tasks contain image illustrations related to the tasks (X-rays, images of skin rashes, organ images, etc.), which are helpful for students in addition to the task text itself (which often small case descriptions about patients). This multimodality could not be "seen" or interpreted by GPT-4 in spring 2023, and thus, for these tasks, it could only respond based on text descriptions. Nevertheless, we see that GPT-4 achieves 87.3% correct answers (96 out of 110) on this exam, and when looking at the detailed and reasoned responses it provides, this shows a good academic level. Below I present a dialogue (A) with GPT-4 regarding these results.

Table 1

GPT-4's Performance on Exams and Tests inside and outside Academia in Norway

Field	Correct (%)	Incorrect (%)
Medicine (entire exam)	87.3	12.7
Nursing (entire exam)	96.2	3.8
Psychology (sample checks)	95	5
Military Conscription (IQ-test) (sample checks)	90	10
Driving Test (Car) (sample checks)	96	4
Swedish Scholastic Aptitude Test (SweSAT) (sample checks)	93.3	6.7
Citizenship Test (entire test)	100	0
Dentistry (sample checks)	90.5	9.5
Teacher education (Mathematics) (entire exam)	100	0
Average Precision Rate	**94.26**	

Summary of Phases A and B

Overall, GPT-4 shows advanced capabilities in understanding and generating accurate responses across diverse and complex tasks, particularly in Norwegian-language exams and medical image analysis.

The results indicate that in the medical exam, descriptive and procedural knowledge are in a dialectical relationship as GPT-4 cannot answer the exam questions without possessing both types of surface and deep knowledge. Exams in nursing, psychology, dentistry, the Swedish Scholastic Aptitude Test (SweSAT), and teacher education also exhibit this combination to some extent. Other tests, like the military conscription test (IQtest), also include this combination but are more oriented towards general knowledge. It can be added that a smaller version than GPT-4, GPT-3, managed 73 out of 80 tasks in the SweSAT in 2022 (Svensson, 2022), and many were surprised at how well it handled abstract metaphors. The driving test and citizenship test primarily assess descriptive knowledge. From this, we can see that the GPT-4's descriptive and procedural abilities can also be related to Anna Sfard's (1998) two metaphors for learning (acquisition metaphor and participation metaphor) but in a more situated context within school or academia, not limited to a specific exam or test situation inside and outside academia.

Overall, GPT-4s scores across the nine different exams/tests demonstrate its ability to handle multilinguist and relatively complex Norwegian-language questions, at times at a high academic level. Additionally, phases A and B show that it also handles multimodal image analysis very well. This suggests a need for a broader epistemological discussion about new forms of AI-generated communities of practice (CoP and whether GPT-4 can be considered a highly capable dialogue partner and tutor for Norwegian-speaking students preparing for this form of summative assessment (school exams). These findings are consistent with our tentative knowledge summaries and case studies (Krumsvik, 2024, 2025a, 2025b, 2025c), which find a similar trend across various English-language exams/tests internationally (Ray, 2023).

Discussion

The results from testing GPT-4 on various Norwegian-language exams and tests on multilingual and cognitive capabilities in such contexts, aligns with our previous findings (Krumsvik, 2024, 2025a, 2025b, 2025c). GPT-4s cognitive capabilities also align with the pre-print from Bubeck et al. (2023), Ray (2023) and Deng et al. (2025). Particularly notable in this study is GPT-4's multilinguistic abilities and performance on Norwegian-language exams, despite the model primarily being trained on English-language data. This indicates a good ability to generalize knowledge across languages. Exams such as medicine, nursing, and psychology include both descriptive and procedural knowledge, requiring deeper understanding and processing. GPT-4's ability to handle such complex tasks suggests that the model can go beyond mere memorization of facts and engage in more sophisticated cognitive processing. Such findings are supported within ITS by VanLehn (2011), who emphasizes the importance of deep learning in effective tutoring systems. VanLehn (2011) points out that human tutoring has an effect size of d = 0.79, while ITSs show a similar effect size of d = 0.76 in his study, and in Tlili et al. (2025) metanalysis this is g=1.07.

The results of our study suggest that GPT-4, as part of an ITS, can offer a comparable level of support as human tutors. This is especially relevant in light of previous research showing that traditional classroom instruction often does not reach the same level of effectiveness as one-on-one tutoring (Bloom, 1984). At the same time, it is important to note that, according to Ma et al. (2014), there are some distinctive features of ITS (as mentioned earlier) that GPT-4 does not inherently possess. These are especially oriented towards calculating inferences from student responses, constructing multidimensional models of the student's learning status, and placing the student's current learning status in a multidimensional domain model. This can be partially achieved by establishing a domain-specific "chatbot within the chatbot" by integrating a training basis on top of GPT-4 and simultaneously embedding a script in this chatbot, tuning it more specifically

towards having an ITS-related functionality (along with established ITSs). Integrating GPT-4 in ITS-related areas can potentially expand tutoring opportunities in the educational sector. With GPT-4's ability to generate educational content, analyze student input, and offer real-time feedback, GPT-4 can significantly enhance tutoring opportunities and AI-CoP both for students who have Norwegian as their native language, but also for foreign students who may interact with GPT-4 in English about Norwegian-language exam questions, tests, etc. This can be a good sparring partner for such language thresholds and be a valuable supplement for foreign students in addition to other measures and conventional tutors in higher education.

Conclusion and Implications

The research question in position paper focused on how capable GPT-4 is of answering exams and tests in Norwegian and what implications could this have for education. Despite some limitations, the position paper confirms that GPT-4 has significant multilinguistic and cognitive capabilities, making it a valuable tool both inside and outside academia as a sparring partner in various exam and test contexts. With an average precision rate of 94.26%, the model demonstrates the ability to answer both more factual questions and more complex and varied questions in a manner comparable to human performance in such contexts. Given that GPT-4 also masters Norwegian well, it is particularly relevant in the nine exam and test contexts studied, which are primarily in Norwegian, targeting the Norwegian educational and societal context (the exception being the SweSAT, where GPT-4 also shows strong proficiency in Swedish). GPT-4's good performance in Norwegian as early as 2023 is probably because, e.g. medical knowledge is globally standardized and largely overlaps with the English-language material the model was trained on. It handles written Norwegian and technical terminology well, and many exam tasks require pattern recognition rather than deep reasoning—an area where GPT-4 excels. The exams were highly standardized and not dependent on Norwegian-specific legal or cultural context, allowing the model to apply its global knowledge base effectively in Norwegian. This suggests that large language models can play a complementary role in various tutoring contexts in higher education, in developing AI-CoPs, and in the further development of ITS.

In summary, such language models are intellectual artifacts mastering contextual language games (Wittgenstein, 1997), representing a significant leap from earlier language models, ITS, and chatbots. But it requires mastery of the granularity in prompts based on "chain of thought" prompting, which often shows variability among students, citizens, and learners. This illustrates that such digital competence will become increasingly important both inside and outside academia in the coming years to fully exploit the potential of such language models for both formative and summative assessment contexts in the future. With such reservations, one

implications of the findings might be that non-Norwegian-speaking students, university staff, and citizens in general who wish to learn Norwegian now have a highly competent "Norwegian teacher" by their side with GPT-4. However, AI is at the same time an ethical minefield and the study also underscores the need for vigilance and careful implementation to mitigate biases and ethical issues associated with AI use in education.

Methodological Considerations – Strengths and Weaknesses

As previously mentioned, the case study underpinning this position paper is based on pre-testing of GPT-4 in the absence of actual student participation. As such, it carries several limitations, yet also demonstrates certain strengths. Given that the AI field is a "moving target" developing very rapidly, case studies with triangulation, cumulativeness, and the possibility for retesting over a year can be an effective design in such research settings. The ability to track progress over an entire year provides valuable insights into the model's ability to adapt and improve over time. Although this case study has several strengths, there are also important limitations that need to be considered related to the methodological choices made. For instance, a full-scale testing of all nine exams and test areas was not conducted. Instead, a series of sample checks were carried out in this case study in five of the test areas, and such sample checks have several limitations that should be noted. The tasks were not translated to English and kept in Norwegian. Additionally, although the study followed the development over one year, this may still be a relatively short period to fully understand the long-term effects and improvements in AI models. Longer follow-up periods could provide more comprehensive insights. The results from the case study cannot be generalized and derive their strength from the depth perspective. The selected sample also has its biases. These limitations highlight the need for caution in interpreting the results and the importance of further research to validate and extend the findings.

Declaration of AI Use

This article explores the use of GPT-4, and artificial intelligence (AI) is therefore the object of study in this case-based research. As such, GPT-4 has been used in various ways throughout the research process, including pre-testing, documentation, and analytical reflection. However, all parts of the article—including the structure, argumentation, interpretation of findings, and final wording—have been designed, authored, and critically reviewed by the authors themselves.

References

Anderson, J. R. (2005). *Cognitive Psychology and Its Implications* (6th ed.). Worth Publishers.

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). *Combining human expertise with artificial intelligence: Experimental evidence from radiology (Working Paper No. 31422).* National Bureau of Economic Research. https://www.nber.org/papers/w31422

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4-16. https://doi.org/10.3102/0013189X013006004

Brin, D., Sorin, V., Vaid, A., Soroush, A., Glicksberg, B. S., Charney, A. W., Nadkarni, G., & Klang, E. (2023). Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports*, *13*, 16492. https://doi.org/10.1038/s41598-023-43436-9

Brodeur, P. G., Buckley, T. A., Kanjee, Z., Goh, E., Ling, E. B., Jain, P., Cabral, S., Abdulnour, R.-E., Haimovich, A. D., Freed, J. A., Olson, A., Morgan, D. J., Hom, J., Gallo, R., McCoy, L. G., Horvitz, E., Mombini, H., Lucas, C., Fotoohi, M.,…Rodman, A. (2024). Superhuman performance of a large language model on the reasoning tasks of a physician. *ArXiv,* 2412.10849. *https://arxiv.org/abs/2412.10849*

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *ArXiv,* 2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Creswell, J. W., & Guetterman, T. C. (2021). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (6th ed.). Pearson.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons,A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *Journal of the American Medical Association (JAMA) Network Open, 7*(10*),* Article e2440969. doi:10.1001/jamanetworkopen.2024.40969

Hirunyasiri, D., Thomas, D. R., Lin, J., Koedinger, K. R., & Aleven, V. (2023, July 7). Comparative analysis of GPT-4 and human graders in evaluating praise given to students in synthetic dialogues [Conference paper]. In *Proceedings of the AIED 2023 Workshop: Towards the Future of AI-Augmented Human Tutoring in Math Learning* (CEUR Workshop Proceedings, Vol. 3491, pp. 1–12). CEUR-WS.org. https://ceur-ws.org/Vol-3491/paper4.pdf

Jin, H. K., Lee, H. . & Kim, E. (2024). Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis. *BMC Med Educ 24*, 1013. https://doi.org/10.1186/s12909-024-05944-8

Karthikesalingam, A. & Natarajan, V. (2024, January 12). *AMIE: A research AI system for diagnostic medical reasoning and conversations*. Google Research. https://blog.research.google/2024/01/amie-research-ai-system-for-diagnostic_12.html

Kim, H.-E., Kim, H. H., Han, B.-K., Kim, K. H., Han, K., Nam, H., Lee, E. H., & Kim, E.-K. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. *The Lancet Digital Health*, 2(3), e138–e148. https://doi.org/10.1016/S2589-7500(20)30003-0

Krumsvik, R. J. (2024). Artificial intelligence in nurse education – a new sparring partner? GPT-4 capabilities of formative and summative assessment in National Examination in Anatomy, Physiology, and Biochemistry. *Nordic Journal of Digital Literacy*, *19*(3), 172–186. https://doi.org/10.18261/njdl.19.3.5

Krumsvik, R. J. (2025a). GPT-4's capabilities in handling essay-based exams in Norwegian: An intrinsic case study from the early phase of intervention. *Frontiers in Education, 10,* 1444544. https://doi.org/10.3389/feduc.2025.1444544

Krumsvik, R. J. (2025b). Chatbots and academic writing for doctoral students. *Education and Information Technologies, 30*, 9427–9461. https://doi.org/10.1007/s10639-024-13177-x

Krumsvik, R. J. (2025c). GPT-4's capabilities for formative and summative assessments in Norwegian medicine exams—an intrinsic case study in the early phase of intervention. *Frontiers in Medicine, 12*, 1441747. https://doi.org/10.3389/fmed.2025.1441747

Liu M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *J Journal of Medical Internet Research, 26,* e60807. doi: 10.2196/60807.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901-918. edu-a0037123.pdf

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., Hou, L., Cheng, Y., Liu, Y., Mahdavi, S. S., Prakash, S., Pathak, A., Semturs, C., Patel, S., Webster, D. R.,… Natarajan, V. (2025). Towards accurate differential diagnosis with large language models. *Nature, 642*, 451–457. https://doi.org/10.1038/s41586-025-08869-4

Svensson, E. (2022, June 27). Does GPT-3 know Swedish? *Medium*. https://medium.com/@emil_svensson/does-gpt-3-know-swedish-3e22bfd7b58f

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *ArXiv*, 2303.13375. https://doi.org/10.48550/arXiv.2303.13375

Phung, T., Pădurean, V.-A., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., & Soares, G. (2023). Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume* 2 (pp. 41-42). https://doi.org/10.1145/3568812.3603476

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Lee, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., Griesel, R., Ng, A. Y., Boyles, T. H., & Lungren, M. P. (2020). CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digital Medicine, 3*, 115. https://doi.org/10.1038/s41746-020-00322-2

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, *27*(2) 4-13. https://doi.org/10.3102/0013189X027002004

Stake, R. E. (1995). *The art of case study research*. Sage.

Stake, R. E. (2006). *Multi case study analy*sis. Guilford Press.

Tlili, A., Saqer, K., Salha, S., & Huang, R. (2025). Investigating the effect of artificial intelligence in education (AIEd) on learning achievement: A

meta-analysis and research synthesis. *Information Development*. https://doi.org/10.1177/02666669241304407

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221. https://doi.org/10.1080/00461520.2011.611369

Wittgenstein, L. (1997). *Filosofiske undersøkelser* (M. B. Tin, Trans.). Pax.

Author Details

Rune Krumsvik
University of Bergen
NORWAY
Rune.Johan.Krumsvik@uib.no

Lise Jones
University of Bergen
NORWAY