Large Language Models as Educational Evaluators: Reasoning and Explainability in Low-Resource Language Assessment

Ekhi Azurmendi

HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract

With large language models, significant advances have been made in various language processing tasks, such as sentiment analysis or automatic translation. However, there are tasks that still prove difficult for these models, and the automatic evaluation of written compositions is one of them. In this task, the model must evaluate a composition according to guidelines or criteria. Different evaluation systems have been attempted, but there is still much to investigate. This project aims to develop an automatic evaluation system for compositions, focused on the Basque language. The goal is to evaluate compositions following the system's guidelines and provide feedback on errors made. Additionally, the model should identify students' weaknesses and create exercises to address their deficiencies, contributing to the learning process. To develop this system, we will use advanced techniques to overcome the limitations of language models, hallucinations, and the generation of incorrect information, such as Retrieval Augmented Generation (RAG) or more general forms of text-conditioned learning. We will work on zero-shot or few-shot learning techniques to follow guidelines not observed during training, as well as efficient parameter adjustment methods, such as supervised fine-tuning (SFT) or reinforcement learning. We will also define and create synthetic data and auxiliary tasks to aid in the model's learning process. We will share with the scientific community the resources generated and conclusions obtained throughout the project.

Keywords

NLP, LLM, automatic essay review, reasoning, explainable AI,

1. Reason for the Proposed Research

The most recent and powerful Large Language Models (LLMs) are capable of evaluating, correcting, and suggesting improvements to user-generated texts. However, they remain susceptible to hallucinations and often struggle with nuanced linguistic reflection. This issue is especially pronounced in low-resource languages, where these models lack proficiency and the ability to understand deep linguistic features [1]. Although many studies aim to teach LLMs to follow predefined evaluation rubrics, they use closed models to create the datasets automatically and to evaluate their models [2]. Additionally, the development of tailored exercises to help writers improve their skills has been a focus of several research efforts using generative LLMs, but further research is needed to effectively adapt these tasks to new domains and languages with limited or no available training data.

The objective of this PhD project is to explore new methods for adapting LLMs to the educational domain, specifically for rubric-based evaluation and the generation of personalized writing exercises based on the learner's needs. The research will focus on low-resource languages, with a particular emphasis on Basque, where there is a significant lack of publicly available educational resources.

2. Background and Related Work

Thanks to LLMs, Natural Language Processing (NLP) has experienced unprecedented advancement [3, 4]. These models are trained on large amounts of text to learn language representation, and then transfer that knowledge to new contexts, trained with few manually annotated texts. These techniques

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

ekhi.azurmendi@ehu.eus (E. Azurmendi)

0009-0008-4113-890X (E. Azurmendi)

improve performance as the number of parameters increases [5, 6], as shown by results obtained in different benchmarks such as SuperPEG [7], MMLU [8] and the BIG-bench evaluation benchmarks [9].

Despite their unprecedented success, Large Language Models (LLMs) have significant limitations in following orders in special or highly specialized contexts [10]. An obvious limitation of LLMs is the phenomenon of hallucinations, where they provide erroneous information to users. Several efforts have been made to overcome these limitations: techniques such as Retrieval Augmented Generation (RAG) [11] or more general forms of textual conditioning learning [12], better known as prompt learning or in-context learning, offer great flexibility to adapt models to different domains and tasks. Conventional fine-tuning can also help address this problem, but training on constantly changing data is not a practical solution due to the high computational cost. Therefore, techniques such as the aforementioned RAG are very useful when developing applications for the real world [13].

Suitable recipes for effectively carrying out continuous learning to keep LLM skills up to date have not yet been clarified. Techniques such as RAG are appropriate for feeding models with information extracted from external documents, and facilitate bringing the capabilities of huge LLMs to smaller models [14, 15]. In this way, we can introduce new information not encoded in parameters without training, which reduces update costs. Moreover, some studies have shown that the evaluation of essays based on ideal references enhanced models' performance [2], so RAG techniques are suitable to find reference essays.

One current trend is the simultaneous application of strategies that combine textual conditioning learning with SFT. According to recent studies, LLMs have demonstrated generalization capability using hybrid techniques when following guidelines not observed in training, surpassing existing zero-shot capabilities, that is, improving models without learning examples [2, 16]. However, even with the spectacular results obtained, it is still unclear how knowledge obtained in one domain can be transferred to another specialized domain. For example, adaptation techniques to a specialized domain such as language learning have been little studied.

In addition to the previous techniques, it has been observed that using external tools to complete information that models don't have encoded is a powerful strategy [17, 18, 19]. There are numerous external tools: information retrieval, search engines, symbolic modules or code interpreters, for example. The use of these tools opens multiple research avenues, in which models can be interactively trained using reinforcement learning to adapt to these tools [20].

Although these lines of research generally aim to avoid hallucination and reasoning problems, they can be used to create effective, powerful, and dynamic applications and systems.

Another way to improve the skills of LLMs is to use reasoning strategies [21, 22**?**], so that more appropriate responses can be generated in exchange for more computational resources. Reasoning strategies are completed through chains of reasoning, that is, the problem is divided into several steps to facilitate its resolution. Previous studies have shown that LLMs are capable of simple reasoning [23], but have problems when performing complex reasoning.

For example, these models accurately respond to the birth and death dates of historical figures, but often have problems when asked about how old these people were when they died. To address these problems, reasoning chains are appropriate strategies, since by solving step by step, a more suitable final result can be obtained.

However, recent researchers have discovered new methods to improve the reasoning capabilities of these models by applying reinforcement learning techniques [24]. The latest open source models are able to outperform proprietary models in several benchmarks such as scientific question answering or language understanding [25].

Learning by text commands can be considered as a method that allows interaction between people and computers. Work done in recent years has used learning from commands written in natural language to guide computers towards different real-world tasks [26, 5].

In relation to the human-computer interaction environment, education is an important field of application where LLMs can have a great impact. The work carried out has shown that LLMs can be helpful in writing or reading in the educational environment [27, 28]. Innovative research has also been conducted [29, 10], using LLMs as aid in the classroom environment, collaborating in teacher-student

interaction, offering specialized teaching, or for automatic assessment of essays. The automatic creation of adapted exercises becomes increasingly interesting, due to the creative competencies of LLMs. The automatic creation of distractors, for example, has shown usable results in practice [30, 31]. Although the competencies of LLMs to create good exercises have improved greatly, it is not clear how models can be dynamically adapted to the specific needs of students, to improve exercises and feedback. The automatic creation of multiple-choice questions has taught us that we already have tools to put this into practice [30, 31]. Even taking into account the competencies of language models, it is still not clear how to adjust LLMs to create appropriate and high-quality exercises. The creation of these variable domain exercises adapted to students is important to provide the most appropriate help in the learning process.

With Basque as the focus, automatic assessment systems for texts written by students have been developed using traditional machine learning techniques [32]. These systems are based on the extraction of linguistic features that take into account the evaluation criteria, subsequently using a classifier to determine the level. However, traditional systems have shown problems adapting to new domains.

Using deep learning techniques, attempts have been made to improve the competencies of these systems [33]. The weak point of these latest systems lies in the reasoning of responses, as the model is not able to determine the errors or weaknesses identified through a textual description or give indications that help in the writing process.

Several attempts have been made to develop approaches to directly generate exercises from texts in Basque [34]. The authors aimed to create exercises from free texts using rules and traditional learning techniques, but in this approach the exercises are not reformed based on the user's errors, that is, the system is not dynamic.

3. Description of the Proposed Research, Including the Main Hypotheses for Research

Advances in NLP in the educational field have been achieved thanks to the surprising competencies that LLMs have demonstrated. Very evident improvements have been achieved in tasks such as automatic text evaluation, automatic exercise creation, or automatic text correction, among others. Despite these advances, systems created through LLMs have shown limitations: 1) Annotated data is needed to adjust the models, but the number of annotated texts in Basque in the educational field is low; 2) The adjustment of models has high computational costs and there are problems when dynamically adapting to new domains; 3) Despite the reasoning competencies of LLMs, there are still no adequate recipes for developing systems to adapt to student needs.

The objective of this project is to adapt LLMs to the educational domain to perform evaluations following guidelines, explain errors or improvement needs to the user through explanations, and create exercises or instructions dynamically adapting to the user's needs. This main objective can be divided into continued tasks or sub-objectives:

Add guideline-following capacity to LLMs. In this way, reasoning competencies of model evaluations would be developed and would be flexible to adapt to different evaluation criteria. It will be essential to translate and adapt to Basque the techniques used in the current state of the art. We will base on two main approaches:

- Use zero-shot or few-shot learning techniques, so the model has flexibility to adapt to new domains when little annotated data is available. We will use RAG, textual conditioning and command training methods to carry out this sub-objective.
- Study supervised learning methods by command, so that LLMs learn to follow domain instructions. Given the small amount of data, synthetic data or auxiliary tasks must be created to address this problem.

Develop LLMs that adapt to the needs of users. To meet this objective, the model must have the ability to plan, reason, explain and make appropriate comments. New Reinforcement Learning techniques, such as Direct Preference Optimization (DPO) [35] or Group Relative Policy Optimization (GRPO) [24]

will be taken into account, as well as other state-of-the-art techniques used in reasoning, including reasoning strategies.

Develop a model that, independent of the domain, is capable of making comments based on student errors and generating exercises. It is closely related to the previous objectives, as the model will need reasoning and planning capabilities to successfully complete this task. Appropriate evaluation methodologies and datasets must be created so that automatically created exercises are of good quality.

Use appropriate adjustment and training methods to reduce computation costs. Reducing the costs of adjustment, training, and use of LLMs is very important so that applications or real-world uses are as accessible as possible. To carry out this objective, PEFT-type methods will be used, to achieve competencies and the ability to follow orders with reduced costs. We will rely on techniques known as LORA, QLORA, or VeRA to meet this objective.

4. Methodology and the Proposed Experiments

This research project will use the research methodologies and functions presented below to carry out the aforementioned objectives.

The empirical method will be used; that is, the proposed hypotheses will be implemented in a system and evaluated using publicly accessible datasets. In this evaluation, we will make a comparison with systems available in the state of the art, validating the hypothesis when statistically significant improvements are obtained in said comparison.

The objectives we propose in this thesis are ambitious, and it is likely that not all proposed hypotheses will be fulfilled. For this reason, approaches will be tested one by one using the empirical method, and those with the greatest future projection will be explored in depth, leaving the rest aside.

The test banks and evaluation metrics and environments created and built throughout the project will be shared with the scientific community. In this way, we could not only collect comparable results, but also advice and improvements from the community. Although the results obtained may not be as expected, we will meet the set objectives and learn from the comments of the scientific community. We will disseminate the contributions made during the research at high-quality conferences with the community (ICML, IAAA, IJCAI, ICRL, ACL, EACL, NAACL, EMNLP, all SCIE Class 1 - Core A or A*).

Research Tasks (RT) and Research Questions (RQ):

RT0: Prepare the research environment. The first task is related to the preparation of the evaluation environment. The task will be precisely defined and publicly available datasets will be collected. Several works to follow guidelines have already been identified, but we need to check if there are new developments at the start of the thesis. Available LLMs will also be selected and initial experiments will be carried out to define an appropriate baseline. The main research questions in this section will be the following:

- **RQ0.A)** In the educational field, what are the most appropriate evaluation environments and tasks to evaluate LLM competencies?
- **RQ0.B)** What datasets are available and useful to us?
- **RQ0.C)** Of the publicly available LLMs, which are the most suitable for defining the baseline?

RT1: Adapt to follow evaluation guidelines in environments without training examples (zero-shot scenario). The task will focus on training LLMs to follow evaluation guidelines. The goal is to create a model independent of domains and guidelines. That is, the model should be able to adapt to new guidelines. LLMs will be adapted to carry out the task in situations with no examples or few examples. The main research questions in this section will be the following:

- **RQ1.A)** In an environment with no or few examples, what technique is most effective for incorporating domain-associated knowledge into the model?

- **RQ1.B)** Are RAG and textual conditioning learning effective techniques for models to learn to follow guidelines? If so, how can we implement them in the language learning domain?

RT2: Train LLMs by instruction to learn to follow guidelines. In this task, we will study methods to overcome data scarcity, to learn to follow guidelines. In addition, auxiliary tasks will be defined using synthetic data and avoiding the need for manual annotations. These tasks will help in the learning process. We will focus on the following research questions:

- **RQ2.A)** In synthetic data generation, what are the most effective techniques for teaching LLMs to follow orders?
- **RQ2.B)** What auxiliary tasks might be most suitable for teaching LLMs to follow guidelines?

RT3: Align LLMs with user needs. The objective is to investigate different methods for LLMs to carry out appropriate planning and reasoning and provide indications, explanations, and recommendations to users. This task will include the following research questions:

- **RQ3.A)** What is the best way to give feedback to users after reasoning?
- **RQ3.B)** How could we adapt LLMs to generate exercises and comments based on the educational needs and competencies of users?
- **RQ3.C)** Can we train an LLM dynamically and automatically to create exercises and comments?

Annual Research Planning:

First year, foundations: The work to be carried out during the first year will be related to tasks RT0 and RT1. The objective is to prepare the research environment and, therefore, create the evaluation methodology and carry out the first experiments. The following tasks are planned:

1.1) To answer questions RQ0.A and RQ0.B, the evaluation environment will be defined. The necessary datasets will be collected, and, if necessary for the project, a proprietary dataset will be created.
1.2) We will evaluate available LLMs and develop basic techniques to answer RQ0.C.
1.3) To answer RQ1.A, we analyze state-of-the-art models. We will perform a quantitative and qualitative analysis of the shortcomings of these systems: Their behavior across datasets and different tasks will be analyzed in depth.
1.4) To answer RQ1.B, we will try to improve textual conditioning learning methods to adequately follow guidelines.
1.5) We expect to send the answers and conclusions obtained from the different RQs to high-level journals and conferences.

Second year, model adaptation: In the second year, work will be done on tasks RT1 and RT2. The objective is to finish researching training methodologies for LLMs to follow guidelines. The following tasks are planned:

2.1) Taking advantage of the conclusions from the experiments performed, we will try to improve the system created to adequately answer question RQ1.B.
2.2) To answer question RQ2.A, synthetic data methods will be analyzed, as well as the shortcomings of current techniques. Using what was learned from the research conducted, new methods for generating effective synthetic data will be proposed.
2.3) To answer question RQ2.B, on the other hand, we will analyze works to create auxiliary tasks and design and create tasks appropriate to our domain.
2.4) With the learning obtained from questions RQ1 and RQ2, a more powerful and better model will be developed. We will use the data generated in RQ1 and the auxiliary tasks from RQ2 to improve the results of state-of-the-art techniques.
2.5) At least one article will be submitted to a main conference or journal, based on what is obtained when answering these research questions.

Third year, model alignment: During the third year, we will try to complete RT3. Our goal is to align LLMs to the needs of users, so that they generate appropriate responses, improvements, and exercises.

3.1) The work done in section RT0 will be reviewed and new evaluation datasets will be added and updated, if applicable.
3.2) To answer question RQ3.A, datasets associated with symbolic reasoning will be collected and adapted. In addition, synthetic data generation techniques will be applied to improve the reasoning capacity of the models.
3.3) The technologies and methods developed during the second year of the project will interact with the model arising from question RQ3.A, to then answer RQ3.B.
3.4) We will conduct experiments to evaluate the improved model we are going to create, and at the same time pay attention to question RQ3.C.
3.5) The results obtained with the new model will be sent to a main journal and conference.

Fourth year, refinement: During the first month, the results obtained in previous years will be collected and completed. Then, the thesis will be written, and the defense will be prepared. To carry out these objectives, the following tasks have been defined:

4.1) Refine tasks from previous years.
4.2) Submit an article to a journal.
4.3) Write the thesis.
4.4) Prepare the thesis defense.

5. Specific Issues of Research to be Discussed

Our research focuses primarily on evaluating written texts according to specific rubrics, then providing feedback and creating exercises based on the educational needs of the user. Although there are plans to train models to follow guidelines for creating specific and useful feedback [2], we are uncertain about how to adapt these techniques to the educational domain in low-resource languages with data scarcity. Preliminary experiments have shown that models can predict individual marks across different evaluation criteria, but we remain unsure which techniques would be adequate to verbalize the inner reasoning process of the model to create specific feedback while avoiding ambiguous or general comments.

The evaluation of the feedback and generated exercises presents a challenging task in our work. While evaluation based on agreement with GPT-4 or other closed models is widely used in the field [36], the linguistic capabilities of these models for Basque lag behind newer open-source models [1], suggesting they may not be appropriate for evaluating feedback and generated exercises in this context.

Acknowledgments

This PhD will be partially supported by:

- The Basque Government (IKER-GAITU project).
- Ixa group A type research group (IT1570-22)
- Ekhi Azurmendi hold a PhD grant from the Basque Government (PRE_2024_1_0035).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Etxaniz, O. Sainz, N. Perez, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An open language model and evaluation suite for basque, 2024. URL: https://arxiv.org/abs/2403.20266. arXiv:2403.20266.

[2] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, M. Seo, Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL: https://arxiv.org/abs/2310.08491. arXiv:2310.08491.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, 2022. URL: https://arxiv.org/abs/2204.02311. arXiv:2204.02311.

[5] T. Scialom, T. Chakrabarty, S. Muresan, Continual-t0: Progressively instructing 50+ tasks to language models without forgetting, arXiv preprint arXiv:2205.12393 (2022).

[6] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).

[7] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. URL: https://arxiv.org/abs/1905.00537. arXiv:1905.00537.

[8] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, 2021. URL: https://arxiv.org/abs/2009.03300. arXiv:2009.03300.

[9] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. G.-A. et al. 2022, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL: https://arxiv.org/abs/2206.04615. arXiv:2206.04615.

[10] F. Kamalov, D. Santandreu Calonge, I. Gurrib, New era of artificial intelligence in education: Towards a sustainable multifaceted revolution, Sustainability 15 (2023) 12451.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[12] OpenAI, Gpt-4 technical report, https://cdn. openai.com/papers/gpt-4.pdf (2023).

[13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023).

[14] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al., Improving language models by retrieving from trillions of tokens, in: International conference on machine learning, PMLR, 2022, pp. 2206–2240.

[15] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, E. Grave, Atlas: Few-shot learning with retrieval augmented language models, Journal of Machine Learning Research 24 (2023) 1–43.

[16] O. Sainz, I. García-Ferrero, R. Agerri, O. L. de Lacalle, G. Rigau, E. Agirre, Gollie: Annotation guidelines improve zero-shot information-extraction, 2024. URL: https://arxiv.org/abs/2310.03668.

arXiv:2310.03668.
[17] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, M. Lewis, Measuring and narrowing the compositionality gap in language models, arXiv preprint arXiv:2210.03350 (2022).
[18] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, G. Neubig, Pal: Program-aided language models, in: International Conference on Machine Learning, PMLR, 2023, pp. 10764–10799.
[19] R. Liu, J. Wei, S. S. Gu, T.-Y. Wu, S. Vosoughi, C. Cui, D. Zhou, A. M. Dai, Mind's eye: Grounded language model reasoning through simulation, arXiv preprint arXiv:2210.05359 (2022).
[20] R. S. Sutton, A. G. Barto, et al., Reinforcement learning: An introduction, volume 1, MIT press Cambridge, 1998.
[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.
[22] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, Galactica: A large language model for science, arXiv preprint arXiv:2211.09085 (2022).
[23] A. Creswell, M. Shanahan, I. Higgins, Selection-inference: Exploiting large language models for interpretable logical reasoning, arXiv preprint arXiv:2205.09712 (2022).
[24] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL: https://arxiv.org/abs/2402.03300. arXiv:2402.03300.
[25] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, Z. Zhang, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.
[26] J. Dwivedi-Yu, T. Schick, Z. Jiang, M. Lomeli, P. Lewis, G. Izacard, E. Grave, S. Riedel, F. Petroni, Editeval: An instruction-based benchmark for text improvements, arXiv preprint arXiv:2209.13331 (2022).
[27] K. Malinka, M. Peresíni, A. Firc, O. Hujnák, F. Janus, On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?, in: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, 2023, pp. 47–53.
[28] T. Susnjak, T. R. McIntosh, Chatgpt: The end of online exam integrity?, Education Sciences 14 (2024). URL: https://www.mdpi.com/2227-7102/14/6/656. doi:10.3390/educsci14060656.
[29] K. Tan, T. Pang, C. Fan, S. Yu, Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects, arXiv preprint arXiv:2305.03433 (2023).
[30] H. McNichols, W. Feng, J. Lee, A. Scarlatos, D. Smith, S. Woodhead, A. Lan, Automated distractor and feedback generation for math multiple-choice questions via in-context learning, arXiv preprint arXiv:2308.03234 (2023).
[31] S. K. Bitew, J. Deleu, C. Develder, T. Demeester, Distractor generation for multiple-choice questions

with predictive prompting and large language models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2023, pp. 48–63.

[32] J. M. Arriola, M. Iruskieta, E. Arrieta, J. Alkorta, Towards automatic essay scoring of basque language texts from a rule-based approach based on curriculum-aware systems, in: Proceedings of the NoDaLiDa 2023 Workshop on Constraint Grammar-Methods, Tools and Applications, 2023, pp. 20–28.

[33] E. Agirre, I. Aldabe, X. Arregi, M. Artetxe, U. Atutxa, E. Azurmendi, I. De la Iglesia, J. Etxaniz, V. García-Romillo, I. Hernaez-Rioja, et al., Iker-gaitu: research on language technology for basque and other low-resource languages (2024).

[34] N. Perez, M. Cuadros, Multilingual call framework for automatic language exercise generation from free text, in: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 49–52.

[35] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct preference optimization: Your language model is secretly a reward model, 2024. URL: https://arxiv.org/abs/2305.18290. arXiv:2305.18290.

[36] L. Zhu, X. Wang, X. Wang, Judgelm: Fine-tuned large language models are scalable judges, 2025. URL: https://arxiv.org/abs/2310.17631. arXiv:2310.17631.