Prompt Engineering for evaluators: optimizing LLMs to judge linguistic proficiency

Lorenzo Gregori (Università degli Studi di Firenze)

lorenzo.gregori(at)unifi.it

Abstract

Prompt Engineering, the practice of optimizing the question made to a Large Language Model, is closely linked to the evaluation procedures. Depending on the type of task we are performing through LLMs, we can have an evaluation metric with high or low reliability, making Prompt Engineering more or less effective. LLM-as-a-judge represents a possible solution to perform Prompt Engineering in tasks that are hard to evaluate, although the reliability of this practice is not granted, depending on the task and the language model. This paper presents an evaluation of general purpose LLMs in an essay-scoring task using state-of-the-art small models. In particular, the ability of language models to assign proficiency levels to short essays written by Italian L2 learners is evaluated. Test data with expert annotations of CEFR scores are extracted from Kolipsi-II corpus. Several prompting techniques have been used to analyze the impact of Prompt Engineering on this task. Results show a wide difference in accuracy among the three LLMs and that choosing the right prompt radically changes their rating abilities.

Keywords

large language models, prompt engineering, LLM-as-a-judge, evaluation

1 Prompt Engineering and evaluation

Large Language Models (LLMs) have represented a revolution in computational linguistics, changing the research focus from methods to tasks: in fact, an LLM is a unique algorithm that can accomplish many different tasks with state-of-the-art performance (Örpek et al. 2024). Before LLMs, computational linguistics studies were focused on developing language models, trained to perform a single task, while nowadays a unique pre-trained model can address most of the tasks. Within this new paradigm, prompt engineering (PE) became a fundamental research field, developing practices and techniques aimed at optimizing the creation of prompts provided to the models. Brown et al. (2020) performed a comprehensive evaluation of GPT-3 model, showing that PE has a relevant effect on the accuracy,[1] and this effect grows as the model's size increases.

According to the transformers architecture (Vaswani et al. 2017), an LLM receives in input the numeric representation of a tokenized text (input embedding) and outputs a probability distribution of the token vocabulary; a new token is sampled at every iteration from this probability distribution. In this process, the

[1] Average accuracy measured on SuperGLUE benchmark (Wang et al. 2019) containing a wide variety of language understanding tasks.

AILing

AI-Linguistica.
Linguistic Studies on AI-Generated Texts and Discourses

Gregori, Lorenzo. 2025.
Special Issue: AI-Driven Linguistic Studies, Vol.2. No. 2.
DOI: 10.62408/ai-ling.v2i2.22
ISSN: 2943-0070

input text provided to the model, i.e. the prompt, is a crucial component: it's encoded into the input vectors and, consequently, it affects the output tokens. In fact, several studies show that LLMs are extremely sensitive to prompts, and even minor input variations can change the accuracy of LLM answers (see Voronov et al. 2024; Alzahrani et al. 2024; Sclar et al. 2023).

In the last years, prompt engineering emerged as a research field, developing prompting techniques that can be used to optimize LLM output by tuning the input question. Phoenix and Taylor (2024) define prompt engineering as "the process of discovering prompts that reliably yield useful or desired results". This definition highlights that PE is not made of formal techniques, but it's a continuous process of refinement of questions towards optimal answers. From this definition, it emerges that PE is not applicable in the same way to every task: in particular, it is straightforward with all the tasks in which it's possible to determine exactly if the answer is right or wrong (e.g. classification, multiple choice questions). In this kind of task, the model output is evaluated through a metric against a test set with target annotations. This allows us to compute exactly the model performance on a task and to refine the provided prompt to obtain better results.

Conversely, natural language generation tasks – like machine translation, summarization, automatic report generation – are harder to evaluate. In fact, the test sets for these tasks, if available, contain possible examples of correct output, that need to be compared to the real output through a metric. BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and BERTscore (Zhang et al. 2020) are examples of common metrics for Natural Language Generation (NLG) tasks, but several studies have shown that they often have a medium or low correlation with human evaluation (Reiter 2018; Deriu et al. 2021; Briman and Yildiz 2024).

Moreover, LLM can also be used for more creative tasks, like narrative generation, for which the test set does not exist. In general, in NLG we need a different kind of evaluation: not to match a reference annotation, but to provide a score based on the analysis of syntax, semantics, content, orthography, etc.

For those tasks for which a metric is not reliable or is not applicable, PE is more complex, because prompts can't be optimized to reduce the task error, that is not measurable. Moreover, performing a human evaluation of generated text for every run is usually not affordable. In these cases, an emerging practice is the use of LLM-as-a-judge (Zheng et al. 2023), asking LLMs to evaluate a text in place of humans. This practice is becoming very popular and recently it has been used for a wide variety of tasks, including machine translation (Huang et al. 2024), summarization (Gao et al. 2023), and essay scoring (Lee et al. 2024), allowing to obtain reliable human-like judgments. LLM-as-a-judge allows having an automatic evaluation for every AI-generated text, enabling input optimization through PE in all the tasks in which a reliable metric is missing. The drawback of this approach is that we cannot rely on automatic judgments without testing the model's ability as an evaluator.

This paper presents an evaluation of LLMs in judging the proficiency level of Italian through the analysis of short essays written by Italian L2 learners and the

role of prompt engineering in this task. The same task has been addressed before by Yancey et al. (2023), which exploited state-of-the-art commercial models on English learners, obtaining a good agreement with the human raters. Conversely, the current experiment aims to evaluate the performance of small models, with a strong focus on their optimization through prompt engineering.

2 The Kolipsi-II corpus

The Kolipsi-II Corpus (Glaznieks et al. 2023) is a written learner corpus comprising texts from German and Italian L2 speakers in South Tyrol, Italy. Developed during the KOLIPSI II project (Vettori and Abel 2017), it investigates linguistic and socio-psychological aspects of L2 acquisition, specifically among South Tyrolean pupils aged 16-18. Data collection occurred in 2014 and involved two standardized written tasks: (1) a narrative email recounting an event based on a picture story, and (2) an argumentative email discussing negative aspects of social media chats. These tasks, conducted under strict time constraints (25 minutes) and without reference materials, were adapted slightly from the original study.

Learner texts are annotated with CEFR levels, providing expert evaluations of coherence, sociolinguistic appropriateness, lexical accuracy and diversity, grammar, and orthography. The Kolipsi-II Corpus contains detailed metadata about the students: their language background, socio-demographic factors, and educational context.

Kolipsi dataset

Exploiting the rich corpus metadata, a representative sample of Kolipsi-II corpus has been extracted, by filtering the students with the following characteristics:

- L1 language is German
- Language group affiliation is German
- Both parents' L1 language is German
- The average grade of the previous school year is between 6 and 7

The derived dataset contains ~150,000 tokens, belonging to 816 texts written by 408 students (see Table 1): each student wrote a narrative and an argumentative essay. Each text is evaluated with CEFR levels for each of the following parameters:

- Sociolinguistic appropriateness
- Coherence and cohesion
- Grammar correction.
- Lexical accuracy
- Lexical diversity
- Orthography accuracy

For this experiment, an overall CEFR score is computed by averaging the scores of the six parameters. The average is computed by converting the CEFR scores in

integer values from 1 to 6 (A1 = 1, A2 = 2, and so on), calculating the mean rounded to the nearest integer, and converting back the numeric values to the CEFR scores.

Table 1: Kolipsi dataset.

	Kolipsi-2_IT	**Dataset**
# tokens	400,000	151,472
# texts	2,063	816
# writers	1,035	408

The dataset is highly unbalanced in terms of CEFR classification, with 64% of items belonging to the B1 class (Figure 1).

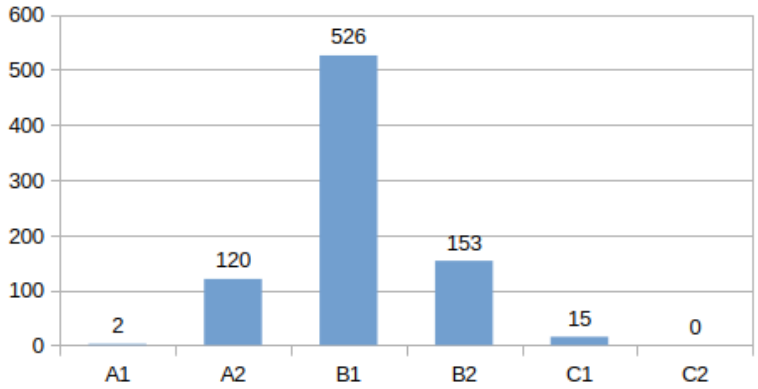

Figure 1: distribution of CEFR classes in Kolipsi dataset.

3 Experiment

The Kolipsi dataset has been exploited to check the ability of LLMs to provide a CEFR rating and to assess them through prompt engineering techniques.

Three LLMs have been used: Llama-3.2-3B-Instruct, Falcon3-3B-Instruct, Qwen2.5-3B-Instruct. All these models are free,[2] small (3 billion parameters), instruction-tuned, and very recent (see Table 2).

Table 2: LLMs used for the experiment.

Model class	**Model version**	**Parameters**	**Instructed**	**Release date**
Llama	3.2	3 billion	Yes	September 2024
Falcon	3	3 billion	Yes	December 2024
Qwen	2.5	3 billion	Yes	September 2024

[2] Freely available through Huggingface website: https://huggingface.co/

Six prompting techniques have been used: Zero-shot, Few-shot (Brown et al. 2020), Self-consistency (Wang et al. 2023), Role prompting (Shanahan et al. 2023), Zero-shot Chain-of-though (Kojima et al. 2023), Few-shot Chain-of-thought (Wei et al. 2022). All the prompts for this analysis are written in Italian. Every text in the dataset (816) has been evaluated by each LLM (3) with 6 prompts, for a total of 14,688 LLM calls.

With zero-shot prompting, the LLMs are simply asked to output a CEFR rating, according to the linguistic proficiency text, after providing some additional details about the task. With Few-Shot prompting, two items are removed from the test set and used in the prompt, providing the model with two examples of texts annotated with human ratings; this works as a mini-training, pushing the model to replicate the behavior just seen. With Role prompting, the model is asked to behave as a teacher of Italian who evaluates a student essay. For Self-Consistency, the zero-shot prompt has been run five times to obtain multiple ratings; then, the most frequent value is selected. Chain-of-thought (CoT) reasoning has been aroused in two different prompts: by explicitly asking the model to proceed step by step (zero-shot CoT), and by providing an example of human-annotated text with the evaluation of every textual aspect (few-shot CoT). The full list of prompts is reported in the Appendix.

Each prompt is run with the hyper-parameters reported in Table 3. The temperature at 1 is useful to enhance the answers variability preserving answer's reliability without having hallucination issues, that are typical with the temperature above 1 (Renze and Guven 2024). A seed has been fixed for experiment replication purposes.

Table 3: Hyper-parameters setup.

	temperature	**top_k**	**seed**
Zero-shot	1.0	500	25
Few-shot	1.0	500	25
Self-consistency	1.0	500	25, 26, 27, 28, 29
Role prompting	1.0	500	25
Chain-of-thought (Z-S)	1.0	500	25
Chain-of-thought (F-S)	1.0	500	25

4 Results

Table 4 shows the number of evaluated mails by every model with every prompt, that is the number of runs in which the LLMs were able to output a score of proficiency according to the CEFR scale. In fact, despite the prompts being designed to be clear in terms of instructions, sometimes the answer does not contain

a CEFR value. This phenomenon occurs mainly with chain-of-thought prompts, where the LLM performs more complex reasoning and has a higher probability of getting lost in its "thoughts". All the outputs are manually revised to extract the CEFR value from generated texts: in fact, with chain-of-thought prompts, and occasionally with other types of prompts, a whole text is produced, instead of a single CEFR value. This is a common LLM behavior that could be solved using a second prompt to extract relevant information; in this case, a manual revision is performed to avoid introducing additional errors from the extraction task.

Considering the task type and the manual revision of every LLM answer, a simplified version of the answer relevance metric is used, assigning 1 if a CEFR value could be derived from the LLM output, and 0 else.

Table 4: Number of relevant answers provided for every item.

	Llama	**Falcon**	**Qwen**
Zero-shot	816 (100%)	799 (98%)	816 (100%)
Few-Shot	782 (96%)	812 (99%)	810 (99%)
Selc-cons.	806 (99%)	801 (98%)	816 (100%)
Role pr.	816 (100%)	793 (97%)	814 (99%)
CoT (Z-S)	639 (78%)	711 (87%)	430 (52%)
CoT (F-S)	614 (75%)	803 (98%)	815 (99%)

Quadratic Weighted Kappa (QWK) metric has been used to measure the agreement between the overall CEFR scores and the predicted ones. Results (Table 5) report a QWK value between -0.1 and 0.1 for every model with every prompting technique, showing a substantial absence of correlation between annotators and models.

The experiment by Yancey and colleagues (2023) obtained a QWK between 0.66 and 0.89 with GPT-4 model, showing that state-of-the-art models are good rating estimators in this task. Moreover, in that experiment, prompt engineering techniques have a strong impact on the output quality. Otherwise, in the current experiment, the prompting impact is not so clear, given that QWK scores are very low.

In addition to QWK, *accuracy±1* scores for each model are also reported: this modified accuracy metric considers a predicted score correct if it is the same value of the reference score or differs by one value from the reference score (e.g. the real score is B2 and the LLM assigns B1 or C1). This metric aims at estimating the plausible evaluations and is useful in this context, because it allows us to measure how LLMs are able to act as human raters, which can disagree with the assigned score.

Small models have fewer reasoning capabilities and are more prone to biased evaluation (Popov et al. 2025). During this experiment, it occurred

frequently that the LLMs reacted to a prompt by answering always with the same rating (or with the same two alternated ratings). These cases are considered biased behavior because the rating does not depend on the text provided in the input, highlighting a lack of analysis.

In this paper, a dataset evaluation made by an LLM through a prompt has been considered biased if the percentage of the two most frequent ratings is greater than 95%. For example, Llama model with Self-Consistency technique answered A2 or B2 in 98.14% of cases,[3] ignoring any other rating. The cases of biased evaluation are removed from the analysis (see Table 5).

Table 5: QWK and accuracy±1 scores; biased evaluations are marked with asterisk.

	Prompt	**QWK**	**accuracy±1**	**biased**
Llama	**Zero-shot**	-0.006	0.631	
	Few-Shot	0.028	0.706	
	Selc-cons.	-0.051	0.772	*
	Role pr.	-0.021	0.659	
	CoT (Z-S)	0.021	0.728	
	CoT (F-S)	0.036	0.734	
Falcon	**Zero-shot**	-0.013	0.854	*
	Few-Shot	-0.004	0.883	
	Selc-cons.	-0.010	0.851	*
	Role pr.	-0.004	0.851	*
	CoT (Z-S)	-0.013	0.861	
	CoT (F-S)	0.039	0.711	
Qwen	**Zero-shot**	-0.007	0.545	*
	Few-Shot	0.069	0.736	
	Selc-cons.	-0.010	0.794	*
	Role pr.	-0.013	0.515	*
	CoT (Z-S)	0.007	0.653	*
	CoT (F-S)	0.060	0.650	

[3] The percentage of A2+B2 in the original dataset is 33.45%.

5 Discussion

Results show poor quality of small foundation LLMs in proficiency rating: they do not exhibit enough capability to align with human ratings. Regarding accuracy, the best results are obtained by few-shot prompting, both the simple version and the one paired with chain-of-thought: this result is confirmed in each model, highlighting the validity of exemplar-based prompting in proficiency evaluation tasks.

Regarding answer relevance, the critical prompts for all the LLMs are the two types of CoT, reaching very poor results: for example, only 52% of relevant answers by Qwen with CoT, and only 75% by Llama with few-shot CoT. Conversely, all the other prompts were able to provide an evaluation for at least 96% of the items. Surprisingly, in two models, Falcon and Qwen, the structured pattern of Few-Shot CoT solved the problem, bringing the answer relevance to 98/99%; this did not happen with Llama, which obtained similar results in both types of CoT.

6 Conclusion and future work

This paper shows that small LLMs are unable to perform a human-like language proficiency evaluation of Italian learner essays. The strong connection between prompt engineering and evaluation is highlighted, presenting the use of prompt engineering to optimize LLMs to evaluate learner texts. Six prompting techniques have been used with three small and free LLMs, obtaining poor results with respect to the previous findings on state-of-the-art commercial models. This confirms the complexity of the proposed task but also suggests new advancement in the analysis: an extension of this work is foreseen, using bigger models (7B) and including an Italian model to see if, and to what extent, an LLM trained on Italian data can perform a more fine-grained analysis of Italian texts and provide a better evaluation.

7 Acknowledgment of AI tools

The freely available LLMs Llama-3.2-3B-Instruct, Falcon3-3B-Instruct and Qwen2.5-3B-Instruct were necessary to conduct this research. They have been downloaded from Huggingface platform, queried through Python libraries and run on Google T4 GPUs.

8 Conflicts of interest

The author declares no conflicts of interest regarding the publication of this contribution.

References

Alzahrani, Norah & Alyahya, Hisham Abdullah & Alnumay, Yazeed & Alrashed, Sultan & Alsubaie, Shaykhah & Almushaykeh, Yusef & Mirza, Faisal & Alotaibi, Nouf & Altwairesh, Nora & Alowisheq, Areeb et al. 2024. When benchmarks are targets: Revealing the sensitivity of Large Language Model leaderboards. arXiv. https://doi.org/10.48550/arXiv.2402.01781

Briman, Mohammed Khalid Hilmi & Yildiz, Beytullah. 2024. Beyond ROUGE: A comprehensive evaluation metric for abstractive summarization leveraging similarity, entailment, and acceptability. *International Journal on Artificial Intelligence Tools* 33(5). https://doi.org/10.1142/S0218213024500179

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda et al. 2020. Language Models are Few-Shot learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Deriu, Jan & Rodrigo, Alvaro & Otegi, Arantxa & Echegoyen, Guillermo & Rosset, Sophie & Agirre, Eneko & Cieliebak, Mark. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54(1). 755–810. https://doi.org/10.1007/s10462-020-09866-x

Gao, Mingqi & Ruan, Jie & Sun, Renliang & Yin, Xunjian & Yang, Shiping & Wan, Xiaojun. 2023. Human-like summarization evaluation with ChatGPT. arXiv. https://doi.org/10.48550/arXiv.2304.02554

Glaznieks, Aivars & Frey, Jennifer-Carmen & Abel, Andrea & Nicolas, Lionel & Vettori, Chiara. 2023. The Kolipsi Corpus family: Resources for learner corpus research in Italian and German. *IJCoL. Italian Journal of Computational Linguistics* 9(2). https://doi.org/10.4000/ijcol.1210

Huang, Xu & Zhang, Zhirui & Geng, Xiang & Du, Yichao & Chen, Jiajun & Huang, Shujian. 2024. Lost in the Source Language: How Large Language Models Evaluate the Quality of Machine Translation. arXiv. https://doi.org/10.48550/arXiv.2401.06568

Kojima, Takeshi & Gu, Shixiang Shane & Reid, Machel & Matsuo, Yutaka & Iwasawa, Yusuke. 2023. Large Language Models are Zero-Shot reasoners. arXiv. https://doi.org/10.48550/arXiv.2205.11916

Lee, Sanwoo & Cai, Yida & Meng, Desong & Wang, Ziyang & Wu, Yunfang. 2024. Unleashing Large Language Models' proficiency in Zero-shot essay scoring. In Al-Onaizan, Yaser & Bansal, Mohit & Chen, Yun-Nung (eds), *Findings of the Association for Computational Linguistics EMNLP 2024*, 181–198. Miami: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.10

Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004* (Barcelona, July 25-26, 2004), 74–81. https://aclanthology.org/W04-1013/ (last access on 29/01/2025).

Nezhad, Sina Bagheri & Agrawal, Ameeta. 2024. What drives performance in multilingual Language Models? In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects VarDial 2024* (Mexico City, June 20, 2024), 16–27. https://doi.org/10.18653/v1/2024.vardial-1.2

Örpek, Zeynep & Tural, Büşra & Destan, Zeynep. 2024. The Language Model revolution: LLM and SLM analysis. In *Proceedings of the 8th International Artificial Intelligence and Data Processing Symposium IDAP 2024* (Online, September 21-22, 2024), 1–4. https://doi.org/10.1109/IDAP64064.2024.10710677

Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei-Jing. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (Philadelphia, July 22, 2002), 311–318. https://aclanthology.org/P02-1040.pdf (last access on 29/01/2025).

Phoenix, James & Taylor, Mike. 2024. *Prompt Engineering for Generative AI. Future-Proof Inputs for Reliable AI Outputs*. O'Reilly Media.

Popov, Ruslan O. & Karpenko, Nadiia V. & Gerasimov, Volodymyr V. 2025. Overview of small language models in practice. In *Proceedings of the 7th Workshop for Young Scientists in Computer Science & Software Engineering CS&SE@SW2024* (Online, December 27, 2024), 164–182. https://ceur-ws.org/Vol-3917/paper28.pdf

Reiter, Ehud. 2018. A structured review of the validity of BLEU. *Computational Linguistics* 44(3). 393–401. https://doi.org/10.1162/coli_a_00322

Sclar, Melanie & Choi, Yejin & Tsvetkov, Yulia & Suhr, Alane. 2023. Quantifying Language Models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. arXiv. https://doi.org/10.48550/arXiv.2310.11324

Shanahan, Murray & McDonell, Kyle & Reynolds, Laria. 2023. Role play with large language models. *Nature* 623(7987). 493–498. https://doi.org/10.1038/s41586-023-06647-8

Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N. & Kaiser, Łukasz & Polosukhin, Illia. 2017. Attention is all you need. In Guyon, Isabel & Von Luxburg, Ulrike & Bengio, Samy & Wallach, Hanna & Fergus, Rob & Vishwanathan, S.V.N. & Garnett, Roman (eds), *Advances in Neural Information Processing Systems* 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (last access on 29/01/2025).

Vettori, Chiara & Abel, Andrea. 2017. *KOLIPSI II: gli studenti altoatesini e la seconda lingua; indagine linguistica e psicosociale = KOLIPSI II: die Südtiroler SchülerInnen und die Zweitsprache; eine linguistische und sozialpsychologische Untersuchung*. Bolzano: Eurac research. https://doi.org/10.13140/RG.2.2.24248.96001

Voronov, Anton & Wolf, Lena & Ryabinin, Max. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. arXiv. https://doi.org/10.48550/arXiv.2401.06766

Wang, Alex & Pruksachatkun, Yada & Nangia, Nikita & Singh, Amanpreet & Michael, Julian & Hill, Felix & Levy, Omer & Bowman, Samuel. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In Wallach, Hanna & Larochelle, Hugo & Beygelzimer, Alina & d'Alché-Buc, Florence & Fox, Emily (eds), *Advances in Neural Information Processing Systems* 32.

https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html (last access on 29/01/2025).

Wang, Xuezhi & Wei, Jason & Schuurmans, Dale & Le, Quoc & Chi, Ed & Narang, Sharan & Chowdhery, Aakanksha & Zhou, Denny. 2023. Self-consistency improves Chain of Thought reasoning in Language Models. arXiv. https://doi.org/10.48550/arXiv.2203.11171

Wei, Jason & Wang, Xuezhi & Schuurmans, Dale & Bosma, Maarten & Ichter, Brian & Xia, Fei & Chi, Ed & Le, Quoc & Zhou, Denny. 2022. Chain-of-Thought prompting elicits reasoning in Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2201.11903

Yancey, Kevin P. & Laflair, Geoffrey & Verardi, Anthony & Burstein, Jill. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In Kochmar, Ekaterina & Burstein, Jill & Horbach, Andrea & Laarmann-Quante, Ronja & Madnani, Nitin & Tack, Anaïs & Yaneva, Victoria & Yuan, Zheng & Zesch, Torsten (eds), In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications BEA 2023* (Toronto, July 13, 2023), 576–584. https://doi.org/10.18653/v1/2023.bea-1.49

Zhang, Tianyi & Kishore, Varsha & Wu, Felix & Weinberger, Kilian Q. & Artzi, Yoav. 2020. BERTScore: Evaluating text generation with BERT. arXiv. https://doi.org/10.48550/arXiv.1904.09675

Zheng, Lianmin & Chiang, Wei-Lin & Sheng, Ying & Zhuang, Siyuan & Wu, Zhanghao & Zhuang, Yonghao & Lin, Zi & Li, Zhuohan & Li, Dacheng & Xing, Eric P. et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv. https://doi.org/10.48550/arXiv.2306.05685

Appendix

Full prompts used for the experiments. The [[mail]] wildcard is used to identify the variable part, i.e. the content of each mail in the dataset. The prompts reported here are related to the narrative texts in the Kolipsi dataset. For the argomentative texts the same prompts are used, with the replacement of “per raccontare un evento accaduto al supermercato” with “per raccontargli la propria esperienza e le proprie opinioni sul mondo di internet e delle chat”.

Zero-shot

A uno studente delle scuole superiori è stato chiesto di scrivere una mail a un amico per raccontare un evento accaduto al supermercato. Lo studente è un apprendente italiano di madrelingua tedesca. La mail che ha scritto è la seguente.

[[mail]]

Dai una valutazione di questa mail utilizzando i valori della scala CEFR (A1, A2, B1, B2, C1, C2) considerando i seguenti aspetti: ampiezza del lessico, padronanza del lessico, correttezza grammaticale, coerenza e coesione testuale, appropriatezza sociolinguistica, padronanza ortografica. Rispondi solo con il giudizio finale, senza aggiungere altro.

Few-shot

A degli studenti delle scuole superiori è stato chiesto di scrivere una mail a un amico per raccontare un evento accaduto al supermercato. Gli studenti sono apprendenti di italiano di madrelingua tedesca. Si richiede di valutare queste mail utilizzando la scala CEFR (A1, A2, B1, B2, C1, C2). Scrivere solo il valore CEFR senza ulteriori aggiunte.

MAIL: Ciao Angelika!
Come stai? Io non mi sento molto bene. Oggi è successo una cosa molto imparazzata a me.Oggi mattina sono andato dal fruttivendolo perché volevo fare una maccedonia per i bambini. Ho preso tutto che ho servato e dopo ho pagato. In auto mi ho riccordato che servo ancora pane e latte. Allora sono andato al supermercato per comprare queste cose. Quando avevo trovato tutto ho pagato alla cassa il pane e il latte. La signora alla cassa ha detto a me che cosa voglio fare con le frutte. Rispondevo che non sono del supermercato, sono da un altro negozio, ma lei non mi voleva credere e allora ha telefonato la polizia. Volevano vedere il scontrino, ma ho perso. Alla fine hanno visto che le frutte non erano le vostre, ma per me era così imbarazzabile che sono andato subito a casa
Ci vediamo!
Maria.
VALUTAZIONE CEFR: A2

MAIL: Caro Giuseppe,
ti devo raccontare cosa è successo a Maria due giorni fa durante che faceva la spesa. Prima è stata dal fruttivendolo dove lavora Michele per comprare delle mele, banane, fragole e delle uva. Ha preso la frutta ed è andata alla cassa per pagare. Dopo è uscita dal negozio ma ha dimenticato di prendere lo scontrino della frutta pagata. Prima di tornare a casa, Maria è ancora andata al supermercato per comprare del pesce e un po' di pane e portava la frutta comprata dentro una borsa con sé. Ha preso il pesce e il pane ed è andata alla cassa. Dopo di aver pagato si é girata e la signora dietro la cassa ha visto la borsa piena di frutta. La cassiera ha voluto che Maria pagasse per la frutta. Maria ha

provato di farla capire che la frutta ha comprato dal fruttivendolo e che non la voleva rubare. Ma senza scontrino non era abile a chiarire la situazione e per non prendere una denuncia doveva pagare per la frutta di nuovo anche al supermercato. Da questo giorno in poi, Maria sicuramente non dimenticerà più a prendere lo scontrino dopo di aver pagato!
Ti auguro ancora un bello fine settimana!
Il tuo caro amico, Moritz!
VALUTAZIONE CEFR: C1

MAIL: [[mail]]
VALUTAZIONE CEFR:

Chain-of-thought (Zero-shot)

A uno studente delle scuole superiori è stato chiesto di scrivere una mail a un amico per raccontare un evento accaduto al supermercato. Lo studente è un apprendente italiano di madrelingua tedesca. La mail che ha scritto è la seguente.

[[mail]]

Dai una valutazione di questa mail utilizzando i valori della scala CEFR (A1, A2, B1, B2, C1, C2) considerando i seguenti aspetti: ampiezza del lessico, padronanza del lessico, correttezza grammaticale, coerenza e coesione testuale, appropriatezza sociolinguistica, padronanza ortografica. Procedi passo dopo passo nella valutazione e rispondi solo con il giudizio finale, senza aggiungere altro.

Chain-of-thought (Few-shot)

Dai una valutazione della seguente mail utilizzando i valori della scala CEFR (A1, A2, B1, B2, C1, C2) considerando i seguenti aspetti: ampiezza del lessico, padronanza del lessico, correttezza grammaticale, coerenza e coesione testuale, appropriatezza sociolinguistica, padronanza ortografica. Alla fine fornisci un valore CEFR globale.

MAIL: Ciao Mario! Oggi ero al supermercato per fare dei acquisti per la famiglia. Quando sono arrivato alla cassa ho visto una scena straordinaria. Una donna alla età di 40 anni ha provato di rubare qualche cosa. La donna alla cassa voleva contattare la polizia, ma io le ho calmata. Maria, così si chiama la donna che rubava, ha cercato di spiegare che cosa è successo. Pochi minuti dopo ci era chiaro che le cose che erano nella tasca, lei ha comprato da un altro supermercato. Maria ci ha dato anche lo scontrino che ha ricevuta. La donna alla cassa si sentiva veramente male e ha cercato di chiedere scusa. Ma Maria era molto arrabbiata e ha detto che non vuole venire mai più!
VALUTAZIONE: ampiezza del lessico: A2, padronanza del lessico: B1, correttezza grammaticale: B1, coerenza e coesione testuale: B1, appropriatezza sociolinguistica: A2, padronanza ortografica: C1.
VALUTAZIONE FINALE: B1

MAIL: [[mail]]
VALUTAZIONE:

Role prompting

Sei un'insegnante di italiano in una scuola superiore tedesca del Südtirol. A un tuo studente hai chiesto di scrivere una mail a un amico per raccontare un evento accaduto al supermercato. Lo studente è un apprendente italiano di madrelingua tedesca. La mail che ha scritto è la seguente.

[[mail]]

Dai una valutazione di questa mail utilizzando i valori della scala CEFR (A1, A2, B1, B2, C1, C2) considerando i seguenti aspetti: ampiezza del lessico, padronanza del lessico, correttezza grammaticale, coerenza e coesione testuale, appropriatezza sociolinguistica, padronanza ortografica. Rispondi solo con il giudizio finale, senza aggiungere altro.