Can LLMs Solve Reading Comprehension Tests as Second Language Learners?

Akio Hayakawa[1], Horacio Saggion[1]

[1]LaSTUS Lab, TALN Research Group, Department of Engineering, Universitat Pompeu Fabra, C/Tànger 122 (08018), Barcelona, Spain

Abstract

The manual evaluation of natural language processing systems is costly and time-consuming, especially when targeting people with specific attributes as evaluators. Current large language models (LLMs) are reported to outperform humans at various tasks, and recently have been used as substitutes for human evaluators. LLMs also have shown the ability to behave as specified in a prompt. This progress raises a fundamental question: can LLMs mimic the behavior of language learners? In this study, we intentionally weaken LLMs aiming to make them simulate language learners on multiple-choice reading comprehension tests. By comparing answer distributions from language learners and LLMs, we observe that prompts designed to weaken the LLMs indeed degrade their performance. However, this degration does not bridge the gap between the original LLMs and language learners, thereby hilighting a critical discrepancy between them.

Keywords

Natural Language Processing, Large Language Models, Question Answering, Reading Comprehension

1. Introduction

In the field of Natural Language Processing (NLP), the evaluation of systems is commonly categorized into two approaches: automatic and manual evaluation. Manual evaluation, which is considered more reliable, involves methods ranging from subjective scoring on scales, such as a 5-point rating, to task-based assessments like solving comprehension questions. Despite its reliability, manual evaluation requires greater time and cost investments [1].

The difficulty of conducting manual evaluation significantly increases when targeting individuals with specific attributes, as access to these groups becomes more difficult. This has resulted in the diminished prioritization of their participation, calling into question the trustworthiness of manual evaluation. For instance, in the text simplification task, which aims to make texts more readable and understandable, children, language learners, and people with disabilities are considered ideal evaluators for the simplicity of texts, as they are presumed to benefit most from the simplification [2]. Nevertheless, studies on text simplification have relied on native speakers or people who do not need simplified texts for manual evaluation [3, 4], rarely involving individuals who need simplification, probably due to significant disparities in accessibility to diverse groups. Indeed, Sauberli et al. [5] recently demonstrated subjective differences in perceived text difficulty between people with and without intellectual disabilities, highlighting the importance of their involvement.

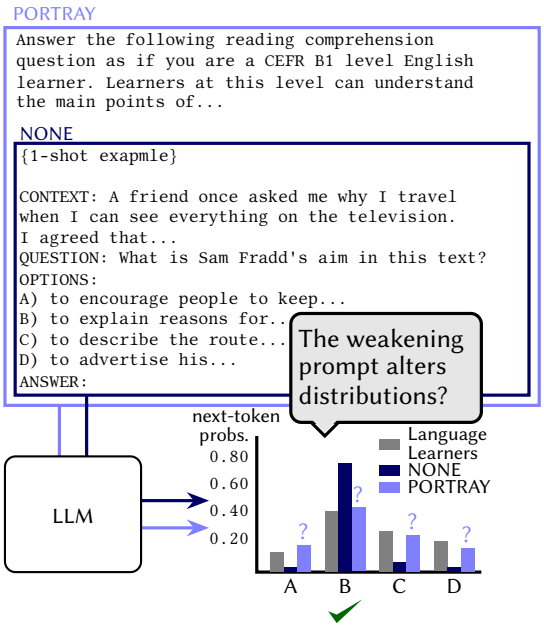

Figure 1: Overview of our experimental setup. We investigate whether it is possible to make next-token probabilities of LLM closer to selection distribution by language learners, by weakening the LLM.

Recent advancements in NLP, especially with Large Language Models (LLMs), may address this bottleneck. One line of work has attempted to substitute manual evaluation with assessments conducted by LLMs [6, 7, 8], seeking immediate and inexpensive annotations of higher

KiL'24: Workshop on Knowledge-infused Learning co-located with 30th ACM KDD Conference, August 26, 2024, Barcelona, Spain
akio.hayakawa@upf.edu (A. Hayakawa);
horacio.saggion@upf.edu (H. Saggion)
https://ahaya3776.github.io/ (A. Hayakawa)

quality. Another set of studies has reported that LLMs are capable of emulating a specific persona by including attributes in a prompt [9, 10].

Therefore, we wonder if LLMs could be prompted to serve as substitutes for specific personas. This study specifically focuses on language learners, investigating whether LLMs can mimic their response patterns. This approach could potentially offer a more accessible means of obtaining evaluations for tasks that ideally require responses from specific target groups, such as predicting the difficulty of questions without a pilot pretesting stage, simply by providing their attributes in the prompt.

To judge the mimicability of LLMs, we compare responses to multiple-choice reading comprehension (RC) tests, which have been widely used to measure language comprehension [11], from language learners and NLP systems. Using the CMCQRD dataset [12], which is a recently released four-choice RC test dataset with selection distributions from language learners, we aim to investigate if LLM output can closely approximate these distributions. While fine-tuning encoder models is one approach to pursuing distributions closer to those of humans [13], prompting LLMs has the potential to target a broader range of personas, suggesting enhanced applicability.

Figure 1 illustrates the outline of our experimental setup. Given that current models in the NLP field often achieve or even surpass human-level performance on various tasks [14], it is reasonable to presume that LLMs could outperform the average language learner on RC tests. Hence, LLMs need to be weakened to mimic language learners. We try several prompting techniques to degrade LLM performance and analyze their effects.

Contrary to our expectations, our preliminary experimental results show that the prompts considered do not lead LLMs to mimic language learners. Furthermore, we observe that the questions LLMs tend to answer incorrectly differ significantly from those that language learners struggle with. This discrepancy suggests a need for deeper analysis when we try to utilize LLM as a replacement for human evaluation.

2. Related Work

2.1. Human Response to Reading Comprehension Dataset

Reading comprehension (RC) tests have been widely used in psycholinguistic studies to assess how well readers, especially language learners, understand the content of a given text [15]. While these studies have seldom made their original data publicly available, research in natural language processing has made standard datasets available to measure the text comprehension abilities of machines [16, 17], sometimes for specific capabilities such as reasoning in HotpotQA [18] and the use of external knowledge in ReClor [19]. However, these datasets are designed only to measure system performance, not for comparison with human responses. As a result, human responses to RC are absent from these datasets. There is limited research that compares responses from machines and humans, and even these studies typically offer only summarized data [20]. This data shortage has hindered research into machine emulation of human response.

In contrast to this scarcity, CMCQRD [12] is a unique RC dataset which includes response data from language learners. CMCQRD adopts a multiple-choice setting like many of the RC datasets mentioned above, and includes the distribution of the choices among options. RC tests and participants are categorized based on the CEFR which is a guideline used to describe achievements of foreign language learners. Among the six reference levels (A1, A2, B1, B2, C1, C2) of the CEFR, independent- (B1, B2) and proficient-level (C1, C2) are considered in the CMCQRD dataset. In other words, each question in this dataset is labeled with a difficulty level ranging from B1 to C2 according to the CEFR, and also includes the selection distribution by language learners whose proficiency corresponds to these labeled levels. This information enables a detailed analysis of the differences between language learners and machines. Liusie et al. [13] compared outputs from an ELECTRA-based classification model with human responses, reporting low similarity due to the model performing worse than language learners.

2.2. Prompts that Alter LLMs' Behaviour in Question Answering

Retrieving distributions for multiple-choice questions from LLMs involves obtaining not only the final answer but also the probabilities associated with each option. While it is nontrivial to extract an answer or a probability because of the auto-regressive nature of text generation by LLMs, Robinson et al. [21] demonstrated that a multiple-choice prompt can lead to a higher probability of generating option symbols as the next token, especially with one or few-shot settings. Unlike a traditional cloze prompt, which selects the option with the highest sequence's probability without giving other options, a multiple-choice prompt provides all options simultaneously and selects the one with the highest probability for the option symbols.

However, even in this setting, it has been reported that LLMs respond less robustly to certain prompts [22, 23]. Utilizing this vulnerability, Santurkar et al. [10] suggested that LLMs can change the distributions of attitude options towards controversial social topics, when given prompts that mimic the behavior of a human group with specific attributes. LLMs' behaviour will also change

when given a degree of certainty like *"Perhaps it's"* [24]. This change was observed in response to context-free open-ended questions, highlighting an opportunity for extended research in multiple-choice RC tests.

3. Experimental Setup

The primary objective of this work is to investigate whether LLMs can mimic the responses of language learners in solving multiple-choice RC tests. In this section, we outline our experimental setup, utilizing the CMCQRD dataset [12], which includes responses from at least 100 language learners per question, providing information about answer probability distributions. Our analysis compares the next-token probability on each option by LLMs with the choice patterns of language learners, aiming to understand the extent of LLMs' capability in emulating learner-like understanding in RC tasks.

Assuming that up-to-date LLMs outperform average language learners, degrading these models is needed to bring their output distributions closer to those of language learners. We employ several methods to weaken the LLM performance and compare the results to the language learners.

Dataset The CMCQRD dataset consists of 4-choice English RC tests, labeled with difficulty levels ranging from CEFR B1 to C2. A subset of CMCQRD includes responses from non-native English speakers whose proficiency aligns with the difficulty label [12, 13]. We refer to this set of responses as the human distribution.

Table 1 shows the statistics of the CMCQRD dataset. The average accuracies of language learners are around 60%, while the accuracies of their mode selections are around 90%. In this experiment, we exclusively use questions at levels B1 and B2 with a human distribution, corresponding to intermediate levels of proficiency. Our focus on these levels is driven by our aim to assess the ability of LLMs to reproduce the challenges faced by language learners who are not fully proficient in reading comprehension.

LLM Settings Since the outputs of LLMs are auto-regressive and free-form, some techniques are required to increase the likelihood of desired tokens in subsequent outputs. To this end, we employ a multiple-choice prompting approach for RC, as described in Robinson et al. [21]. This approach provides LLMs with a single natural language prompt that concatenates the context, a question, options, and an option-symbol-prompting word, such as *"Answer:"*. We take advantage of the next-token probabilities to the distribution by LLM. The logits of next tokens associated with option symbols, {*A, B, C, D*} on 4-choice tests, are normalized using softmax.

Table 1
Statistics of CMCQRD dataset. We use RC tests at B1 and B2 levels with responses.

	w/o responses		w/ responses			
CEFR Level	Num Text	Num QA	Num Text	Num QA	Mode Acc	Avg Acc
B1	5	25	23	115	0.913	0.590
B2	21	160	37	262	0.882	0.594
C1	13	86	12	83	0.880	0.613
C2	3	20	6	42	0.833	0.681

We adopt GPT-4o[1] and LLaMa-2-70B [25] with one-shot prompting. We run LLaMa-2-70B using the HuggingFace library with 4-bit quantization.[2] The temperature parameter is set to 1.0 for both models.

Evaluation To compare human and LLM outputs, we use mode accuracy, average accuracy, and KL divergence following Liusie et al. [13], and also correct/wrong F1 score. Below is the description of these metrics.

1. *Mode Accuracy*: how frequently the most plausible symbol by LLM is the correct answer, denoted as

$$\text{Mode Accuracy} = \mathbb{E}[\text{argmax}_y(p^{\text{LLM}}) = y^{\text{ans}}],$$

   where p represents probabilities for each option and y^{ans} is the correct option.
2. *Average Accuracy*: how frequently the correct option is selected on average by LLM, denoted as

$$\text{Average Accuracy} = \mathbb{E}[y^{\text{LLM}} = y^{\text{ans}}].$$

3. *KL Divergence*: the similarity between two distributions [26], denoted as

$$\text{KL Divergence} = \sum_o l_o \log \frac{l_o}{h_o},$$

   where o represents an option selection, with the LLM and human distribution fixed to l and h, respectively.
4. *Correct/Wrong F1*: the macro-averaged f1 score focused on question-wise correct and wrong consistency on mode options, denoted as

$$\text{Correct/Wrong F1} = \frac{1}{2}(\text{F1}_{\text{correct}} + \text{F1}_{\text{wrong}}),$$

   where each F1 score is calculated based on the elements of confusion matrix, such as $\text{TP}_{\text{correct}} = \sum_i[(y_i^{LLM} = y_i^{ans}) \wedge (y_i^{Human} = y_i^{ans})]$ and $\text{FP}_{\text{wrong}} = \sum_i[(y_i^{LLM} \neq y_i^{ans}) \wedge (y_i^{Human} = y_i^{ans})]$.

Furthermore, we calculate the summation of the probabilities for option symbols appearing as the next token to evaluate the effectiveness of the prompts.

[1] https://openai.com/index/hello-gpt-4o/
[2] https://huggingface.co/meta-llama/Llama-2-70b-hf

Table 2
Result on CMCQRD Dataset. Values on KL and C/W F1 are those compared to Human language learners above.

		B1					B2				
System	Prompt	Mode Acc	Avg Acc	KL↓	C/W F1↑	Sum Prob.	Mode Acc	Avg Acc	KL↓	C/W F1↑	Sum Prob.
Human	-	0.913	0.585	-	-	-	0.885	0.592	-	-	-
GPT-4o	NONE	0.974	0.974	0.570	0.552	0.994	0.931	0.929	0.576	0.633	0.971
	PORTRAY	0.974	0.971	0.566	0.552	0.988	0.927	0.927	0.580	0.606	0.975
	ESL	0.965	0.964	0.563	0.544	0.895	0.927	0.926	0.554	0.651	0.842
	UNCERTAIN	0.713	0.719	0.795	0.471	0.155	0.828	0.805	0.711	0.572	0.228
	MASK	0.922	0.918	0.562	0.512	0.868	0.851	0.852	0.578	0.608	0.798
LLaMa-2-70b	NONE	0.930	0.839	0.338	0.518	0.993	0.854	0.756	0.354	0.611	0.992
	PORTRAY	0.930	0.831	0.320	0.518	0.984	0.847	0.740	0.332	0.604	0.980
	ESL	0.922	0.750	0.211	0.512	0.973	0.851	0.674	0.263	0.658	0.969
	UNCERTAIN	0.922	0.646	0.163	0.512	0.966	0.839	0.556	0.226	0.646	0.971
	MASK	0.843	0.750	0.294	0.553	0.988	0.755	0.644	0.391	0.533	0.983

Prompt Design We employ 4 types of prompt designs below. See Appendix A for the examples.

- NONE: Only the context, question, and candidate answers are given.
- PORTRAY: Similar to Santurkar et al. [10], a role is assigned at the beginning of the prompt, for example, *"Answer the following reading comprehension question as if you are a CEFR B1 level English learner."*, followed by a description of the level defined by CEFR. [3]
- ESL: Bonner et al. [27] suggested that LLMs seem to have the ability to control outputs based on a targeted CEFR level provided in a prompt. We ask LLMs the most plausible answer from language learners at a specific CEFR level, such as *"What do you think is the most plausible answer by CEFR B1 level learners to the following reading comprehension test?"*. In addition, we inject the explanation like *"Given the context and considering that the test takers are at a CEFR B1 level, the most plausible answer they might choolse could be"* after *"ANSWER:"*.
- UNCERTAIN: as reported in Zhou et al. [24], the expression of uncertainty will change LLMs' behavior. We inject the expression like *"I'm not sure because there are some sentences I don't understand, but maybe the answer is,"* after *"ANSWER:"*.
- MASK: Laufer [28] argued that language learners need to know 95% of the vocabulary in a text to comprehend its content. To simulate the scenario where 5% of the vocabulary are not known, top 5% unfrequent words within a context are masked. Unfrequent words in question and options are also masked based on this threshold. The word frequency is calculated based on SUBTLEXus [29].

[3] https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-descriptors

4. Results

Table 2 shows the performance of LLMs on CMCQRD given each prompt. Overall, contrary to our expectations, the results reveal the limited ability of LLMs to mimic language learners when solving multiple-choice RC tests.

LLMs tend not to be distracted. First, the distributions by LLMs, especially from GPT-4o, show more skewness in NONE compared to those from humans. In other words, compared to the small gap between Human and the LLM in the mode accuracy, the average accuracy sees much a wider gap. For GPT-4o, there is almost no difference between these accuracies, which demonstrates that the most plausible next token is only one option symbol regardless of its correctness.

Prompts affect outputs differently across LLMs. The results show the difference in the function of prompts between GPT-4o and LLaMa-2-70b. For LLaMa-2-70b, the sum of the probabilities for option symbols exceeds 95% across all prompts, indicating that the prompts effectively induce the generation of these symbols. On the other hand, GPT-4o behaves differently, particularly with UNCERTAIN prompt, where the probability of generating non-symbol tokens is considerable. This shows that the function of prompts differs across LLMs.

LLaMa-2-70b is better than GPT-4o in weakening. A key distinction between responses from language learners and LLMs is that while both show high Mode Accuracy, LLMs demonstrate substantially higher Average

Accuracy compared to humans, indicating that distributions by LLMs are generally skewed. Therefore, an LLM suited for weakening can maintain Mode Accuracy while reducing Average Accuracy. In this aspect, LLaMa-2-70b is better than GPT-4o. GPT-4o shows minimal changes in Average Accuracy even with weakening prompts, including UNCERTAIN that drops accuracies and Sum Probability. Thus, its distributions remain distinct from language learners, as reflected by the persistently high KL divergence. In contrast, LLaMa-2-70b shows the ability to reduce Average Accuracy while maintaining Mode Accuracy, especially with ESL and UNCERTAIN prompts.

Prompt design plays a crucial role. Prompt designs markedly influence the outputs from LLMs, as exemplified by the difference between PORTRAY and ESL results on LLaMa-2-70b. While both prompts are designed to emulate language learner-like outputs and include the description of the targeted CEFR level, PORTRAY fails to weaken performance, whereas ESL leads to reductions in Average Accuracy and KL Divergence. This suggests that there is much room for prompt engineering in other designs, including UNCERTAIN.

Language Learners and LLM mistake different questions. Whereas KL divergence measures the similarity between two distributions, Correct/Wrong F1 score directly measures the consistency of most plausible answers by humans and LLMs. LLMs show a low F1 score regardless of the prompt given, indicating a discrepancy between questions that lead to human errors and those that lead to LLM errors. LLaMa-2-70b observes the largest drop in KL divergence with UNCERTAIN prompt compared to NONE. However, this does not correspond with a substantial improvement in the F1 score, suggesting that the LLM does not mimic human error patterns effectively. Since distributions by LLMs are generally skewed compared to those by language learners, the reduction of KL divergence is achievable by simply increasing the temperature parameter. This result reveals the importance of not only comparing distributions but also examining the consistency of the mode answers to mimic humans.

5. Discussion

Our results so far seem to demonstrate the inability of LLMs to mimic human language learners when solving RC tests, even when provided with weakening prompts. In particular, we identify differences in the questions that language learners and the LLMs tend to answer incorrectly. In this section, we turn our attention to an analysis of the underlying factors for these discrepancies.

We analyze the influence of the complexity of context on accuracy gaps between language learners and LLM (NONE-Human), and also the gaps between the LLM with and without a weakening prompt (NONE-UNCERTAIN). We select LLaMa-2-70b because of its ability to be weakened. Among the features used in prior research by Sugawara et al. [20], we select Passage Length, FKGL[30], and Word Frequency as indicators of complexity. Correlations are measured between these indicators and the accuracy gaps for each individual question.

Table 3
Correlation between the gap and complexity measures. N, H, and U mean NONE, Human, and UNCERTAIN, respectively. * means statistical significance on $p < 0.05$.

		N-H	N-U	Avg
Δ Average Accuracy	B1	0.254	0.193	-
	B2	0.164	0.200	-
Passage Length	B1	-0.14	0.01	342.2
	B2	0.14*	0.04	656.7
FKGL	B1	0.31*	0.23*	9.69
	B2	0.05	-0.02	9.22
Word Freq (per 1k words)	B1	-0.18	-0.08	6.53
	B2	-0.01	0.13*	6.44

Table 3 shows the correlations, some of which are statistically significant. For Passage Length, there is a weak positive correlation with the gap between NONE and Human at the B2 level, which means that the longer the context, the harder it is for language learners to answer correctly compared to the LLM. This implies that a longer context may hinder B2 level language learners from finding the evidence needed to answer more than it does the LLM. FKGL, a readability metric based on the number of words and syllables per sentence, shows a weak-to-moderate positive correlation with the gap between LLM and human, and also the gap between LLMs with and without uncertainty prompt. Since FKGL is designed to show a lower value on easier texts, these statistically significant gaps imply that the LLM shows a higher accuracy in more complex contexts. UNCERTAIN prompt can slightly smooth this trend, but it does not enable the LLM to emulate the tendency of language learners. Finally, for Word Freqency, there is a weak positive correlation with gap between NONE and UNCERTAIN at B2 level. This may imply that UNCERTAIN weaken LLMs more when a context is composed of more common words.

Overall, these surface-level complexity indicators are not sufficient to explain the difference between language learners and LLMs. We reserve deeper analysis, such as semantic considerations, for our further research.

6. Conclusion

In conclusion, our research reveals that LLMs does not behave as second language learners even with potentially

performance-weakening prompts we provide. We also observe that the performance varies depending on the model and prompts used, even though a limited set of models and prompts are considered. Expanding the variety of these elements, including prompts with more sophisticated approaches such as chain-of-thought [23] and automatic prompt tuning [31], will be critical for a more comprehensive evaluation of the mimicability.

Our findings demonstrate that discrepancies between language learners and LLMs in terms of easiness of questions, highlighting the necessity for micro-level analysis. Nonetheless, the limited size of CMCQRD dataset used in this research presents challenges in drawing comprehensive conclusions. The development of datasets incorporating diverse personas beyond language learners is essential when trying to use LLMs as the complement of human evaluators.

Acknowledgments

The authors acknowledge the support from Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021) and from Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MCIN/AEI /10.13039/501100011033.

References

[1] S. Gehrmann, E. Clark, T. Sellam, Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text, 2022. arXiv:2202.06935.

[2] N. Grabar, H. Saggion, Evaluation of automatic text simplification: Where are we now, where should we go from here, in: Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, ATALA, Avignon, France, 2022, pp. 453–463. URL: https://aclanthology.org/2022.jeptalnrecital-taln.47.

[3] L. Martin, A. Fan, Éric de la Clergerie, A. Bordes, B. Sagot, Muss: Multilingual unsupervised sentence simplification by mining paraphrases, 2021. arXiv:2005.00352.

[4] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations, 2020. arXiv:2005.00481.

[5] A. Sauberli, F. Holzknecht, P. Haller, S. Deilen, L. Schiffl, S. Hansen-Schirra, S. Ebling, Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities, 2024. arXiv:2402.13094.

[6] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences 120 (2023) e2305016120.

[7] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. arXiv:2303.16634.

[8] D. Dillion, N. Tandon, Y. Gu, K. Gray, Can ai language models replace human participants?, Trends in Cognitive Sciences 27 (2023) 597–600.

[9] E. Hwang, B. P. Majumder, N. Tandon, Aligning language models to user opinions, 2023. arXiv:2305.14929.

[10] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, T. Hashimoto, Whose opinions do language models reflect?, 2023. arXiv:2303.17548.

[11] S. Liu, X. Zhang, S. Zhang, H. Wang, W. Zhang, Neural machine reading comprehension: Methods and trends, Applied Sciences 9 (2019) 3698.

[12] A. Mullooly, O. Andersen, L. Benedetto, P. Buttery, A. Caines, M. J. F. Gales, Y. Karatay, K. Knill, A. Liusie, V. Raina, S. Taslimipoor, The Cambridge Multiple-Choice Questions Reading Dataset, Cambridge University Press and Assessment, 2023. URL: https://www.repository.cam.ac.uk/handle/1810/358683. doi:10.17863/CAM.102185.

[13] A. Liusie, V. Raina, A. Mullooly, K. Knill, M. J. F. Gales, Analysis of the cambridge multiple-choice questions reading dataset with a focus on candidate response distribution, 2023. arXiv:2306.13047.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[15] E. H. Jeon, J. Yamashita, L2 reading comprehension and its correlates: A meta-analysis, Language learning 64 (2014) 160–212.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, USA, 2016, pp. 2383–2392.

[17] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, Race: Large-scale reading comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, USA, 2017, pp. 785–794.

[18] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, Hotpotqa: A dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,

USA, 2018, pp. 2369–2380.
[19] W. Yu, Z. Jiang, Y. Dong, J. Feng, Reclor: A reading comprehension dataset requiring logical reasoning, in: International Conference on Learning Representations, International Conference on Learning Representations, USA, 2019.
[20] S. Sugawara, N. Nangia, A. Warstadt, S. Bowman, What makes reading comprehension questions difficult?, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, USA, 2022, pp. 6951–6971.
[21] J. Robinson, C. M. Rytting, D. Wingate, Leveraging large language models for multiple choice question answering, 2023. arXiv:2210.12353.
[22] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438. URL: https://aclanthology.org/2020.tacl-1.28. doi:10.1162/tacl_a_00324.
[23] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.
[24] K. Zhou, D. Jurafsky, T. Hashimoto, Navigating the grey area: How expressions of uncertainty and overconfidence affect language models, 2023. arXiv:2302.13439.
[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.
[26] T. M. Cover, Elements of information theory, John Wiley & Sons, USA, 1999.
[27] E. Bonner, R. Lege, E. Frazier, Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching., Teaching English with Technology 23 (2023) 23–41.
[28] B. Laufer, What percentage of text-lexis is essential for comprehension?, Special language: From humans thinking to thinking machines (1989) 316.
[29] M. Brysbaert, B. New, Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english, Behavior research methods 41 (2009) 977–990.
[30] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, 1975.
[31] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, Y. Yang, Connecting large language models with evolutionary algorithms yields powerful prompt optimizers, 2024. URL: https://arxiv.org/abs/2309.08532. arXiv:2309.08532.

A. Prompt Examples

Table 4
Examples of designed prompts.

NONE
CONTEXT: I won't pretend being a flight attendant is easy. But since I started the job, I've been everywhere, from the US to Australia. I work with incredible people, I have a lot of time off, and life is never boring - which ... *QUESTION: What does Jack say about attending his job interview?* *A) He was surprised at the age range of people there.* *B) He made sure he seemed different from the others.* *C) He wondered whether he had enough qualifications.* *D) He realised there were too many people for the jobs available.* *ANSWER:*\n
PORTRAY
Answer the following reading comprehension questions as if you are a CEFR B1 level English learner. Learners at this level can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. But sometimes it may be difficult to understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. {Same as NONE from *CONTEXT:* to *ANSWER:*\n}
ESL
You are an ESL teacher. What do you think is the most plausible answer by CEFR B1 level learners to the following reading comprehension test? Learners at this level can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. But sometimes it may be difficult to understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. {Same as NONE from *CONTEXT:* to *D) he ...*} *ANSWER:* *Given the context and considering that the test takers are at a CEFR B1 level, the most plausible answer they might choose could be:*\n
UNCERTAIN
{Same as NONE from *CONTEXT:* to *D) he ...*} *ANSWER:* *I'm not sure because there are some sentences I don't understand, but maybe the answer is:*\n
MASK
CONTEXT: I won't [MASK] being a flight [MASK] is easy. But since I started the job, I've been everywhere, from the US to Australia. I work with incredible people, I have a lot of time off, and life is never [MASK] - which ... *QUESTION: What does Jack say about attending his job interview?* *A) He was surprised at the age range of people there.* *B) He made sure he seemed different from the others.* *C) He [MASK] whether he had enough qualifications.* *D) He realised there were too many people for the jobs available.* *ANSWER:*\n