Improve Performance of Fine-tuning Language Models with Prompting

Zijian Győző Yang, *Member, HTE* and Noémi Ligeti-Nagy

Abstract—**This paper explores the effectiveness of prompt programming in the fine-tuning process of a Hungarian language model. The study builds on the prior success of prompt engineering in natural language processing tasks and employs the prompting method to enhance the fine-tuning performance of a huBERT model on several benchmark datasets of HuLU. The experimentation involves testing 45 prompt combinations for the HuCoPA dataset and 15 prompt variations for the HuRTE and HuWNLI datasets. The findings reveal that the addition of an instructional text consistently produces the best results across all winning cases, and that the [CLS] token produces the best results in the separator token experiments. The most significant enhancement was observed in the HuWNLI dataset, with an increase in accuracy from 65% to 85%. These results demon- strate that the addition of instruct text is crucial and sufficient in enabling the language model to effectively interpret and solve the Winograd Schemata problem. These results showcase the potential of prompt programming in enhancing the performance of language models in fine-tuning tasks, and highlight the importance of incorporating task-specific instructions to improve model interpretability and accuracy.**

Index Terms—**BERT, prompting, fine-tuning**

I. Introduction

PROMPTING is a technique used to guide language models in generating specific types of language. With prompting, a user provides a starting point or “prompt” for the language model, and the model generates text that continues from that point. Prompting can be used to control the topic, style, tone, and other aspects of the generated text.

There are several types of prompting techniques that have been studied in the context of language models. One common technique is prefix-based prompting, where the user provides a few words or a sentence as the starting point for the generated text. Another technique is conditional prompting, where the user specifies a condition or constraint that the generated text must satisfy, such as a certain topic or sentiment. Additionally, prompting can be done through natural language prompts, multiple-choice prompts, or other forms of input.

Prompting can be a powerful tool for controlling the output of language models and making them more useful for specific applications. However, there are also challenges associated with prompting. One challenge is designing effective prompts that achieve the desired result. Another challenge is understanding how the language model is interpreting the prompt and generating the resulting text.

There is a growing body of research on prompting in the context of language models [1]. Some studies have focused on improving the effectiveness of prompting [2], [3], while others have explored the ethical implications of using prompts to control the output of language models [4]. Overall, prompting is an important area of research in the field of natural language processing, and it has the potential to shape the future of human-computer interaction and content generation.

Drawing inspiration from the success of prompt engineering, the present study adapted this approach to enhance the fine-tuning performance of a Hungarian language model and associated benchmarks, employing the prompting method. By doing so, our experimentation aimed to build on the prior success of this method in natural language processing tasks.

The currently most performant language model for the Hungarian language is huBERT [5]. Despite being a smaller model, huBERT outperforms HILBERT [6] in available tests, likely due to being trained on more data. Yang et al. [7] have recently introduced three large models trained on large amount of Hungarian data (PULI GPT-3SX, PULI GPT-2, PULI BERT-Large) and evaluated all the above mentioned models on the datasets of HuLU, the Hungarian Language Understanding Benchmark Kit [8], [9].

HuLU was created on the basis of the GLUE [10] and SuperGLUE [11] benchmark databases. The main purpose of HuLU is to enable a standardized evaluation of neural language models in a simple way while also enabling multi-perspective analysis. It also compares the performance of various language models on various tasks. The HuLU comprises 7 corpora containing annotation for various standard language comprehension tasks. As usual, these corpora are divided into training, validation and test sets. The subcorpora of HuLU are either translated datasets (Hungarian Choice of Plausible Alternatives Corpus – translated from CoPA [12] –, Hungarian Recognizing Textual Entailment dataset – translated from the RTE1, RTE2, RTE3 and RTE5 datasets [13], [14], [15], [16] –, Hungarian version of the Stanford Sentiment Treebank – sentences translated from the SST5 dataset [17] –, Anaphora resolution datasets for Hungarian as an inference task [18] – the examples translated from the Winograd schemata and the WNLI dataset [19], [10]) or datasets created from scratch the design of which follows some English datasets (Hungarian CommitmentBank Corpus – designed based on Commitment-Bank [20] –, Hungarian Corpus of Linguistic Acceptability – designed based on COLA [21] –, Hungarian Corpus for Reading Comprehension with Commonsense Reasoning [22] – designed based on ReCoRD [23]).

The primary objective of our research is to harness the

Zijian Győző Yang and Noémi Ligeti-Nagy: Language Technology Research Group, Institute for Language Technology and Applied Linguistics, Hungarian Research Centre for Linguistics, Budapest, Hungary.
E-mail: {yang.zijian.gyozo, ligeti-nagy.noemi}@nytud.hu,

Manuscript received April 19, 2005; revised August 26, 2015.

DOI: 10.36244/ICJ.2023.5.10

power of prompt programming to refine the performance of a Hungarian language model, a step forward that has not been fully explored yet. We chose this area given the promise that prompt engineering holds in enhancing the interpretability and accuracy of language models, especially when dealing with complex language problems. We consider this research vital as it stands to bridge a significant gap in language modeling, potentially setting a new standard for fine-tuning processes in language model development across multiple languages, thus paving the way for better language processing applications.

The structure of our paper is as follows: Section I provides a brief background, discusses previous solutions, and outlines the structure of this paper. In Section II, we present the data and methods used. In Section III, a detailed description of our prompt experiments is provided. Section IV contains the results and evaluations, and Section V presents a concise conclusion.

II. Data and Methods

A. Datasets

In our experiment, we conducted fine-tuning of a huBERT model on several benchmark datasets of HuLU. Our research encompassed experimentation on the HuCoPA, HuRTE, and HuWNLI datasets. The HuCoPA dataset [24] comprises 1,000 instances, each consisting of a premise and two alternatives. The task involves selecting the alternative that describes a situation that stands in a causal relation to the situation described by the premise (example (1)). The train, validation and test sets contain 400, 100 and 500 instances, respectively, following the splits of the original English dataset (as in the GLUE benchmark).

HuRTE [25] is the Hungarian version of the Recognizing Textual Entailment dataset of GLUE, comprising 4,504 instances. Each instance contains a sometimes multi-sentence premise and a one-sentence hypothesis, and the task is to determine whether the former entails the latter or not. The task is a binary classification problem (see example (2)). The train, validation and test sets contain 2 131, 242 and 2 131 instances, respectively.

The HuWNLI dataset [26] comprises the collection of the Hungarian Winograd Schemata [27], extended with the set of sentence pairs of the test set of the WNLI dataset of GLUE, and transformed into a natural language inference task. The NLI format was created by replacing the ambiguous pronoun with each possible referent (see example (3)). The data is distributed among three splits: training set (562), validation set (59) and test set (134).

(1) premise: *A testem árnyékot vetett a fűre.* 'My body cast a shadow over the grass.'
choice 1: *Felkelt a nap.* 'The sun was rising.'
choice 2: *A füvet lenyírták.* 'The grass was cut.'
question: *cause*
label: 1 (the number of the more plausible choice)

(2) premise: *Még nem találtak tömegpusztító fegyvereket Irakban.* 'No weapons of mass destruction have yet been found in Iraq.'
hypothesis: *Tömegpusztító fegyvereket találtak Irakban.* 'Weapons of mass destruction have been found in Iraq.'
label: 0 (1, if the premise entails the hypothesis, 0 otherwise.)

(3) sentence 1: *A férfi nem tudta felemelni a fiát, mert olyan nehéz volt.* 'The man couldn't lift his son because he was so heavy.',
sentence 2: *A fia nehéz volt.* 'His son was heavy.',
label: 1 (1, if sentence 1 entails sentence 2, 0 otherwise.)

B. Fine-tuning process

In our fine-tuning process, we employed identical hyperparameter settings across all cases, and fine-tuned all models for a period of 20 epochs. Our comparison was based on selecting the highest result scores. For the experiments, we used 2 NVIDIA A100 GPUs. The modified hyperparameters are as follows:

- HuCoPA: sequence length: 128; batch size: 8 per GPU; learning rate: 2e-5;
- HuRTE: sequence length: 512; batch size: 32 per GPU; learning rate: 2e-5;
- HuWNLI: sequence length: 256; batch size: 6 per GPU; learning rate: 8e-6.

For fine-tuning our language model, we used the scripts provided by Hugging Face [28]. In the case of HuCoPA, we treated the task as a multiple choice task, while for HuRTE and HuWNLI, we employed the text classification script to train our models. Initially, we used a learning rate value of 2e-5 for all cases. However, further experimentation with HuWNLI revealed that 8e-6 yielded the best results and thus became the preferred choice.

In our experiments, we fine-tuned our models on the training set. Subsequently, we conducted experiments on the validation set and selected the checkpoint that yielded the highest results. Finally, we evaluated this selected checkpoint on the test sets.

III. Prompt experiments

In our current research, we have explored various possible prompt templates, ranging from not using prompts at all, to adding only a separator token between the sentences, and even to utilizing complex prompt templates with multiple sentence-long instructions. In the case of using prompts, we explored several versions of the separator token or text. When using text as a separator, there were instances where it was necessary to modify the syntax of the input sentence, such as converting the original sentence to lowercase. We even experimented with multiple instruction texts, which contain a detailed description of the current task to be solved. All the instructions were in Hungarian. Some examples of the different types of prompts used with the HuCoPA dataset is provided below for your reference:

- Original text from HuCoPA:
  - premise: *A sofőr felkapcsolta az autó fényszóróit.* ’The driver turned on the car’s headlights.’
  - choice 1: *Mennydörgést hallott.* ’He/she heard a thunderclap.’
  - question: cause
- Input using a separator token: *A sofőr felkapcsolta az autó fényszóróit.* **[SEP]** *Mennydörgést hallott.* – ’The driver turned on the car’s headlights. **[SEP]** He/she heard a thunderclap.’
- Input using text as separator: *A sofőr felkapcsolta az autó fényszóróit.* ***Mert*** *mennydörgést hallott.* – ’The driver turned on the car’s headlights. **Because** he/she heard a thunderclap.’
- Input using an instruct text: ***Döntsd el, hogy következik-e az első mondat a második mondatból. Első mondat:*** *A sofőr felkapcsolta az autó fényszóróit.* ***Második mondat:*** *Mennydörgést hallott.* – ’**Decide whether the first sentence is entailed by the second one. First sentence:** The driver turned on the car’s headlights. **Second sentence:** He/she heard a thunderclap.’

In the samples above, the text/token that is added to the original input text is bolded. Using this additional information or these instructions, we can assist the language models in their fine-tuning training. In some cases, when using text as a separator (as you can see above), we had to modify the syntax of the input sentence.

We conducted experiments in all cases using the following category of prompts:

- $[empty]$: The examined texts were concatenated without using any separator token.
- *Separator token*: Either/both the $[CLS]$ or $[SEP]$ token was inserted as a separator between the two examined texts.
- *Conjunction phrase*: Hungarian conjunction word/phrase was employed as a separator token text (details can be found in Table I).
- *Question sentence*: A question sentence was used as a separator token text (see Table I).
- *Instruct text*: Instruct text was added to the beginning of the input text (see Table II). The *question sentence* can be an instruct text as well.
- *Mix*: Different prompt types were mixed, by combining the use of a separator token with the instruct text, as an example.

In Table I and Table II we provide a comprehensive list of the various prompt texts we experimented with. Specifically, for the HuCoPA dataset, we tested 45 different prompt combinations, whereas for the HuRTE and HuWNLI datasets, we tested 15 prompt variations.

IV. Results

In the results section, we have chosen to present only the best scores obtained from each prompt category to enhance the readability of our findings. This decision was made in consideration of the 75 experiment subscores that were obtained, which could otherwise result in excessive data presentation that may obscure the key insights. Figure 1 displays the highest results attained for each category on the test sets. Our experimentation showed that in all corpora, the highest results were achieved through the combination of prompts. Notably, we achieved state-of-the-art results in all three examined benchmarks. To further validate the effectiveness of our prompting method, we submitted our results to the HuLU benchmark competition [29], where our approach outperformed the dedicated three benchmarks, as shown in Table III. The three mixed prompt winners are listed below, the original input texts are marked ({ ... }) as variables:

TABLE I
Conjunction phrase and question sentence as prompts in the case of the different datasets

	Conjunction phrase	Question sentence
HuCoPA	*mert/ezért* *Mert/Ezért* ’because/because of this’	*Oka?/Hatása?* ’Cause of this?/Result of this?’ *Mi az oka?/Mi a hatása?* ’What is the cause of this?/ What is the result of this?’ *Ez a következtetés helyes?* ’Is this conclusion correct?’
HuRTE	*Tehát* ’Therefore’ *Ezért* ’Because of this’ *Ebből következik, hogy* ’This implies that’	*Ez a következtetés helyes?* ’Is this conclusion correct?’
HuWNLI	*Tehát* ’Therefore’ *Ezért* ’Because of this’ *Ebből következik, hogy* ’This implies that’	*Ez a következtetés helyes?* ’Is this conclusion correct?’

- HuCoPA: *Ez a következtetés helyes?* ’Is this conclusion correct?’ { ... (premise text) } *Mert/Ezért* ’Because/Because of this’ { ... (choice sentence text) }
  (To make the sentence grammatically correct, the choice sentence is lowercased.)
- HuRTE: *A következő példákban egy premissza és egy hipotézis található. A premissza több mondatból is állhat. A feladat az, hogy el kell dönteni, a hipotézis következik-e a premisszából: azaz ha a premissza igaz, akkor a hipotézis is igaz.* ’The following examples consist of a premise and a hypothesis. The premise may consist of multiple sentences. The task is to determine whether the hypothesis is entailed by the premise: that is, if the premise is true, then the hypothesis is also true.’ [CLS] *premissza:* ’premise’ { ... (premise text) } [CLS] *hipotézis:* ’hypothesis’ { ... (hypothesis text) }
- HuWNLI: *Az alábbi példákban két mondat látható. El kell dönteni, hogy a második mondat következik-e az első mondatból.* ’The following examples consist of two sentences. The task is to determine whether the second sentence is entailed by the first.’ [CLS] *első mondat:* ’first sentence:’ { ... (sentence1 text) } [CLS] *második mondat:* ’second sentence:’ { ... (sentence2 text) }

As evident from the winning prompts, the addition of an instructional text was consistently observed across all winning cases. This observation aligns with our expectations, as instructional texts typically provide a detailed description of the task at hand, thereby aiding the language models in their training.

TABLE II
INSTRUCT TEXTS USED IN THE EXPERIMENT

	Instruct text
HuCoPA	1.) *Az alábbi példákban van egy mondat, egy kérdés (Ok vagy Hatás), és két lehetséges alternatíva. A feladat az, hogy a két lehetséges alternatíva közül ki kell választani azt, amelyik valószínűbb válasz a kérdésre.* *Mondat: ... Kérdés: ... 1. alternatíva: ... 2. alternatíva: ...* 'The following examples consist of a sentence, a question (Cause or Effect), and two possible alternatives. The task is to select the alternative that is more likely to be the answer to the question. Sentence: ... Question ... 1st alternative: ... 2nd alternative: ...' 2.) *Az alábbi példákban van egy mondat, és két lehetséges folytatás. A feladat az, hogy a két lehetséges folytatás közül ki kell választani azt, amelyik valószínűbb folytatása a mondatnak.* *Mondat: ... 1. folytatás: ... 2. folytatás: ...* 'In the following examples, there is a sentence and two possible continuations. The task is to select the alternative that is more likely to be the continuation of the sentence. Sentence: ... 1st continuation: ... 2nd continuation: ...'
HuRTE	*A következő példákban egy premissza és egy hipotézis található. A premissza több mondatból is állhat. A feladat az, hogy el kell dönteni, a hipotézis következik-e a premisszából: azaz ha a premissza igaz, akkor a hipotézis is igaz.* *Premissza: ... Hipotézis: ...* 'The following examples consist of a premise and a hypothesis. The premise may consist of multiple sentences. The task is to determine whether the hypothesis is entailed by the premise: that is, if the premise is true, then the hypothesis is also true. Premise: ... Hypothesis: ....'
HuWNLI	*Az alábbi példákban két mondat látható. El kell dönteni, hogy a második mondat következik-e az elsőből: azaz ha az első mondat igaz, akkor ebből következik, hogy a második mondat is igaz.* *Első mondat: ... Második mondat: ...* 'The following examples consist of two sentences. The task is to determine whether the second sentence is entailed by the first: that is, if the first sentence is true, then it follows that the second sentence is also true. First sentence: ... Second sentence: ...'

In the separator token experiments, our findings indicate that in all cases, the [CLS] token produced the best results.

The greatest enhancement was attained on the HuWNLI dataset: the results increased from the preceding 65% accuracy to 85%, thereby yielding a markedly superior outcome compared to the previous attempts (see Table III for the comparison of the results on the HuLU datasets). The results indicate that the addition of an instruct text (i.e., the description of the given task) was crucial and sufficient in enabling the language model to effectively interpret and solve the Winograd Schemata problem.

TABLE III
HULU COMPETITION

	HuCoPA (MCC / acc)	HuWNLI (acc)	HuRTE (MCC / acc)
huBERT	56.1 / 78.0	64.93	48.7 / 74.1
PULI BERT-Large	41.4 / 76.6	65.67	51.7 / 75.9
huBERT - Prompt	**56.4 / 78.2**	**85.80**	**53.4 / 76.5**

In Table IV, a snippet of our HuCoPA experiment is presented. The rows represent different prompt sets, while the columns represent epoch numbers (only the first 10 epochs are shown). Upon examining the values in the first epoch, it is evident that the model learned the task at varying speeds depending on the prompt set. The 24th prompt set achieved a precision value of 74.44 in the first epoch, whereas the 17th prompt set only reached 56.99. By the tenth epoch, all models obtained acceptable results; however, there still remains a difference of 7.3 between the highest and lowest values (77.77 – 85.00). An interesting observation is that the 23rd prompt set achieved the highest value in epoch 7.

Due to the nature of our experiments being focused on fine-tuning tasks, we were unable to directly compare our results with large language models and their applications, such as ChatGPT [30].

A. Discussion of the results on the HuWNLI dataset

As highlighted earlier, considerable progress is evident on the HuWNLI dataset. An accuracy of 65% had previously been recorded as the highest, achieved by a BERT-Large model fine-tuned, however, our mixed setting experiment (utilizing an instruct text and the `[CLS]` token) yielded a fine-tuned model with an accuracy score of 85%.

Aside from being an allusion to Turing's imitation game [31], the term ”Turing Test” is broadly applied to any test devised to gauge a computer's ”intelligence”. Winograd schemata are frequently dubbed the new ”Turing Test”. They consist of sentence pairs as closely related in content as possible (with a difference of one word or phrase), having identical target pronouns that refer back to different precursors (example 4).

(4) The city councilmen refused the demonstrators a permit
because they [feared/advocated] violence.
Who [feared/advocated] violence?
a. The city councilmen
b. The demonstrators

Levesque and colleagues [19] suggested a set of Winograd schemata as a fresh AI testing method, inspired by the Turing Test. A Winograd schema must fulfill three conditions to be included in the challenge:

1) it should be easily discernible by a human reader
2) it should not be decipherable by selectional restrictions
3) it should not be searchable on Google

The strength of this new challenge lies in its simplicity: the schemata answer is a binary decision. Furthermore, it's

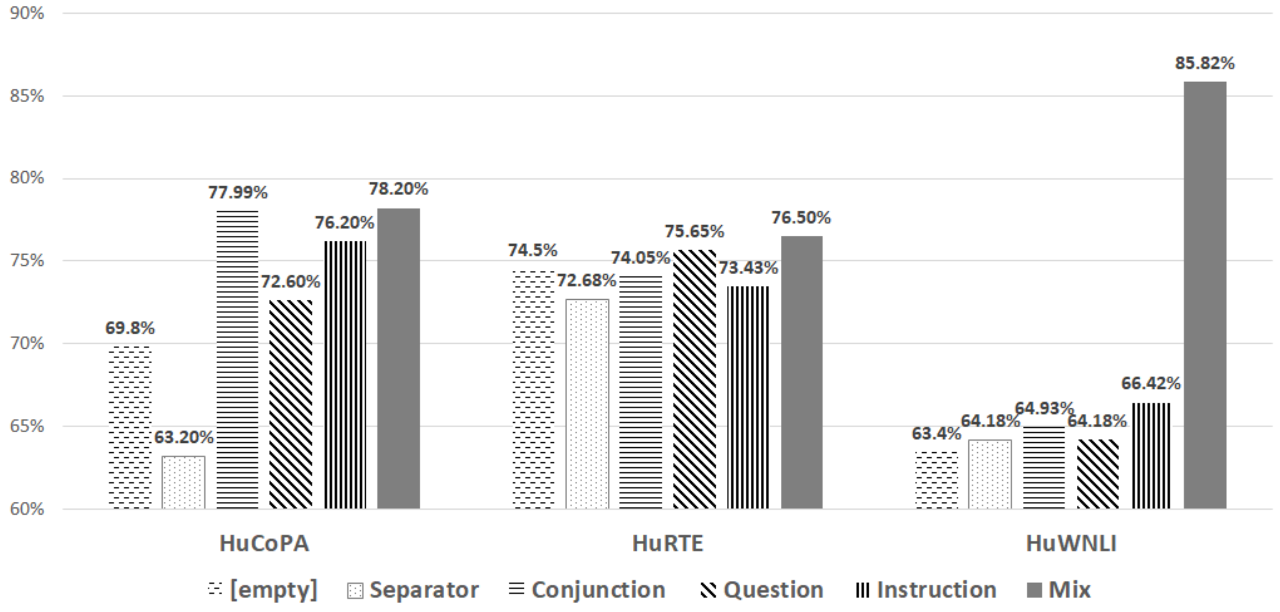

Fig. 1. Performance of the model on three datasets fine-tuned with different prompting strategies

TABLE IV
EXPERIMENTS ON THE HUCOPA VALIDATION SET

id	1 epoch	2 epoch	3 epoch	4 epoch	5 epoch	6 epoch	7 epoch	8 epoch	9 epoch	10 epoch
1	66.55	72.00	79.66	81.77	80.44	80.11	80.33	79.88	80.22	80.22
2	65.55	75.88	81.77	83.77	83.33	82.55	82.99	81.66	81.44	81.66
3	69.99	76.55	81.66	81.99	83.88	82.11	83.66	83.33	82.44	82.77
4	65.44	76.33	82.77	84.55	83.44	84.11	83.66	81.77	81.55	82.11
5	69.11	76.99	82.99	81.77	83.99	82.11	83.55	82.22	82.55	82.55
6	63.88	72.11	78.55	78.66	80.00	78.66	78.66	79.33	79.11	79.22
7	70.44	73.11	80.55	82.88	82.44	80.55	81.33	81.33	81.77	81.33
8	65.22	75.00	79.55	80.44	81.77	79.33	80.88	81.00	80.66	80.77
9	65.77	74.88	80.33	81.99	81.77	81.44	81.00	80.55	80.77	80.77
10	65.88	74.22	81.11	82.88	81.66	81.33	81.88	81.00	81.22	80.66
11	63.33	71.22	77.11	79.77	80.11	80.33	79.55	80.22	80.11	80.11
12	65.33	73.22	76.22	80.33	80.33	78.77	79.00	79.22	79.33	79.22
13	67.44	75.88	80.55	81.33	82.44	82.77	81.77	81.55	81.77	81.99
14	66.66	76.99	78.55	80.66	81.66	81.44	81.55	80.77	80.00	80.33
15	68.66	77.88	81.55	82.99	80.77	82.11	83.33	82.99	81.33	81.77
16	63.44	73.66	79.00	80.66	82.66	81.55	81.99	81.99	80.88	81.00
17	56.99	69.22	70.55	75.55	75.77	77.88	76.99	77.77	77.44	77.77
18	62.33	66.44	72.55	78.11	78.77	79.00	78.11	77.22	77.88	78.00
19	64.77	73.11	76.11	80.55	81.33	79.88	80.33	80.88	80.77	80.44
20	64.11	70.55	79.11	78.55	79.88	81.11	81.00	81.33	79.88	79.77
21	68.11	76.88	80.55	81.77	82.22	81.00	82.66	82.11	81.66	81.99
22	69.11	76.99	84.88	84.88	82.44	82.77	83.44	82.66	83.22	83.33
23	67.77	77.11	83.33	82.99	84.88	84.44	85.11	84.88	84.66	85.00
24	74.44	79.88	84.88	83.33	83.77	84.66	83.77	83.99	84.77	84.78

illuminative: any layperson can deduce that a program that fails to find the right answer lacks sufficient ”intelligence”, i.e., it falls short of human understanding. Lastly, the schemata are demanding: anaphora resolution, while easy for a human, continues to challenge cutting-edge algorithms. This can be attributed to the fact that only world knowledge and reasoning can aid in addressing these issues.

The GLUE and SuperGLUE benchmarks include the WNLI dataset, which features Winograd schemata as sentence pair classification. Here, authors form sentence pairs by substituting the ambiguous pronoun with each possible referent. The task involves predicting whether the sentence, with the pronoun replaced, is implied by the original sentence. In addition, a compact evaluation set composed of new examples taken from fiction books is used alongside the publicly accessible Winograd schemata. This dataset has been shown to be one of GLUE’s most challenging, with ELECTRA first breaking the 90% accuracy barrier in 2019 [32].

Regardless of the impressive accuracy rates that neural models can now achieve on this dataset, commonsense reasoning remains a significant hurdle in AI (for a comprehensive analysis, refer to [33]). Our experiment supports these findings as it emphasizes the significance of the environment where models are fine-tuned and evaluated. This is one of the major criticisms of the Winograd schemata (and other datasets used for evaluating language model performance) highlighted in [33].

As can be seen in Figure 1, we could only approximate the 90% accuracy with the mix setting, all other prompt structures resulted in an accuracy around 65%. The cause of this may lie in the instruction text itself: we narrow the task with this text, aiding the model to focus on the entailment. The `[CLS]` token also assists in setting the boundaries. These two elements – the instruction text and the separator token – appear to be sufficient for the model to pass this test.

Our findings also concur with the aforementioned results, as we observe a substantial leap with the right configuration (prior to ELECTRA's 91.8%, scores on the WNLI dataset in GLUE were hovering between 65-70%). In order to match up with the English results, further enhancements are required. This could be achieved by experimenting with various neural models or by modifying the fine-tuning process and the prompting environment.

V. Conclusion

In this paper, we investigated the effectiveness of different prompting techniques for fine-tuning huBERT on three datasets of HuLU. We experimented with several types of prompts, including conjunction phrases, question sentences, and instruct texts. Our results demonstrate that prompting can significantly improve the performance of huBERT on these datasets.

Overall, our findings suggest that prompt engineering is a promising area of research for improving the performance of language models on specific tasks. By providing targeted prompts that guide the generation of language, we can achieve better results on tasks such as text classification and natural language inference.

In our experiments, we found that the best results were obtained adding instruction text and separator text as prompts. This suggests that combining different types of prompts can be an effective strategy for improving the performance of fine-tuning language models on specific tasks.

Our study has several limitations that should be addressed in future research. For example, we only investigated a limited set of prompting techniques, and there may be other approaches that are even more effective. Additionally, our study only focused on three datasets of HuLU, and it is unclear how well our findings generalize to other datasets and languages.

In the future, we plan to conduct experiments by fine-tuning huBERT using Parameter-Efficient Fine-tuning techniques (such as Lora [34], Prompt tuning [35], etc.). Additionally, we aim to expand our research to include large language models.

In conclusion, our study highlights the potential of prompting techniques for fine-tuning language models on specific tasks. Further research is needed to explore the effectiveness of different prompting strategies, experiments with fuzzy or voting methods [36] and to investigate the generalizability of our findings to other datasets and languages.

References

[1] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023. [Online]. Available: **doi**: 10.1145/3560815

[2] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3045–3059. [Online]. Available: https://aclanthology.org/2021.emnlp-main.243 **doi**: 10.18653/v1/2021.emnlp-main.243

[3] X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. [Online]. Available: https://aclanthology.org/2021.acl-long.353; **doi**: 10.18653/v1/2021.acl-long.353

[4] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel, "Ethical and social risks of harm from Language Models," 2021.

[5] D. M. Nemeskey, "Introducing huBERT," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet, 2021, pp. 3–14.

[6] Á. Feldmann, R. Hajdu, B. Indig, B. Sass, M. Makrai, I. Mittelholcz, D. Halász, Z. Gy. Yang, and T. Váradi, "HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben," in *XVII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet, 2021, pp. 29–36.

[7] Z. Gy. Yang, R. Dodé, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, Á. Kőrös, L. J. Laki, N. Ligeti-Nagy, N. Vadász, and T. Váradi, "Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*. Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Intézet, 2023, pp. 247–262.

[8] N. Ligeti-Nagy, G. Ferenczi, E. Héja, K. Jelencsik-Mátyus, L. J. Laki, N. Vadász, Z. Gy. Yang, and T. Váradi, "HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából," in *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2022, pp. 431–446.

[9] N. Ligeti-Nagy, E. Héja, L. J. Laki, D. Takács, Z. Gy. Yang, and T. Váradi, "Hát te mekkorát nőttél! – A HuLU első életéve új adatbázisokkal és webszolgáltatással," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: JATEPress, 2023, pp. 217–230.

[10] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446, **doi**: 10.18653/v1/W18-5446

[11] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[12] M. Roemmele, C. Bejan, and A. Gordon, "Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning," *AAAI Spring Symposium - Technical Report*, 01 2011.

[13] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL Recognising Textual Entailment Challenge," in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 177–190.

[14] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, "The Second PASCAL Recognising Textual Entailment Challenge," in *Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*, Venice, Italy, 2006, pp. 1–9.

[15] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The Third PASCAL Recognizing Textual Entailment Challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, ser. RTE '07, 2007, pp. 1–9.

[16] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini, "The Fifth PASCAL Recognizing Textual Entailment Challenge," in *Proceedings of the TAC Workshop*, 2009.

[17] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://aclanthology.org/D13-1170

[18] N. Vadász and N. Ligeti-Nagy, "Winograd schemata and other datasets for anaphora resolution in hungarian," *Acta Linguistica Academica*, vol. 69, no. 4, 2022, in press. **DOI**: 10.1556/2062.2022.00575

[19] H. J. Levesque, E. Davis, and L. Morgenstern, "The Winograd Schema Challenge," in *Proceedings of the Thitteenth International Conference on Principles of Knowledge Representation and Reasoning*, ser. KR'12. AAAI Press, 2012, pp. 552–561.

[20] M.-C. de Marneffe, M. Simons, and J. Tonhauser, "The commitmentbank: Investigating projection in naturally occurring discourse," *Proceedings of Sinn und Bedeutung*, vol. 23, no. 2, pp. 107–124, Jul. 2019. [Online]. Available: https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/601

[21] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *arXiv preprint arXiv:1805.12471*, 2018. **DOI**: 10.48550/arXiv.1805.12471

[22] Z. Gy. Yang and N. Ligeti-Nagy, "Building machine reading comprehension model from scratch," *Annales Mathematicae et Informaticae*, pp. 1–17, 2023. [Online]. Available: **DOI**: 10.33039/ami.2023.03.001

[23] S. Zhang, X. Liu, J. Liu, J. Gao, K. Duh, and B. V. Durme, "ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension," 2018.

[24] Hungarian Research Centre for Linguistics, "Hungarian Choice of Plausible Alternatives Corpus." [Online]. Available: https://github.com/nytud/HuCoPA

[25] H. R. C. for Linguistics, "Hungarian Recognizing Textual Entailment dataset." [Online]. Available: https://github.com/nytud/HuRTE

[26] H. R. C. for Linguistics, "Anaphora resolution datasets for Hungarian as an inference task." [Online]. Available: https://github.com/nytud/HuWNLI

[27] N. Vadász and N. Ligeti-Nagy, "Winograd schemata and other datasets for anaphora resolution in Hungarian," *Acta Linguistica Academica*, vol. 69, no. 4, pp. 564–580, 2022. [Online]. Available: https://akjournals.com/view/journals/2062/69/4/article-p564.xml **DOI**: 10.1556/2062.2022.00575

[28] Hugging Face, "Examples." [Online]. Available: https://github.com/huggingface/transformers/tree/main/examples/pytorch

[29] H. R. C. for Linguistics, "Hungarian Language Understanding Benchmark Kit." [Online]. Available: https://hulu.nytud.hu

[30] OpenAI, "Chatgpt." [Online]. Available: https://chat.openai.com

[31] A. Turing, "Computing Machinery and Intelligence," Mind, vol. 59, no. 236, pp. 433–460, 1950.

[32] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=r1xMH1BtvB

[33] V. Kocijan, E. Davis, T. Lukasiewicz, G. Marcus, and L. Morgenstern, "The Defeat of the Winograd Schema Challenge," 2023. **DOI**: 10.1016/j.artint.2023.103971

[34] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[35] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang, "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 61–68. [Online]. Available: https://aclanthology.org/2022.acl-short.8, **DOI**: 10.18653/v1/2022.acl-short.8

[36] T. Tajti, "New voting functions for neural network algorithms," *Annales Mathematicae et Informaticae*, pp. 229–242, 2020. [Online]. Available: **DOI**: 10.33039/ami.2020.10.003

Zijian Győző Yang is research fellow at Hungarian Research Centre for Linguistics. He obtained his PhD degree in Human Language Technology with summa cum laude in 2019. His research areas include large language models, machine translation and evaluation, text summarization, sentiment analysis and text classification.

Noémi Ligeti-Nagy completed her PhD in Computational Linguistics in 2021 and is currently a Research Fellow at the Hungarian Research Centre for Linguistics. Her academic pursuits revolve around the design and development of language corpora, the benchmarking of language processing systems, and the investigation and evaluation of neural language models.