PROMPT TEMPLATE AND METRIC EXPLANATIONS FOR LLM EVALUATIONS

The prompt template for the evaluations of different LLMs is provided as follows. The components that need to be adapted to different problems and solutions are marked with a light gray background. The official solution, along with the final answer, serves as a reference for LLMs to evaluate the student's solution and generate feedback.

Now you play the role of an instructor and need to provide feedback for a student's homework solution. This homework is from a course on circuit analysis. The official solution and student solution are provided in the LaTeX form as follows. The official solution is always correct and can serve as a benchmark for homework assessment. You can provide feedback based on the following aspects. Your feedback should be detailed and precise.

1. Is the student's solution complete? In other words, does the student's solution answer the question?
2. Does the student use the correct method?
3. Are the student's final answers to the problem correct?
NOTE: The correct final answer(s): {FINAL ANSWER GOES HERE}
4. Is there any arithmetic error? Note that the student may use different variable notations from those in the standard solution. These notation differences only should not be regarded as errors, as long as the other parts are correct.
5. Are the units of all variables identified clearly and correctly throughout the calculation process?

[Notes]
You need to consider the following aspects when giving feedback about the student's solution.
1. There might be some typos in the student's solutions. The student can be regarded as solving the problem correctly if all other steps except for the typo are correct.
2. Be careful to check the calculations. The numbers and their signs MUST be correct, and the errors of them cannot be regarded as typos.
3. The rounding errors during the calculation process should not be considered as calculation errors.
4. The equivalency between decimals and fractions should not be regarded as errors.
5. If you think the student's solution is correct, you may just provide concise and brief assessments. When the student's solution is wrong, you need to provide detailed analyses about why the solution is wrong.

[Official Solution]
{OFFICIAL SOLUTION IN LaTeX GOES HERE}

[Student's Solution]
{STUDENT'S SOLUTION IN LaTeX GOES HERE}

The prompt designed above assesses the student's homework in the following metrics, which cover the essential aspects of the homework assessment when the student's solutions were graded by the teaching assistant, who is an author of this paper.

i) *Completeness — The completeness of the student's solution:* This metric evaluates whether the student has completely answered all questions in the problems.

   Justification: Completing circuit analysis problems in full is the minimum requirement for students when they work on assigned homework. When LLMs are presented with both the official solution and the student's solution, applying completeness as a metric specifically prevents LLMs from marking a student's solution as correct if it is only partially complete and the completed portion is accurate. Note that this metric is not intended to assess the correctness of the final answer itself. A student's solution is considered correct as long as it reaches the final step of the solution and provides an answer to the problem.

ii) *Method — The correctness of the method:* This metric evaluates whether the student uses the correct method, regardless of arithmetic errors or typos.

   Justification: A correct solution from a student often implies that they used a reasonable method. However, when a student's solution is incorrect, they may have used either an appropriate or an inappropriate method. Therefore, including a metric for the method in the prompt will help students determine whether they have adopted the correct approach.

iii) *Final Answer — The correctness of the final answers:* This metric evaluates whether the final answers provided by the student are correct.

   Justification: The final answer requires the LLMs to compare the student's final answer with that of the official solution. While the problem-solving process is also important, the final answer, typically brief, is straightforward for LLMs to assess. Including the assessment of final answers enhances the quality of LLM evaluations, as LLMs tend to provide consistent feedback. For example, if a student's final answer does not match the correct answer, LLMs are likely to offer critical feedback, encouraging them to identify the student's incorrect steps.

iv) *Arithmetic — Arithmetic error:* This metric evaluates whether there is any arithmetic error in the student's solution.

   Justification: Accurate arithmetic calculations are essential for deriving correct answers in circuit analysis problems. When a student's solution contains an arithmetic error, identifying these errors can be very beneficial for learning. Therefore, we include arithmetic as one of the metrics to evaluate the LLMs' ability to provide constructive feedback on any arithmetic errors in students' solutions.

v) *Unit — Units of variables:* This metric evaluates whether the students use appropriate units for the circuit variables in their solutions.

   Justification: Correctly using units for circuit variables is important for students to master basic circuit concepts. Therefore, we include a metric to check whether units are appropriately used in students' solutions.

To the best of our knowledge, we are the first to select these metrics for evaluating LLM performance specifically within the domain of homework feedback for an undergraduate-level circuit analysis course. The prompt used in our assessment is closely tailored to these evaluation metrics. Although there is relevant research on evaluating LLMs in related areas, which reports results based on various metrics, our focus remains distinct.

For instance, Zhu *et al.* [3] assessed the assignment completion ability of LLMs in the context of middle school students' learning. Their scoring rubric included criteria such as completion, creativity, accuracy, material refinement, depth, logic, and page aesthetics. Similarly, Chiang *et al.* [1] used LLMs to generate open-ended stories and evaluated them based on grammaticality, cohesiveness, likability, and relevance. In another example, Wang *et al.* [2] explored

LLM performance in Python code generation using various prompt-engineering strategies; their evaluations focused on metrics like pass rate, time spent, and Pylint score.

In summary, the choice of evaluation metrics and associated prompt designs is generally task-specific and remains flexible in the evolving landscape of LLM development.

REFERENCES

[1] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937* (2023).

[2] Tianyu Wang, Nianjun Zhou, and Zhixiong Chen. 2024. Enhancing computer programming education with llms: A study on effective prompt engineering for python code generation. *arXiv preprint arXiv:2407.05437* (2024).

[3] Yumeng Zhu, Caifeng Zhu, Tao Wu, Shulei Wang, Yiyun Zhou, Jingyuan Chen, Fei Wu, and Yan Li. 2024. Impact of assignment completion assisted by Large Language Model-based chatbot on middle school students' learning. *Education and Information Technologies* (2024), 1–33.