Reasoning with Preference Constraints: A Benchmark for Language Models in Many-to-One Matching Markets

Recent advancements in reasoning with large languages models (LLMs) have shown strong performance on complex tasks in mathematics, including combinatorial optimization. For combinatorial optimization, reasoning benchmarks typically concern heuristic approximations of the optimal cost under simple constraints [2]. Such problems featuring complex constraints remain underexplored, even though they are important for testing LLMs' ability to interpret and enforce them. We focus on preference constraints because, as LLMs are increasingly used for decision-making, they must be able to respect user preferences. Matching problems, which involve satisfying explicit lists of agents' preferences and capacity constraints, have received little attention. The many-to-one matching problem, a generalization of the one-to-one case (known as stable marriage problem), has concrete applications, e.g., in college admission and hospital-residence matching. In this problem, adopting the terminology of the college admission application, there are two sets of agents, students and colleges, that must be matched to each other, while respecting each agent's preference list. The Deferred Acceptance (DA) algorithm [1] solves this problem in polynomial time, finding a student-optimal stable matching solution, i.e., a solution fulfilling the following metrics (1) feasibility: all capacity constraints are respected, (2) stability: there is no pair of student and college who would rather be together than their current assignment. We say matching stable if it also needs to satisfy feasibility and assignment stable if not necessary, (3) optimality: minimizing the rank of students under feasibility and stability constraints. In this article, we introduce a new benchmark dataset designed for controlled experimentation. This controlled environment, where typical instances involve 5 to 20 students, enables detailed analysis of model reasoning and decision-making processes, but can be scale arbitrary with the associated code. Our benchmark aims to evaluate scenarios where students' preferences can be (1) complete: students rank all colleges, (2) incomplete: students rank a fixed number of colleges fewer than the number available (3) flexible: students rank between a minimal number and all colleges. In the dataset, the ratio between the number of students and colleges varies from 1:1 to 4:1, whereas the capacity of the colleges are set below, equal or above the total number of students, so under-capacity scenarios force some students to remain unmatched. To analyze the performance, our dataset includes 8 different prompting strategies among the most popular, such as role prompting, In-Context Learning (ICL) and Chain-of-Thought (CoT). The different LLMs employed are open weighted for accessibility and transparency purposes, like Llama 3 (8B and 70B), Mistral (7B), Qwen (7B), Qwen-QWQ (32B) and GPT-oss (120B), with the last few ones being advanced reasoning models. Our article address the following research questions, among others:

RQ1: Are LLMs able to solve a matching problem? Solving many-to-one matching remains challenging for most LLMs. Reasoning LLMs largely outperform base ones, even if stability and optimality is difficult. The number of students drastically reduces performances of all models, with a more pronounced drop for stricter metrics. However, stronger reasoning helps preserve feasibility. The results are consistent with the complexity of the DA algorithm, $O(S \cdot P)$ with S being the number of students and P the length of their preferences. While incomplete preferences decrease the algorithm complexity, they introduce additional reasoning challenges by breaking completeness, creating more invalid student-college pairs, and requiring the LLM models to reason about the presence of unmatched agents. Despite this increase of complexity for the LLM models, our results are very much consistent for all preferences type. Within a same model, our results highlight that increased model capacity is associated with the emergence of advanced reasoning abilities, enabling performance on more complex tasks. However, the models specifically trained on reasoning largely dominate those that are not, even if the latter have much more parameters. Moreover, some models are specifically good for one particular metric, revealing that some complex metrics can be understood even with limited capacity and training as shown in Figure 1.

RQ2 : Can we improve performance with iterative prompting? Iterative prompting includes adding iteratively feedback to the previous prompt based on the last response. Our findings suggests that while it can enhance performance, especially for the easiest constraint, the technique itself is not very promising and the improvement comes from the fact that we select the best answer among all attempts. Indeed, if we select instead the last attempt out of five, the performance can substantially decrease in comparison to selecting the best one. All models' performance can benefit from having multiple attempts, but the simpler models still do not achieve as good as the reasoning ones without iterative prompting.

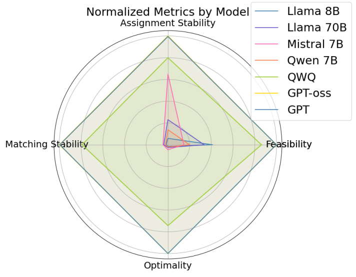

Figure 1: Performances over metrics by model.

[1] David Gale and Lloyd S Shapley. "College Admissions and the Stability of Marriage". In: *The American Mathematical Monthly* 69.1 (1962), pp. 9–15. DOI: 10.2307/2312726.

[2] Chengrun Yang et al. *Large Language Models as Optimizers*. 2024. arXiv: 2309.03409 [cs.LG]. URL: https://arxiv.org/abs/2309.03409.