that the aggregated chain is logically incorrect and does not adhere to any of the rules of the `Game of 24`.

However, we argue that spurious feedback will not be amplified over iterations; instead, thanks to the iterative mechanism of BoT, its negative impact on the generated reasoning steps can be mitigated or even entirely rectified in subsequent iterations. The main reason is that the generated wrong reasoning steps will be further analyzed to produce new experiences to be added to the prompt. Specifically, as these reasoning steps contain obvious mistakes that are easy to identify, LLMs are prone to generating correct error analysis and providing effective advice for revisions. With this new experience included in the prompt, BoT is capable of generating correct reasoning steps. As demonstrated by the experience in Table 8, BoT produces detailed error reports and revision suggestions, resulting in a rational thought generation process illustrated in Table 7.

The advantage of BoT, which leverages iterations to mitigate the detrimental effects of erroneous feedback, is evident in Figure 4. Notably, the performance of BoT exhibits consistent enhancement with an increasing number of iterations. This underscores both the significance of accumulating experience iteratively and the capacity of subsequent experiences to rectify prior errors.

F MORE RESULTS ON MATH

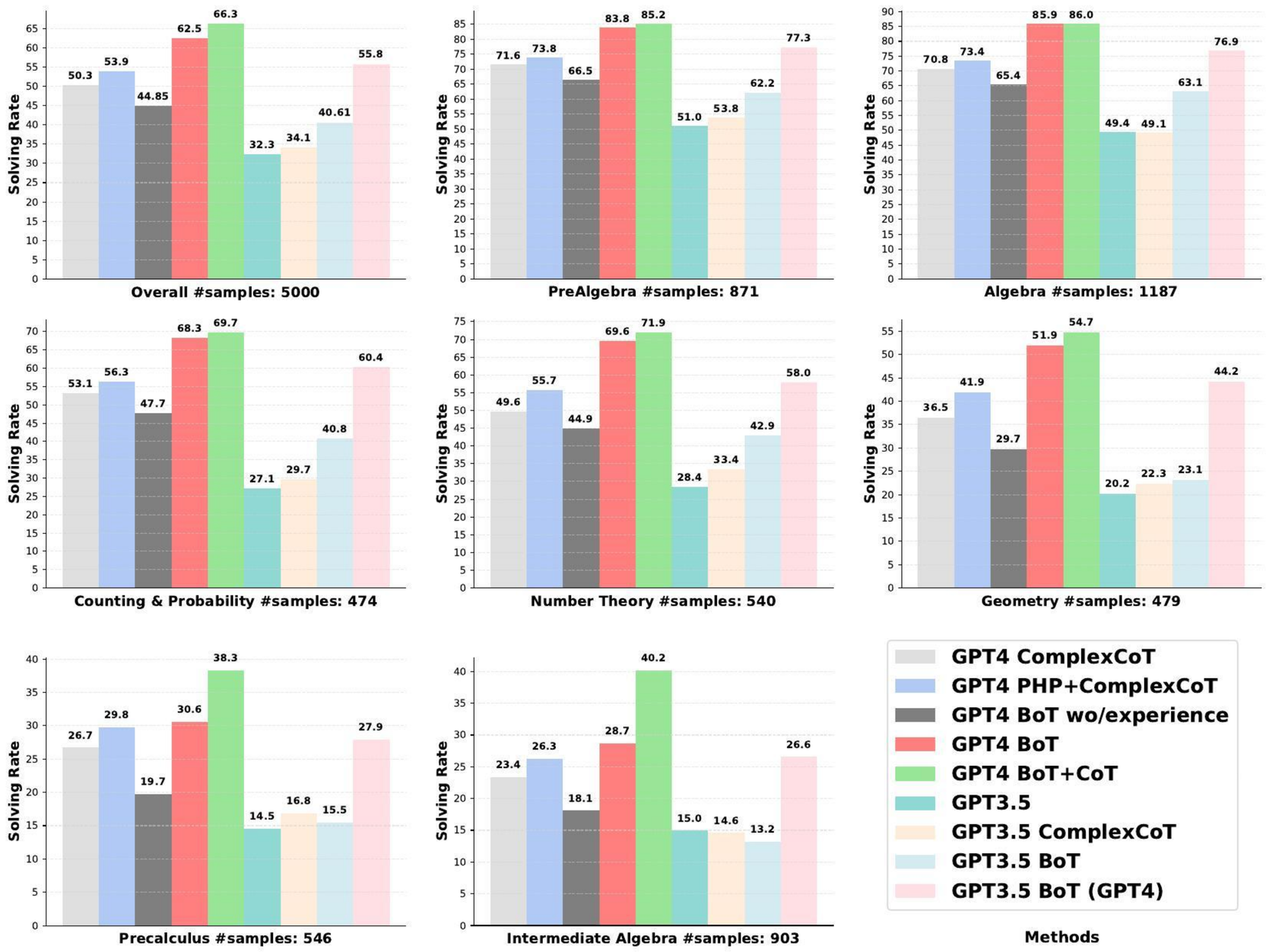

Figure 5: Solving rates on all the problems from different categories of the `MATH` dataset with different methods. The comparsion between these methods are performed on the categories, including PreAlgebra, Algebra, Counting & Probability, Number Theory, Geometry, Precalculus, and Intermediate Algebra, of the test set. The sub-figure with the 'Overall' shows the solving rate computed on all the problems across all categories.

In Figure 5, we provide the solving rate of different methods in each category of the `MATH` dataset. The diverse range of mathematical problems in these categories poses a significantly more challenging benchmark for mathematical reasoning. Thus, the complexity and diversity of the problems in `MATH` require a wide spectrum of reasoning capabilities for solutions. Consequently, a detailed examination of our approach and its comparison with other methods in this context yields valuable insights.