proof from each thread, and (2) Best@32 – the score of the best proof per problem, selected by self-assigned scores across all threads. The self-selected best proofs achieve significantly higher verification scores than the thread average, demonstrating our generator's ability to accurately assess proof quality. Furthermore, Pass@1 improves substantially as maximum sequential attempts increase, showing that self-verification effectively guides iterative improvement. These results confirm that our generator can reliably differentiate between high-quality and flawed proofs, and leverage this self-awareness to systematically improve its mathematical reasoning.

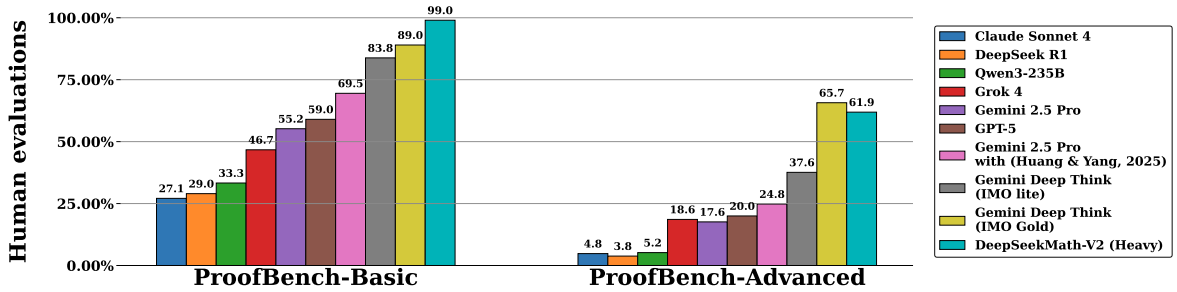

Figure 3 | Expert evaluation results on the Basic and Advanced subsets of IMO-ProofBench. All results are sourced from Luong et al. (2025), with the exception of DeepSeekMath-V2, which was evaluated by our experts following the grading guidelines.

3.3.3. High-Compute Search

To solve the most challenging problems, we scaled both verification and generation compute – using extensive verification to identify subtle issues and parallel generation to explore diverse proof strategies.

Contest	Problems	Points
IMO 2025	**P1** , **P2** , **P3** , **P4** , **P5**	83.3%
CMO 2024	**P1** , **P2** , **P4** , **P5** , **P6**	73.8%
Putnam 2024	**A1** ~ **B4** , **B5**, **B6**	98.3%

Table 1 | Problems in gray are **fully solved** , while underlined problems received **partial credit**.

Our approach maintains a pool of candidate proofs for each problem, initialized with 64 proof samples with 64 verification analyses generated for each. In each refinement iteration, we select the 64 highest-scoring proofs based on average verification scores and pair each with 8 randomly selected analyses, prioritizing those identifying issues (scores 0 or 0.5). Each proof-analysis pair is used to generate one refined proof, which then updates the candidate pool. This process continues for up to 16 iterations or until a proof successfully passes all 64 verification attempts, indicating high confidence in correctness. All experiments used a single model, our final proof generator, which performs both proof generation and verification.

To validate our results, mathematical experts assessed the highest-scoring proofs. As shown in Table 1, our approach solved 5 of 6 problems from IMO 2025 and 4 problems plus partial credit on another from CMO 2024, achieving gold medal performance in both pinnacle high-school competitions. On Putnam 2024, the preeminent undergraduate mathematics competition, our model solved 11 of 12 problems completely and the remaining problem with minor errors, scoring 118/120 and surpassing the highest human score of 90. Figure 3 shows the results on IMO-ProofBench. Our approach outperforms DeepMind's DeepThink (IMO Gold) on the basic set and remains competitive on the advanced set, while substantially outperforming all other baselines. We observe that the hardest IMO-level problems remain challenging for our model.