F The Editing Distance Results by Problem Sets

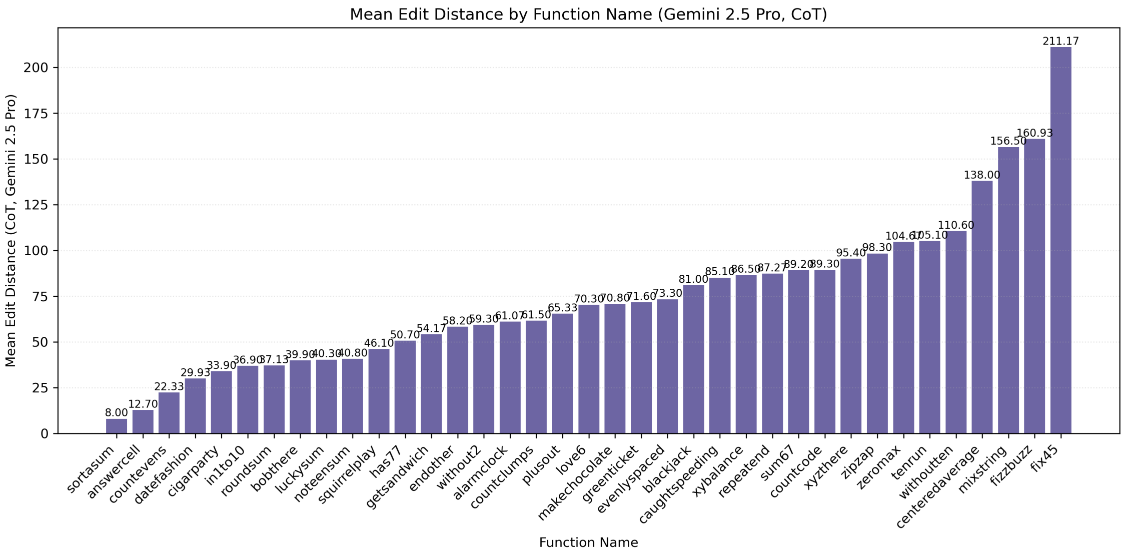

Fig. 6. Problem-Level Variation in Mean Intra-Model Edit Distance (Gemini 2.5 Pro, CoT Prompting).

Note: This figure reports the average pairwise edit distance among multiple LLM-generated code solutions for each function name (problem). Results are specific to the Gemini 2.5 Pro model under the chain-of-thought (CoT) prompting technique. Larger values indicate higher diversity across generated outputs, while smaller values suggest more consistent code generations.

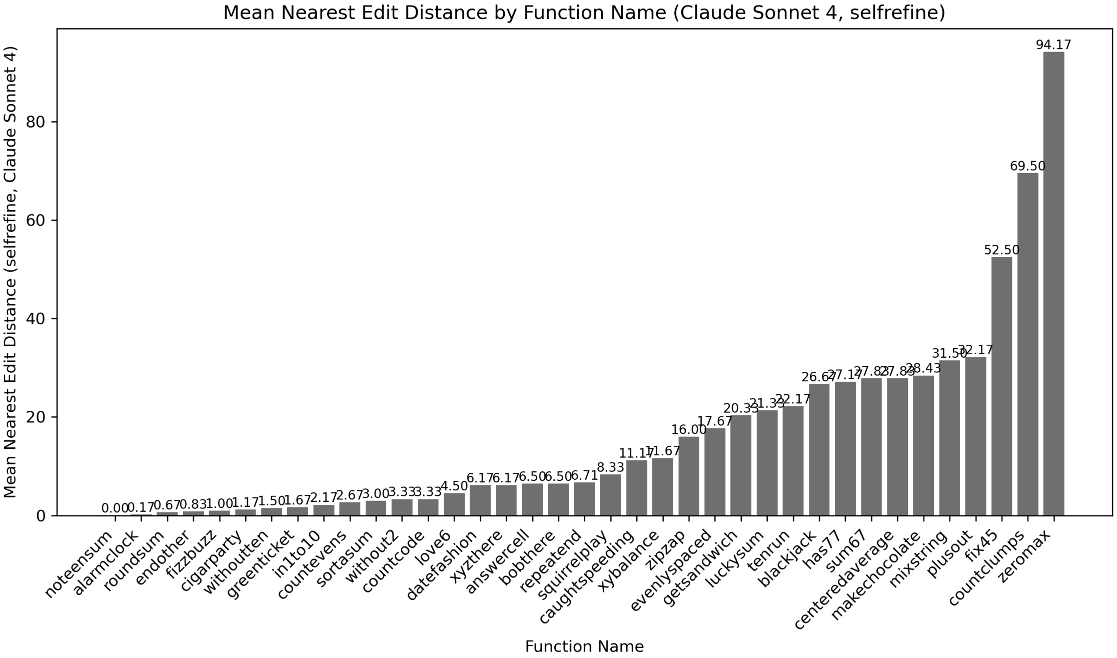

Fig. 7. Problem-Level Variation in Mean Nearest Edit Distance (Claude Sonnet 4, Self-Refine Prompting)

Note: This figure shows the average nearest edit distance between LLM-generated code and human references across different function names (problem). Results are specific to the Claude Sonnet 4 model under the self-refine prompting technique, illustrating substantial variation in function-level alignment. Higher values indicate greater divergence from human submissions, highlighting functions with more challenging alignment.