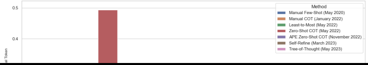
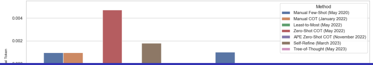
The Practicality of Prompt Engineering

- What prompt engineering methods are most effective for a given cost (length/complexity, financial)?
- Data: 100 programatically-conducted conversations
- Contribution - new metrics, significance testing, run on new and old models
- Statistically Significant Improvements (Paired Tests): GSM8K All, Creative Writing GPT-4 Manual CoT, Zero-Shot CoT
- Gains on CW are growing but gains on GSM8K are shrinking in transition from TD3 to GPT-4. Methods below sorted by technique age
- Low variance is achievable for CW manual few shot, manual CoT with GPT-4
- GPT-4 - 3 cents per 1K tokens input, 6 cents per 1K tokens output. TD3 - 2 cents per 1K tokens.


al Token
Method
Manual Few-Shot (May 2020)
Manual COT (January 2022)
Least-to-Most (May 2022)
Zero-Shot COT (May 2022)
APE Zero-Shot COT (November 2022)
Self-Refine (March 2023)
Tree-of-Thought (May 2023)


al Token
Method
Manual Few-Shot (May 2020)
Manual COT (January 2022)
Least-to-Most (May 2022)
Zero-Shot COT (May 2022)
APE Zero-Shot COT (November 2022)
Self-Refine (March 2023)
Tree-of-Thought (May 2023)