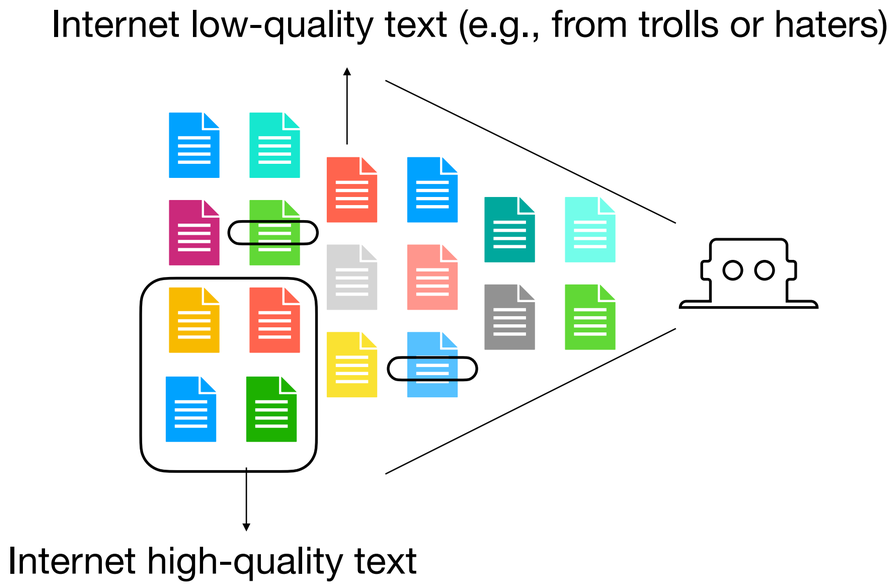
Prompt Engineering

Haw-Shiuan Chang

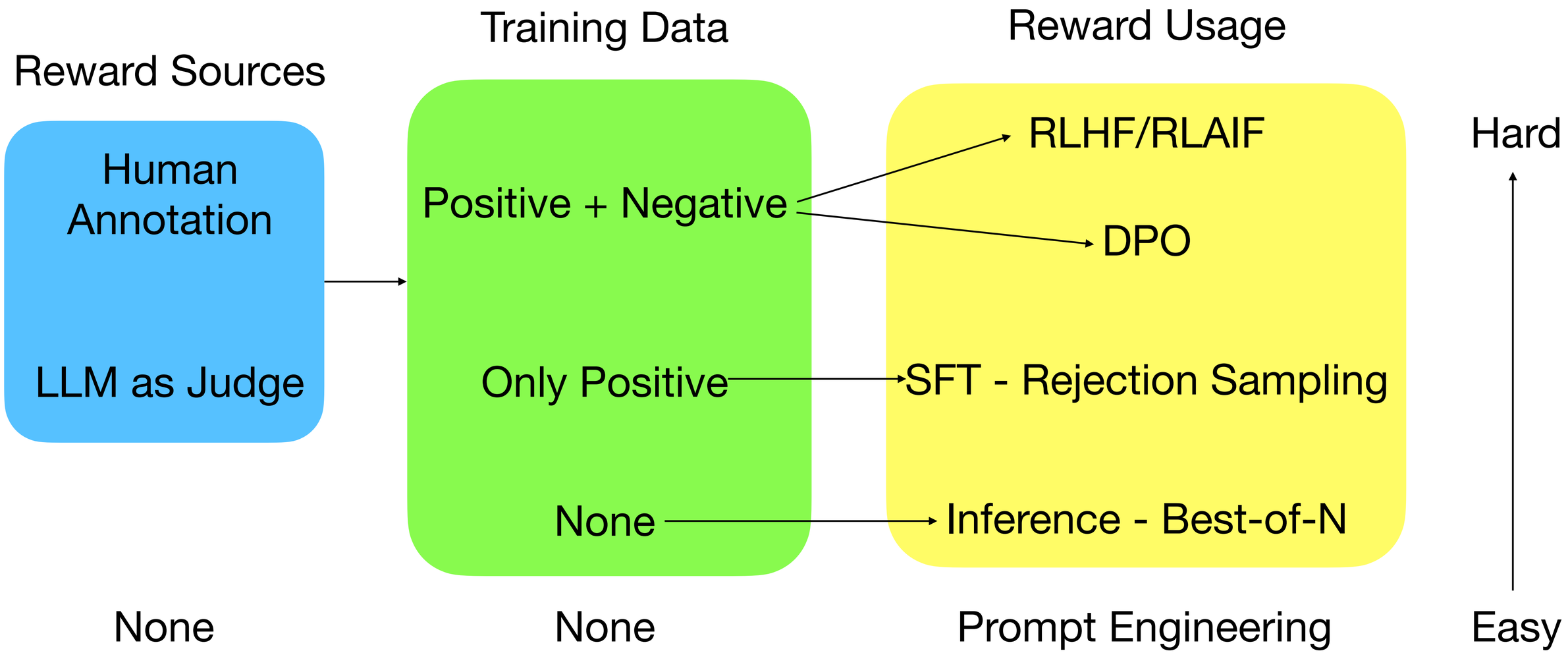
LLM Development



- Architectures
  - MLP
  - RNN
  - Transformer
- Training Stages
  - Pretraining
  - Supervised Fine-tuning (SFT)
  - Alignment
    - Learning from Human Feedback (LHF)
    - Reasoning

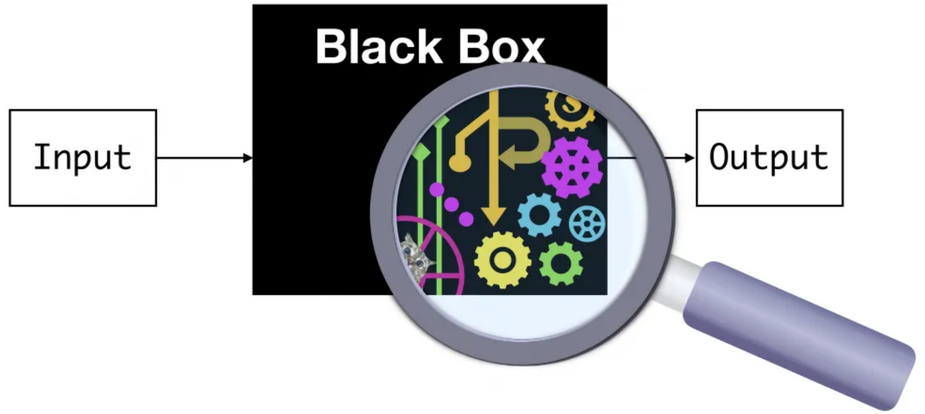
How do you improve LLM's performance without training?

Available Methods

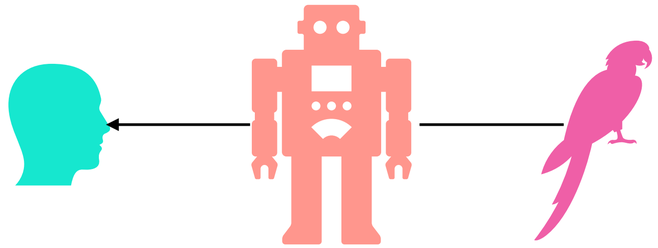


Inference-time Improvement

- **Prompt engineering**
  - **In-context learning**
- Decoding
- **Agentic**
  - RAG
  - Tools
  - Assistant
  - Multi-LLM collaboration



Human brain is also almost a black box



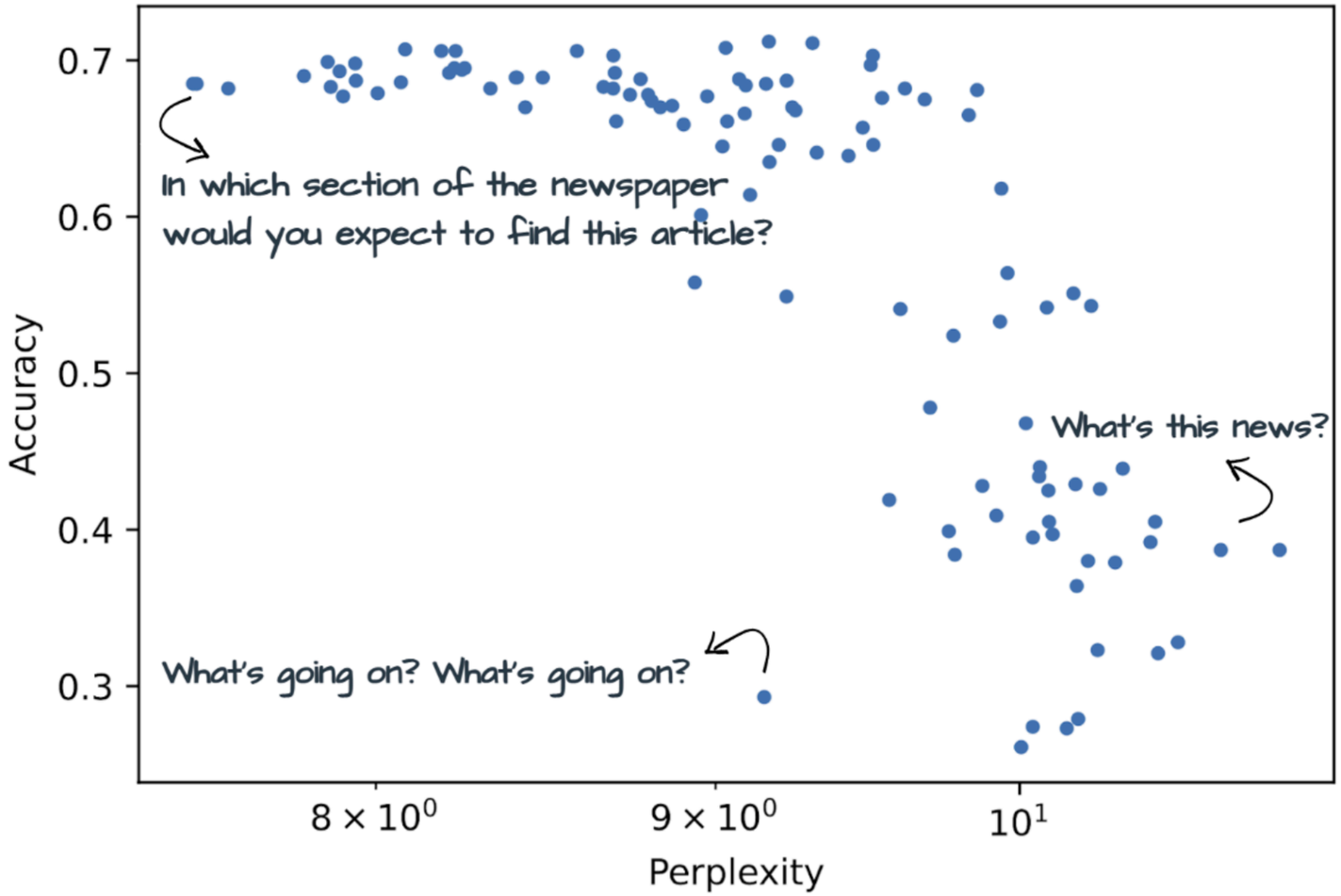
Lots of cognitive science

https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Prompt Engineering

- What prompts are better?
- Chain of Thought
- Self-Consistency and Tree of Thoughts / Beam Search
  - Similar to Best of N
- In-context Learning

Which Prompts are better?



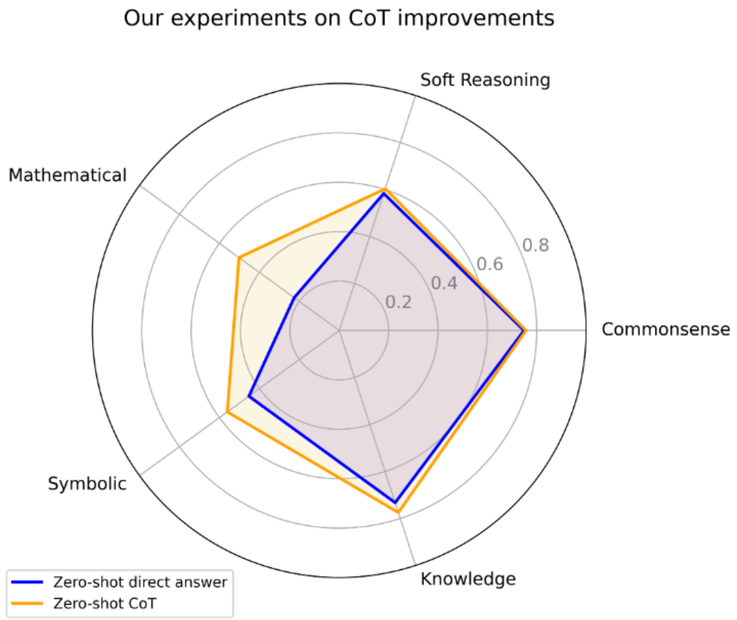
Demystifying Prompts in Language Models via Perplexity Estimation (https://arxiv.org/pdf/2212.04037)

Which Prompt is better?

- User: Please generate a sci-fi story
- LLM: OOXX
- User: Please revise the story to reveal a big secret of the main character
- LLM: XXOO
- User: Please revise the story to …

- User: Please generate a sci-fi story
- Constraint 1: Please revise the story to reveal a big secret of the main character
- Constraint 2: Please revise the story to …
- LLM: XXOO

Chain of Thoughts



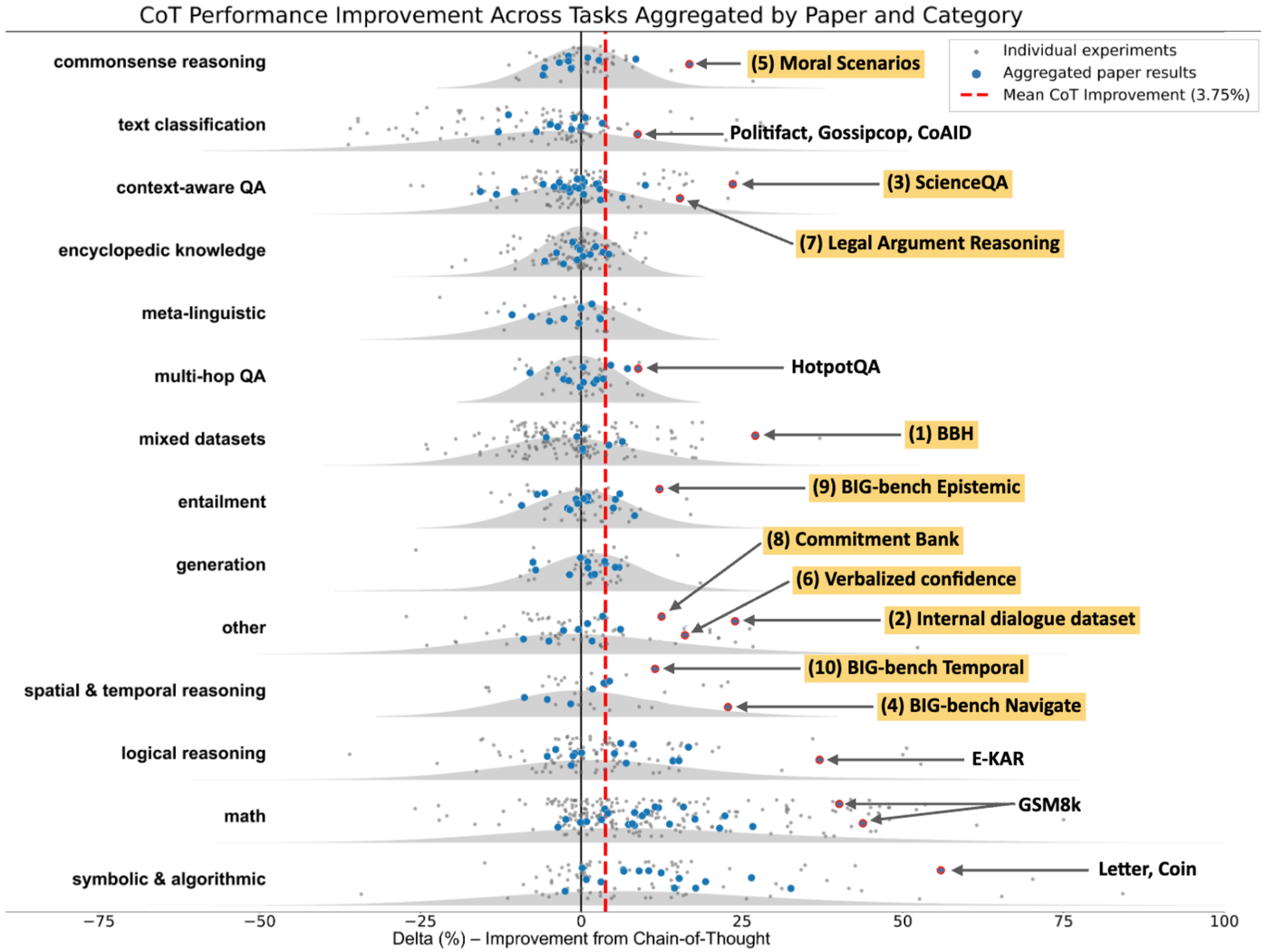


Figure 2: Results from our meta-analysis (grey dots) aggregated by paper and category (blue dots).

Self Consistency and Tree of Thoughts

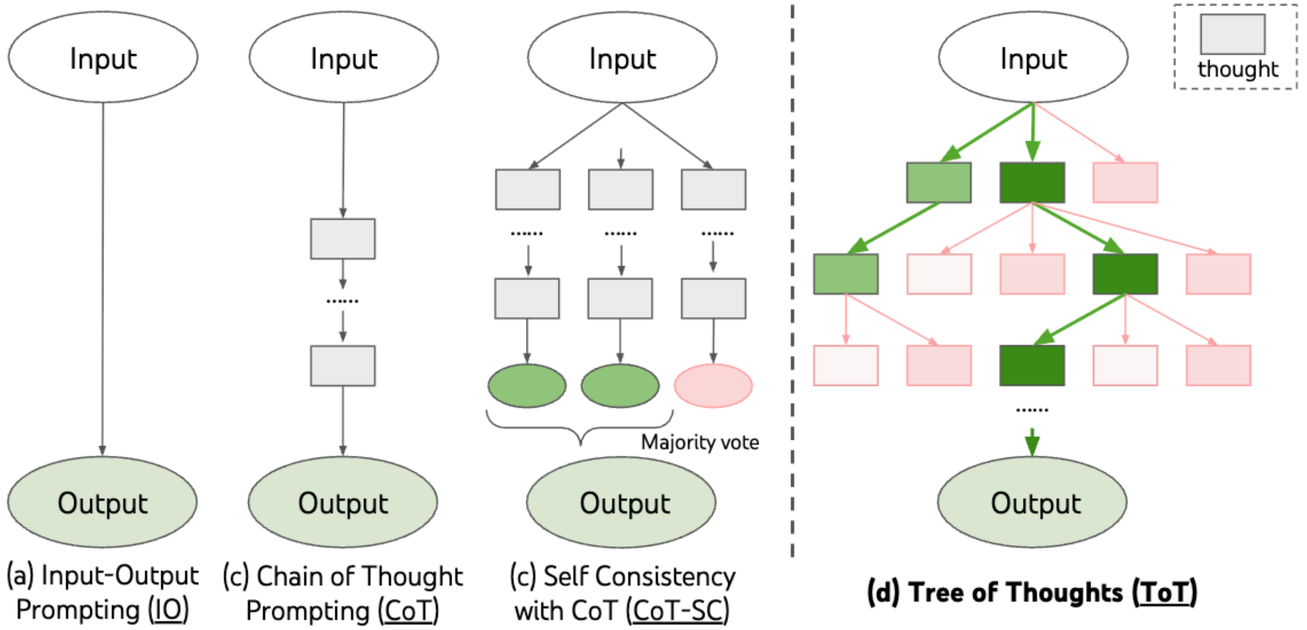
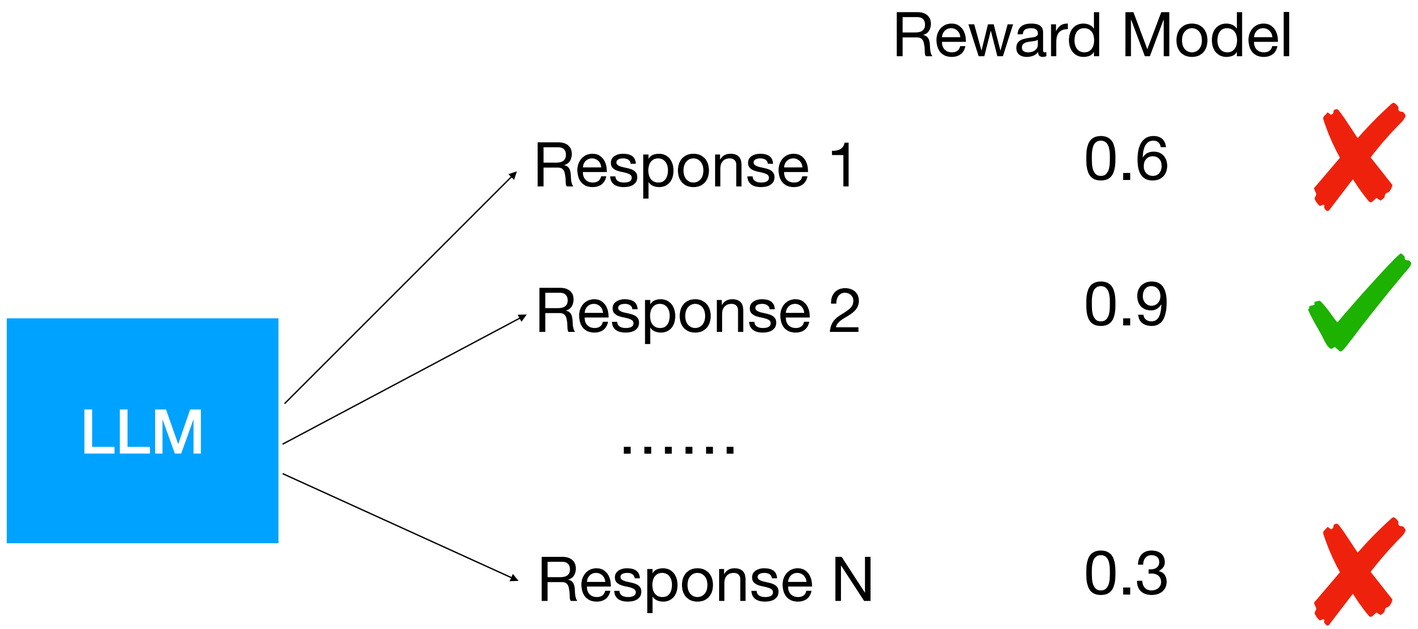


Figure 1: Schematic illustrating various approaches to problem solving with LLMs. Each rectangle box represents a *thought*, which is a coherent language sequence that serves as an intermediate step toward problem solving. See concrete examples of how thoughts are generated, evaluated, and searched in Figures 2,4,6.

Best of N



The reward model could be anything. For example, LM probability (beam search), answer quality scorer, profanity/toxicity filter, sentiment classifier, PRM

Best-of-N vs Beam Search

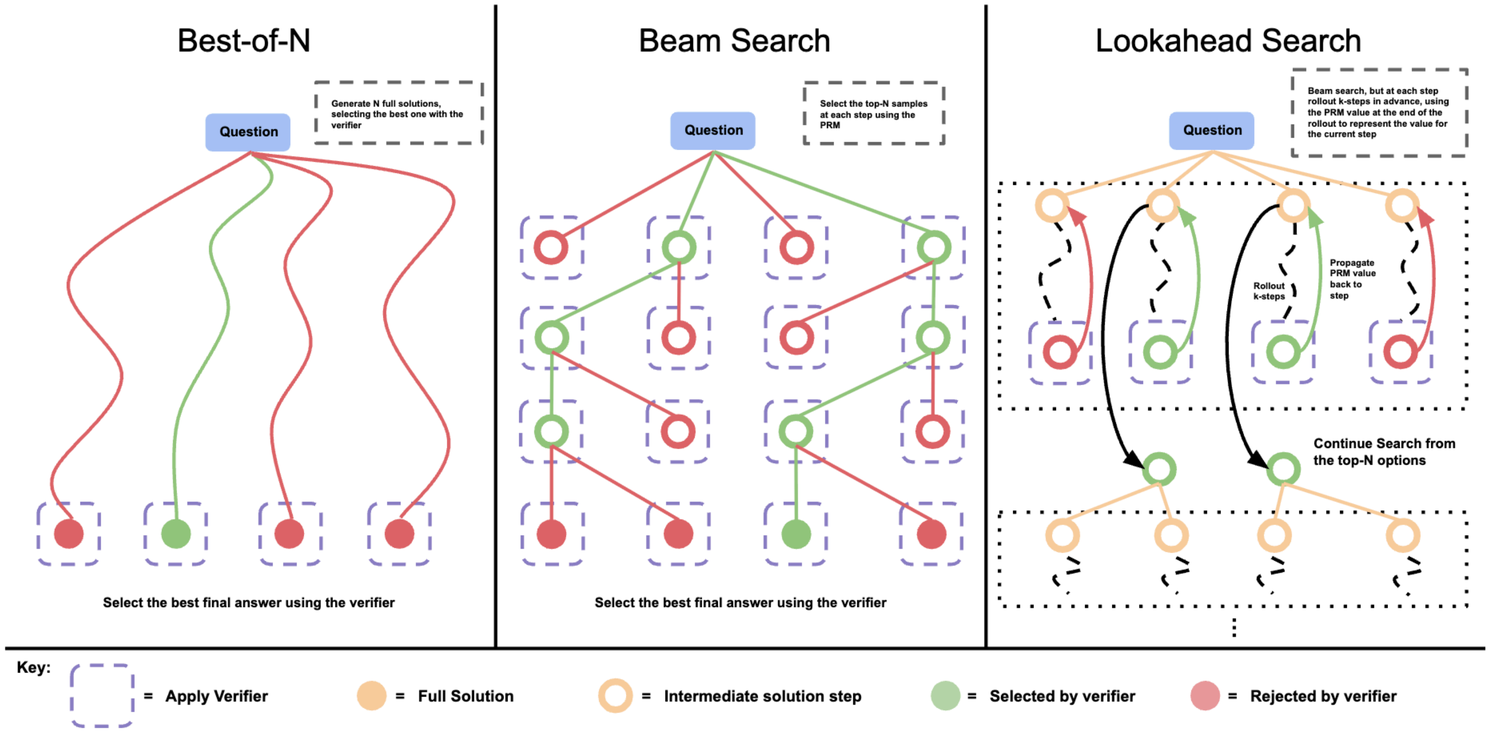
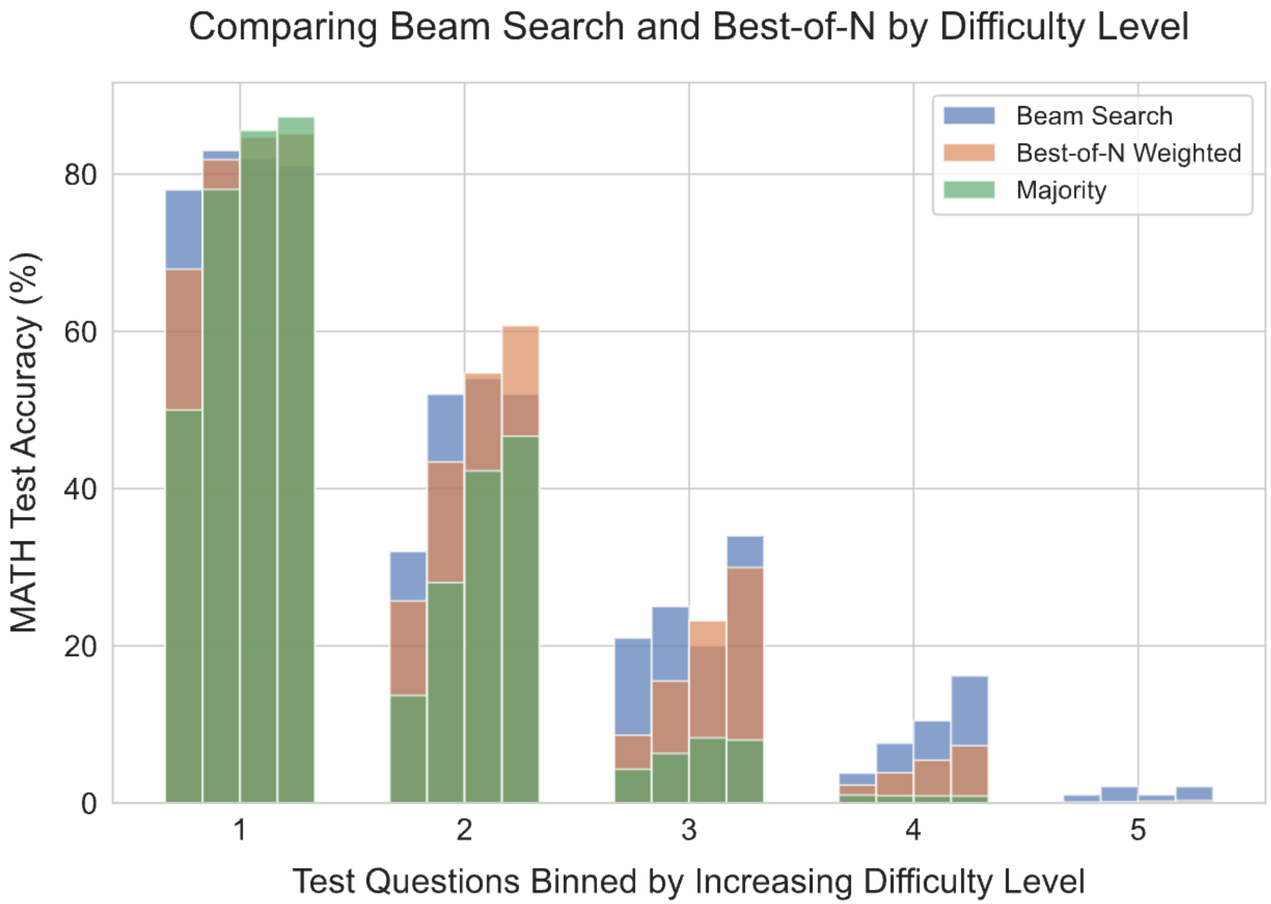


Figure 2 | ***Comparing different PRM search methods.*** **Left:** Best-of-N samples N full answers and then selects the best answer according to the PRM final score. **Center:** Beam search samples N candidates at each step, and selects the top M according to the PRM to continue the search from. **Right:** lookahead-search extends each step in beam-search to utilize a k-step lookahead while assessing which steps to retain and continue the search from. Thus lookahead-search needs more compute.

Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters (https://arxiv.org/pdf/2408.03314)

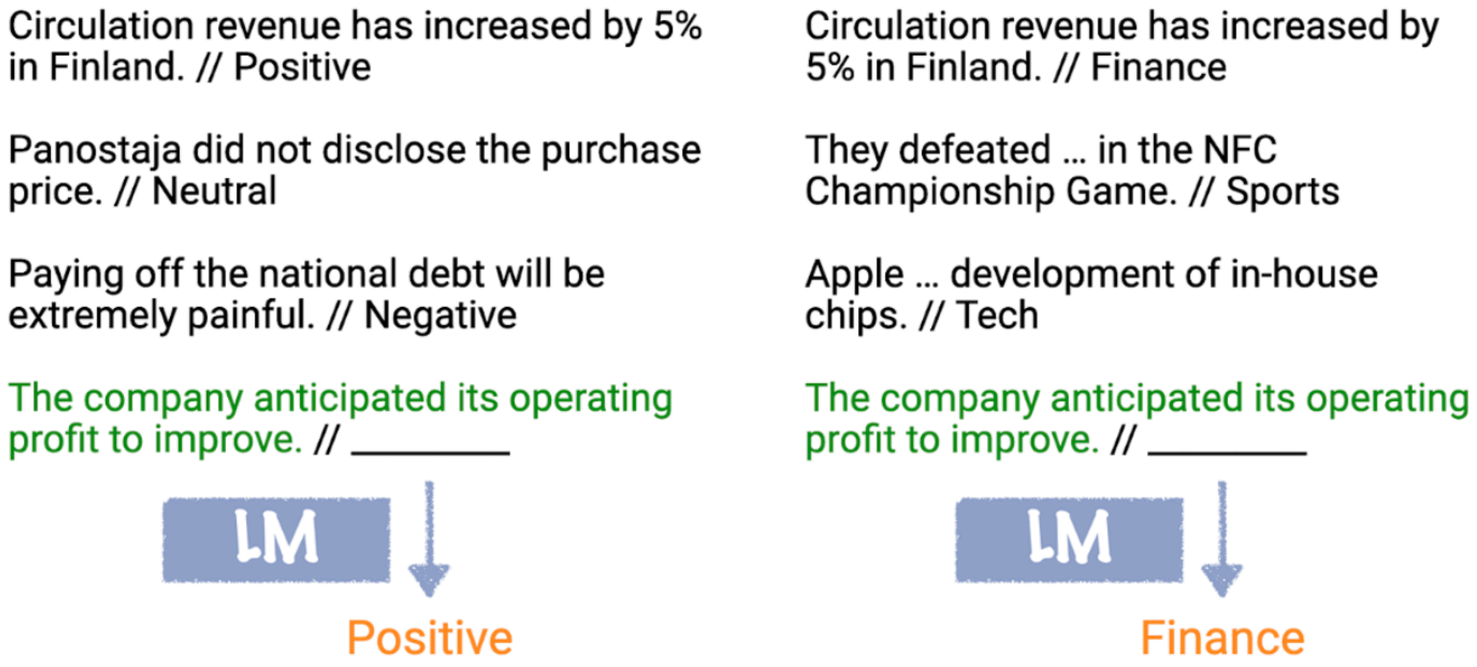
Usefulness of Guidance Depends on the Difficulty



- For LLM rather than LRM

In-context learning:

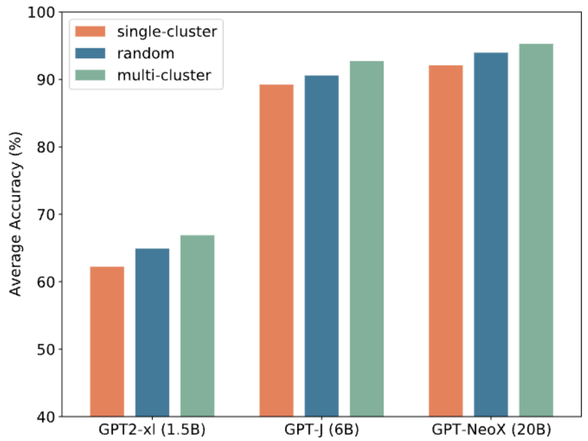
LLMs can solve novel downstream tasks by just conditioning on a few demonstrations of the task in its prefix



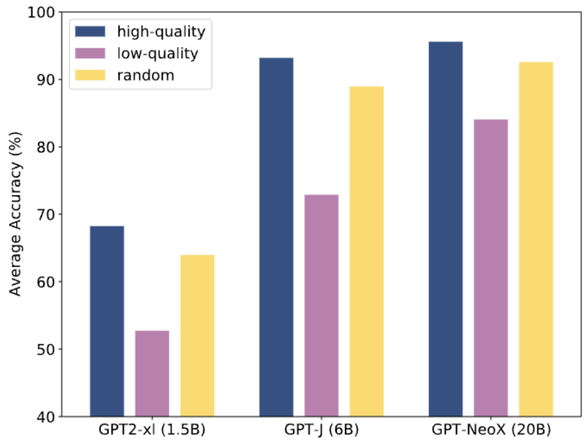
https://ai.stanford.edu/blog/understanding-incontext/

Which Examples should we choose?

- Principle
  - Diversity
  - Quality
  - Relevancy/Coverage to query
- The optimal strategy seems to be task-dependent



(a) Impact of semantic diversity



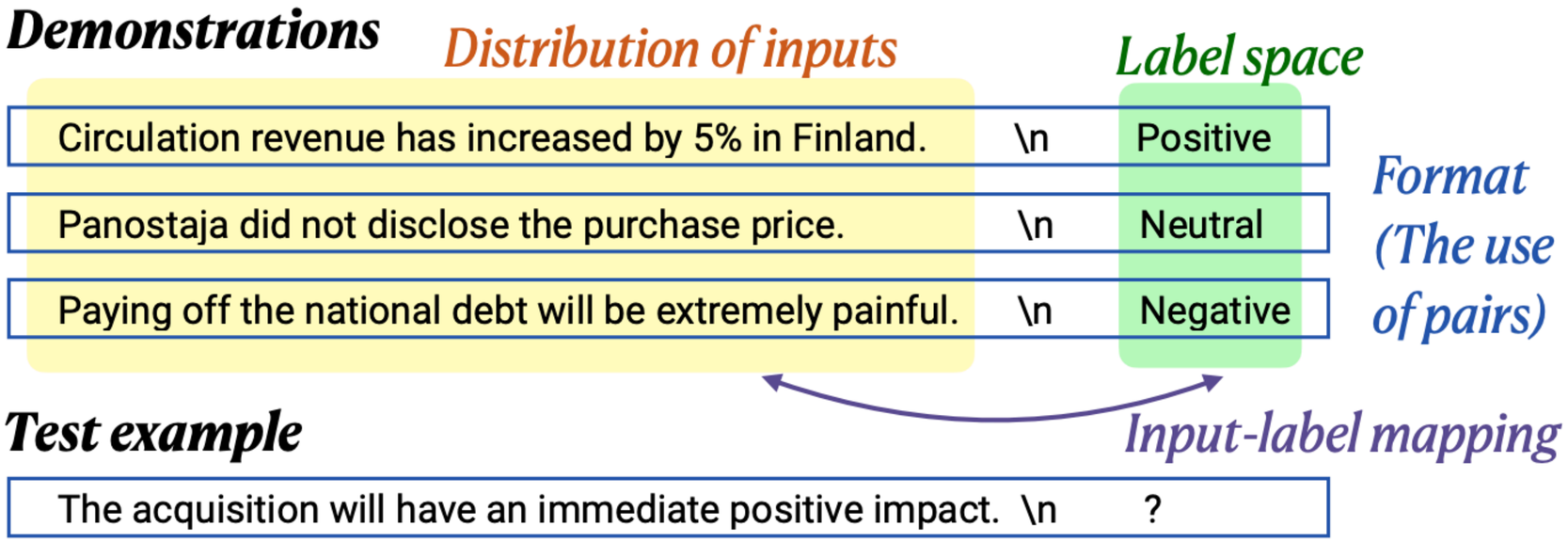
(b) Impact of instance quality

Representative Demonstration Selection for In-Context Learning with Two-Stage Determinantal Point Process (https://aclanthology.org/2023.emnlp-main.331.pdf)

In-context Learning with Retrieved Demonstrations for Language Models: A Survey (https://arxiv.org/pdf/2401.11624v1)

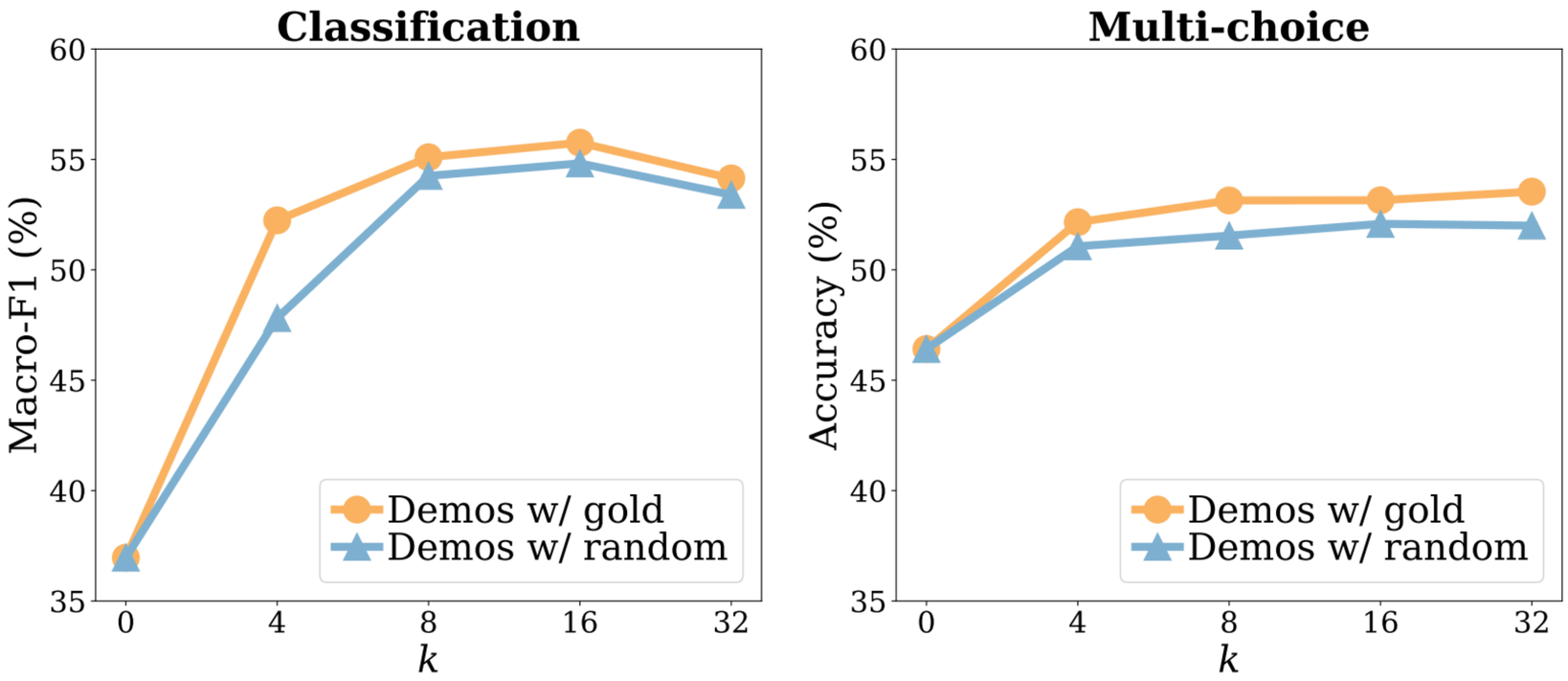
https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/

What's in a demonstration?



Min et al., EMNLP 2022

How important is input-label mapping?



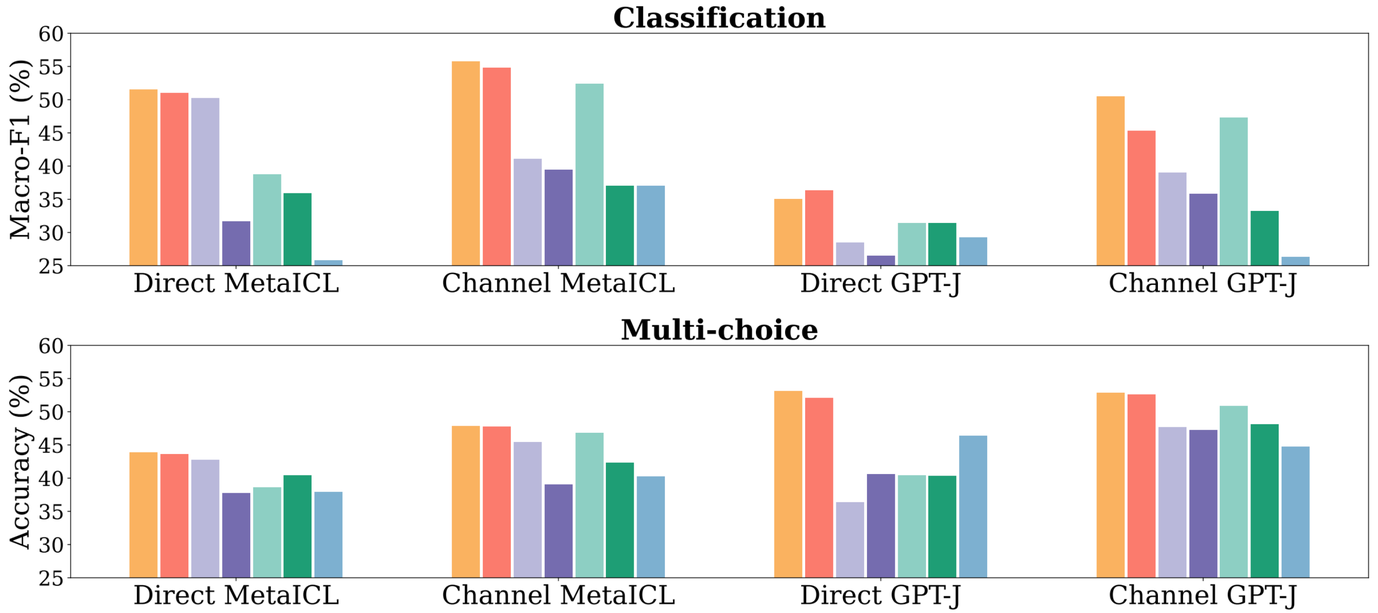
Min et al., EMNLP 2022

What about other aspects of the demonstrations?

Demos w/ gold labels	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✓) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n positive Panostaja did not disclose the purchase price. \n neutral
Demos w/ random labels	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n neutral Panostaja did not disclose the purchase price. \n negative
OOD Demos w/ random labels	(*Format* ✓ *Input distribution* ✗ *Label space* ✓ *Input-label mapping* ✗) Colour-printed lithograph. Very good condition. Image size: 15 x 23 1/2 inches. \n neutral Many accompanying marketing claims of cannabis products are often well-meaning. \n negative
Demos w/ random English words	(*Format* ✓ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n unanimity Panostaja did not disclose the purchase price. \n wave
Demos w/o labels	(*Format* ✗ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. Panostaja did not disclose the purchase price.
Demos labels only	(*Format* ✗ *Input distribution* ✗ *Label space* ✓ *Input-label mapping* ✗) positive neutral

Min et al., EMNLP 2022

What about other aspects of the demonstrations?



	F	*L*	*I*	*M*
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
OOD + Random labels	✓	✓	✗	✗
Random labels only	✗	✓	✗	✗
Random English words	✓	✗	✓	✗
No labels	✗	✗	✓	✗
No demonstrations	✗	✗	✗	✗

F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

Min et al., EMNLP 2022

Followup work tells a different story:

- Yoo et al., 2022 shows that input-label mappings matter quite significantly when using different experimental conditions and eval metrics
- Madaan and Yazdanbakhsh (2022) show that random rationales degrade chain-of-thought performance, but other modifications to the rationale (e.g., wrong equations) don't affect it too much

A Survey to Recent Progress Towards Understanding In-Context Learning (https://arxiv.org/pdf/2402.02212)

Question

- After SFT and RLHF, does prompt engineering become more important or less important?
- Why?

- Prompt engineering becomes less important
- For familiar prompts, the LLMs might output good answers
- For unfamiliar prompts, the LLMs might output good answers or bad answers
  - Bad answers will be filtered out in SFT/RLHF