Tutor-ICL: Guiding Large Language Models for Improved In-Context Learning Performance

Ikhyun Cho¹, Gaeul Kwon², and Julia Hockenmaier¹ ¹Siebel School of Computing and Data Science, University of Illinois at Urbana-Champaign, ²Department of Computer Science, Grinnell College



Target: touchscreen functions Sentiment:

Standard Template

We improve ICL classification performance by adding three novel components to the standard template.

Experimental Settings

Tasks and Datasets

- **1. Aspect-Based Sentiment Classification** (SemEval-14-Laptops/Restaurants)
- 2. News Topic Classification (AGNews)
- **3. Question Type Classification** (TREC-QC)
- 4. Emotion Classification in Dialogues (EmoC)

Models

- 1. Encoder-Decoder LLMs (Flan-T5-XL (3B), XXL (11B))
- 2. Decoder Only LLMs (Llama2-7B, 13B, Llama3-8B-Instruct)

Tutor-ICL

1. Comparative Answer Format (CAF)

Results **1.** Tutor-ICL achieves consistent improvements across various models and tasks: Flan-T5-XXL 11B (F1 score) +6.05TREC QC Baseline + GAT Baseline + Summary Tutor-ICL Llama3-8B-Instruct (F1 score) +10.96 +12.70 +7.95 TREC QC Rest14 Lap14 AGNews Tutor-ICL Baseline + Summary 2. With Tutor-ICL, LLMs make better use of exemplars: Oracle Accuracy (Llama3-8B-Instruct) Gold Instance @ firs 99.94 (+ 0.83) Gold Instance @ last 98.07 (+ 3.26) 98.61 Gold Instance @ first 99.76 (+ 1.15) Gold Instance @ last 96.10 (+ 2.38) Baseline Tutor-IC We include the test instance & gold answer as one of the ICL exemplars. Tutor-ICL results in accuracy Example 1: closer to 100%. Target: Boot time Conclusion We improve ICL classification performance by adding three mew components to the standard ICL template:

Exemplars (N-shot) with Comparative Answer Format (CAF)

Pretend that you are an expert in sentiment and opinion analysis. You need to evaluate the sentiment of the sentence toward the Example 1: Target: Boot time Sentiment: closer to positive than neutral Sentence: Did not enjoy the new Window 8 and touchscreen functions. Target: touchscreen functions Sentiment: Specifically, the goal is to determine the sentiment polarity toward touchscreen functions in the sentence: Did not enjoy the new Window 8 and touchscreen functions. When reading the examples designed to aid your judgment, review the examples based on their contribution to solving the goal. Target: touchscreen functions Sentiment: Pretend that you are an expert in sentiment and opinion analysis. You need to evaluate the sentiment of the sentence toward the specified target, determining whether it is positive, neutral, or negative. Sentence: Boot time is super fast, around anywhere from 35 seconds to 1 minute. Sentiment: positive Let's summarize the examples: Example 1: closer to positive than neutral, ..., Example N: closer to neutral than positive Sentence: Did not enjoy the new Window 8 and touchscreen functions.

Test Instance

2. Glance-At-the-Test (GAT)

Input Instruction

specified target, determining whether it is positive, neutral, or negative. Sentence: Boot time is super fast, around anywhere from 35 seconds to 1 minute. Pretend that you are an expert in sentiment and opinion analysis. You need to evaluate the sentiment of the sentence toward the specified target, determining whether it is positive, neutral, or negative. Example 1: Sentence: Boot time is super fast, around anywhere from 35 seconds to 1 minute. Target: Boot time Sentiment: positive Sentence: Did not enjoy the new Window 8 and touchscreen functions.

Input Instruction Comparative answer format encourages deeper thinking from various answer perspectives. Glance-at-the-Test (GAT) **Exemplars (N-shot)** Test Instance Providing the test instance in advance allows LLMs to focus more on the relevant exemplars using the self-attention Exemplars (N-shot) **Test Instance**

mechanism

3. Summarization

Input Instruction

Summarization

Target: touchscreen functions Sentiment:













```
1. Comparative Answer Format (CAF)
2. Glance-at-The-Test (GAT)
3. Summarization
```