Don’t Just Demo, Teach Me the Principles: A Principle-Based Multi-Agent Prompting Strategy for Text Classification

Peipei Wei, Dimitris Dimitriadis, Yan Xu, Mingwei Shen
Amazon

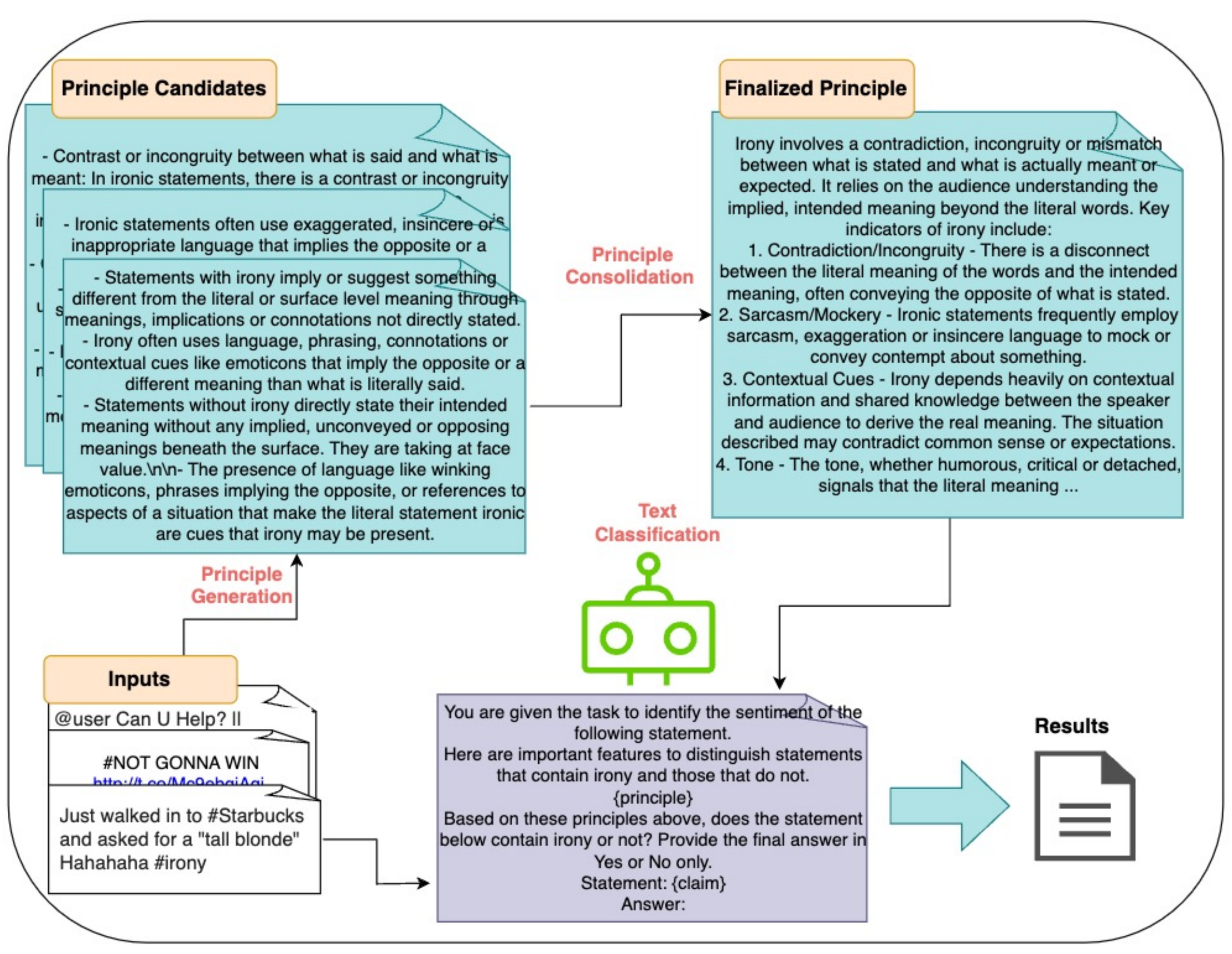
amazon | science

Introduction:

- **In-Context Learning (ICL) & Finetuning:**
  - As one of the emerging capabilities of large language models (LLMs), in-context learning (ICL) allows tasks to be performed via instructions and demonstrations without requiring parameter updates.
  - ICL excels in zero/few-shot settings for tasks such as QA, reasoning etc, but underperforms fine-tuned models in text classification.
- **Challenges:**
  - **Fine-tuned models:** Require costly, time-consuming human annotations.
  - **ICL with LLM:** Reliant on prompt engineering expertise, increased inference costs with demonstrations, and input length constraints.
- **Human-Inspired Solution:**
  - Hypothesis: Injecting **knowledge-intensive principles** into LLMs via ICL could bridge performance gaps in text classification.
  - Propose mimicking **Standard Operating Procedures (SOPs)**—used by domain experts to extract task principles from examples—to enhance LLMs’ task-specific knowledge.
- **Key Research Question:**
  Can task-specific principles, derived from demonstrations, mitigate LLMs’ lack of domain knowledge and improve performance in text classification?

Methodology:

- **Principle-Based Prompting Framework:**
  - **Motivation:** Inspired by humans’ use of abstract principles (vs. memorizing data) for classification tasks.

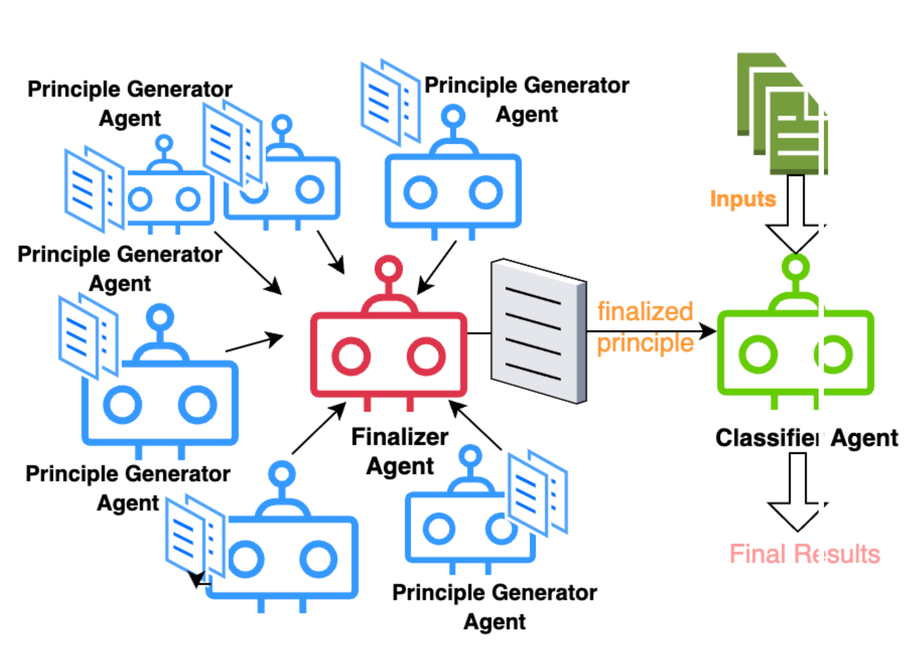


- **Three-Step Workflow:**

Step 1. Principle Generation
- **Process:**
  - LLMs analyze **n = [4, 8, 16]** labeled/unlabeled demonstrations to generate task-level principles.
- **Models:**
  - Tested **6 LLMs** (open/closed-source, varying sizes);
- **Output:** 36 candidate principles per task (varying n, label use, and LLM agents).

Step 2. Principle Consolidation
- **Methods:**
  - **Listwise Ranking:**
    - LLM agents rank top 5 principles; majority voting selects the final principle.
    - Tested with randomized order and different number of demonstrations
  - **Consolidation:**
    - Summarizes and integrates key points from all 36 candidates, resolving conflicts.
  - **Random Selection (Control):** Randomly picks one candidate principle.

Step 3. Text Classification
- **Process:**
  - Optimal principle appended to prompts as context for classification.
  - Tested on **FLAN-T5-XXL** and **FLAN-UL2**
- **Setup:**
  - 5 random seeds; same hyperparameters
  - Also evaluated on internal datasets with human-vs-LLM-generated principles.



Results:

Table 1: Absolute improvements in the macro-F1 scores over the zero-shot vanilla prompting for various single- and multi-agent approaches under the zero-shot settings. Human-crafted principles are only available for two private datasets. Results are averaged across five inferences with different random seeds.

Model	Method		Irony2018	Emotion20	Financial	PC1	PC2	AVG
flan-t5-xxl	single agent	CoT	-9.31	-14.23	1.51	-1.56	17.25	-1.27
		stepback	-2.03	1.68	-3.31	1.36	17.56	3.05
		principle	2.62	8.13	3.40	1.40	12.89	5.69
		principle+human	NA	NA	NA	3.98	14.89	NA
	multi agent	principle+random	0.63	9.74	6.69	2.43	14.16	6.73
		principle + ranking	1.55	9.52	4.16	**3.71**	13.84	6.56
		principle+consolidation	0.45	12.13	4.38	1.43	16.21	6.92
flan-ul2	single agent	CoT	-6.87	0.41	0.96	-0.58	13.46	1.48
		stepback	2.72	0.47	4.18	0.02	13.99	4.28
		principle	4.57	0.02	3.42	-0.2	13.03	4.17
		principle + human	NA	NA	NA	0.90	13.26	NA
	multi agent	principle+random	**5.56**	12.15	11.78	-0.54	19.08	9.61
		principle+ranking	4.96	11.14	11.05	1.57	18.69	9.48
		principle+consolidation	4.77	**15.11**	**14.17**	0.04	**19.37**	**10.69**
RoBERTa	full	finetune	0.44	***17.62**	***16.62**	-5.26	-7.93	4.30
	10%		-19.71	-41.01	-52.41	NA	NA	NA

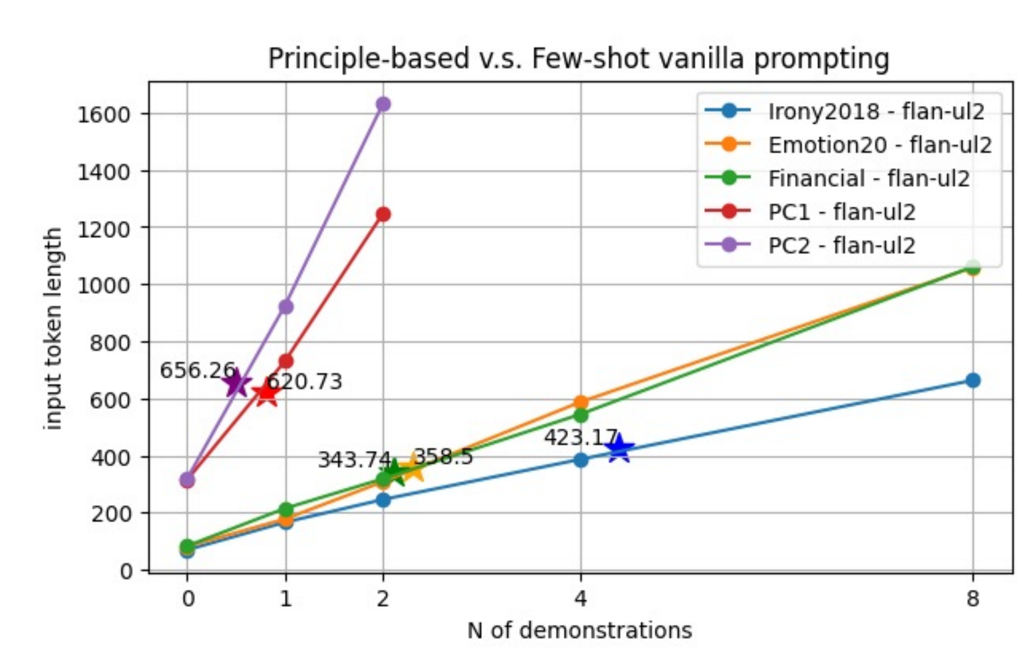
- **Principle-Based Prompting vs. Baselines:**
  - Outperforms **vanilla prompting**, **CoT**, and **stepback prompting** in zero-shot settings for both FLAN-T5-XXL and FLAN-UL2.
  - Achieves **10.69%** (FLAN-UL2) and **6.92%** (FLAN-T5-XXL) average gains over vanilla prompting across five datasets.
  - **Cost Efficiency:** Single-agent principle approach matches **stepback prompting** performance with **half the inference cost**.
  - **Multi-Agent further** boosts performance over single-agent
  - Consolidation (cooperative) outperforms **ranking** (competitive) and **random selection** approaches
- **Principle Quality Comparison:**
  - **LLM-generated principles** matches or outperform **human-crafted principles** on private datasets
- **Key Takeaway:**
  - Principle-based ICL is a **cost-effective alternative** when labeled data is scarce.
  - Maintains **zero-shot/few-shot efficiency** (no fine-tuning) while closing the gap with supervised models.

Principle-based vs. Few-shot ICL:

- **Principle-based approach** achieves **competitive or superior performance** to few-shot ICL with **shorter input tokens**
- **Diminishing returns** with scaling the number of demonstrations
- Principle-based prompting offers a **token-efficient alternative** to traditional few-shot ICL, especially for long-context tasks

Table 2: Absolute improvements in the macro-F1 scores over the zero-shot vanilla prompting for the few-shot versus zero-shot principle-based approaches. Results are averaged across five inferences with different random seeds. n indicates the number of demonstrations per class. For PC1 and PC2, experiments were limited to $n \leq 2$ due to out-of-memory errors caused by long input token lengths.

Dataset	Model	n=1	n=2	n=4	n=8	multiagent principle consolidation
irony2018	flan-t5-xxl	0.62	0.08	0.06	**0.68**	0.45
	flan-ul2	3.63	3.08	3.64	3.66	**4.77**
emotion20	flan-t5-xxl	7.82	4.17	1.92	2.58	**12.13**
	flan-ul2	0.94	1.28	0.32	0.92	**15.11**
financial	flan-t5-xxl	1.57	2.26	2.28	2.70	**4.38**
	flan-ul2	8.22	10.42	11.49	11.32	**14.17**
PC1	flan-t5-xxl	0.22	**1.49**	NA	NA	1.43
	flan-ul2	**0.59**	0.47	NA	NA	0.04
PC2	flan-t5-xxl	**17.36**	17.31	NA	NA	16.21
	flan-ul2	16.98	17.41	NA	NA	**19.37**



Future Work:

- **Methodological Extensions:**
  - **Integration with RAG**
  - **Multi-Label Classification:** e.g., generate per-class principles + retrieval for top-k candidates
  - **Hybrid Approaches:** pairing principles with **example-based explanations** to boost performance.
- **Model & Application Expansion:**
  - **Black-Box LLMs:** Test scalability with models like **GPT-4** to assess broader applicability.
  - **SOP Automation:** Extend the multi-agent framework to auto-generate **domain-specific SOPs** (e.g., legal, medical) from minimal examples.
  - **Beyond Classification:** Apply principle-based approach to **generation tasks** (e.g., summarization, QA) requiring structured reasoning.

AAAI 2025 Workshop on Advancing LLM-Based Multi-Agent Collaboration

https://arxiv.org/abs/2502.07165

*Corresponding author: peipeiw@amazon.com