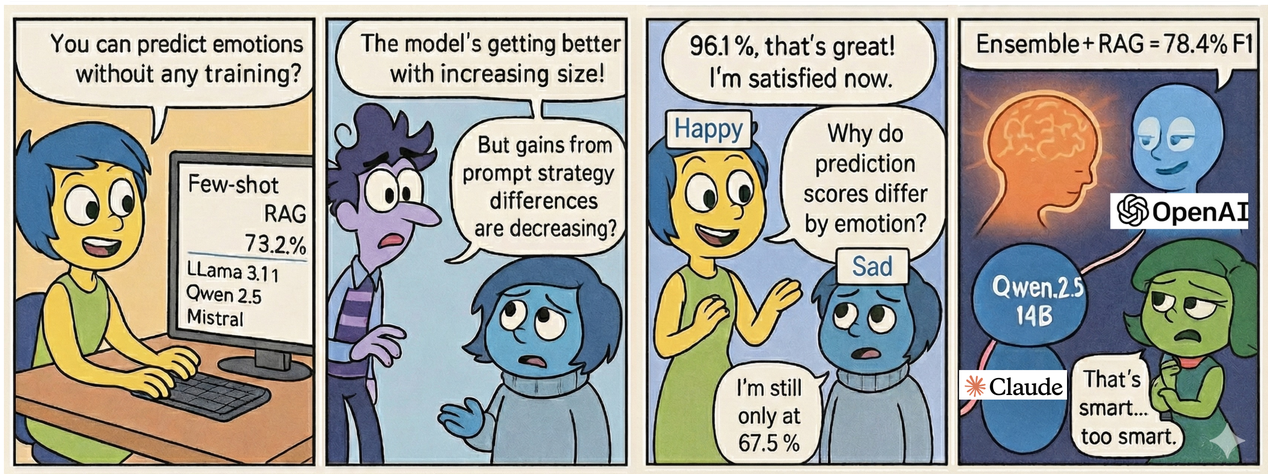
Systematic Benchmarking of Prompt Engineering Strategies for LLM-based Emotion Recognition

RCV 초세대 협력 워크샵

Jungjoon Park[1,2], Nikoo Mashayekh[2], Arasu Narayan[3,4]

[1] KAIST, [2] Neuromatch, [3] Oracle Healthcare, [4] Northeastern University

Graphical Abstract



Motivation

• Traditional **emotion recognition** often relies on **costly**, dataset- and task-**specific fine-tuning**, with **limited generalization** across domains.

The Paradigm Shift: Pre-trained **LLMs** encode rich **affective knowledge**, enabling zero-shot and **few-shot** emotion classification through **prompting**[1].

Research Gap: The systematic interaction between individual **prompt components** and **model scale** on emotion recognition remains underexplored.

Objective: Benchmark commercial and open-source models using a **modular prompt design,** and analyze how scaling effects text-based emotion classification.

Methodology

Dataset

• ISEAR[2] (7,328 samples, Text Modality)

• 7 categorical emotions (Joy, Fear, Anger, Sadness, Shame, Guilt, Disgust).

Models

• **Commercial:** GPT-4o, Claude-3.5 Sonnet.

• **Open-Source:** Qwen 2.5, Gemma 3 (ranging from 1B to 14B parameters).

RAG (Retrieval-Augmented Generation) based Context Retrieval[3]

• Use the **retrieved contexts** as **dynamic few-shot** prompts, enabling nuanced emotion recognition without additional task-specific training.

• **Retrieval:** k-NN (k=7) via FAISS vector database (BGE-M3 embedding).

• **Filtering:** Similarity threshold (L2 distance < 0.1).

Proposer–Aggregator LLM Ensemble[4]

• GPT-4o and Claude-3.5 act as **Proposers**, while Qwen 2.5 14B or Gemma3 12B serves as the **Aggregator** to synthesize the final prediction.

Compositional Prompt

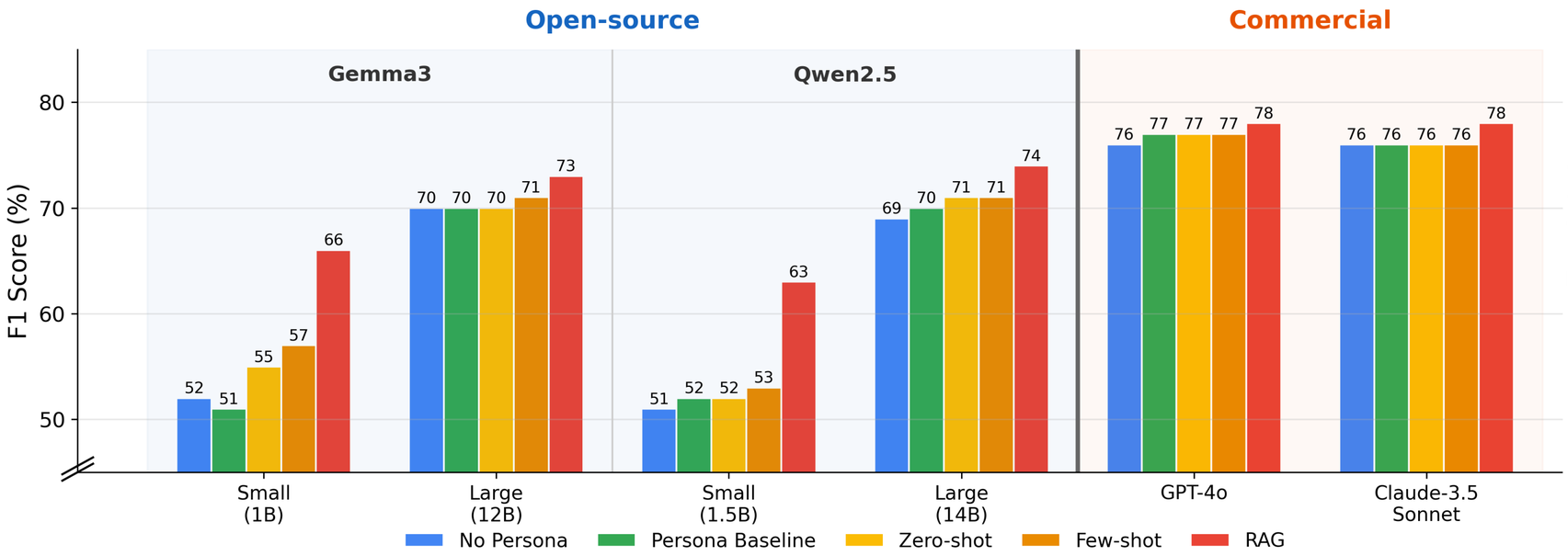
• A 5-stage **modular prompting scheme** incrementally adds persona, definition, and in-context examples.

Experimental Results

Prompting Gains & Scaling Paradox

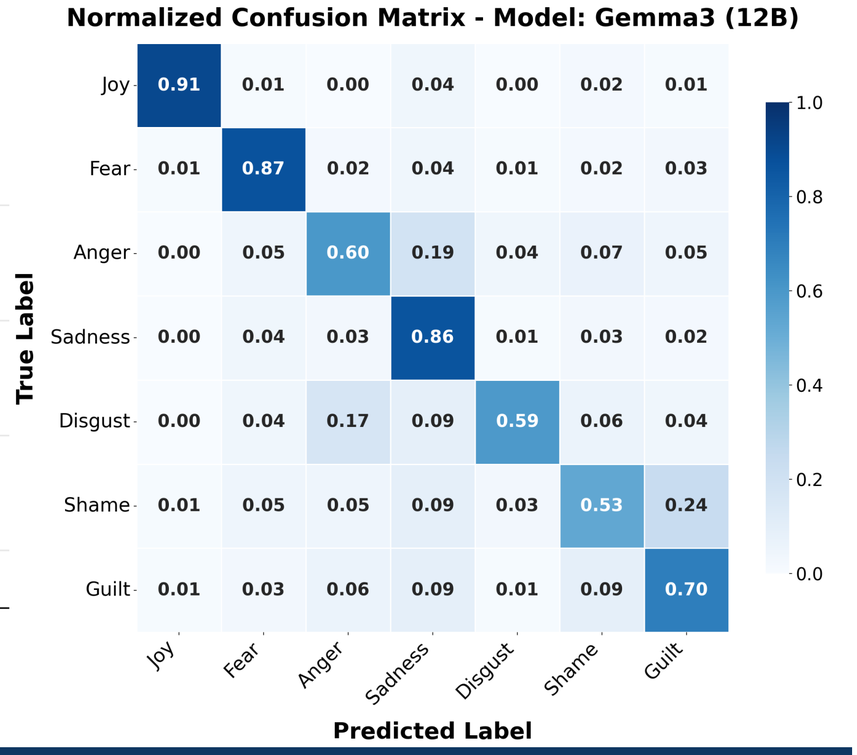
• Macro-F1 score consistently **increases** as more structured contextual modules are incorporated into the prompt.

• Advanced prompting (few-shot, RAG) yields substantial gains for smaller models, **whereas larger models** (≥10B) exhibit robust baseline performance with **diminishing returns** from complex prompts.



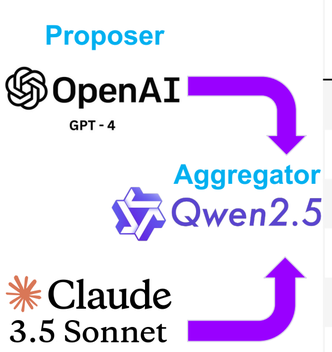
Class-specific performance

• Shame and guilt are among the lowest-scoring and show frequent mutual confusions.



Ensemble-Based SOTA Performance

• Ensemble leverages complementary strengths of commercial LLMs, with Qwen 2.5-14B as aggregator achieving a peak macro-F1 of 78.4% on ISEAR.



	Model	Prompt	joy F1	fear F1	anger F1	sadness F1	disgust F1	shame F1	guilt F1	macro avg F1
0	GPT-4o	RAG	96.1	85.3	71.6	77.4	74.6	67.5	73.5	78.0
8	Claude 3.5 Sonnet	RAG	96.0	85.6	70.3	78.2	74.0	67.5	75.9	78.2
16	Qwen 2.5 14B	RAG	95.0	82.7	64.7	74.0	71.4	62.1	68.0	74.0
17	Qwen 2.5 14B	Ensemble	96.1	85.8	70.7	78.1	74.6	67.8	76.0	78.4
31	Gemma3 12B	RAG	94.0	81.9	63.2	72.5	69.5	62.5	66.9	72.9
32	Gemma3 12B	Ensemble	96.1	85.6	70.6	78.1	74.2	67.4	76.0	78.3

[1] J. Zhou et al., "EmoLLM: Multimodal Emotional Understanding Meets Large Language Models," arXiv:2406.16442, 2024.
[2] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," Journal of Personality and Social Psychology, vol. 66, no. 2, pp. 310–328, 1994.
[3] Z. Huang et al., "Emotional RAG: Enhancing Role-Playing Agents Through Emotional Retrieval," arXiv:2410.23041, 2024.
[4] T.-M. Lin et al., "NYCU-NLP at EXALT 2024: Assembling large language models for cross-lingual emotion and trigger detection," Proc. WASSA, 2024.
[5] J. P. Tangney et al., "Are shame, guilt, and embarrassment distinct emotions?," Journal of Personality and Social Psychology, vol. 70, no. 6, pp. 1256–1269, 1996.

Discussions

• **Shame–guilt confusion** persists due to shared negative valence and overlapping linguistic markers, in line with self-conscious emotion theory[5].

• **Ensemble** results: Commercial model predictions themselves act as strong contextual signals; controlled ablations are needed to isolate their contribution from other prompt components.

• **Generalization:** Apply the modular prompting + ensemble framework to additional text corpora (e.g., GoEmotions) and languages to test robustness beyond ISEAR.

• **Beyond categorical labels:** Extend to **multimodal** and **dimensional** emotion models (e.g., valence–arousal) to represent continuous, context-dependent affective states.