Fine-Grained Emotion Recognition via In-Context Learning

Zhaochun Ren*
Leiden University
Leiden, The Netherlands
z.ren@liacs.leidenuniv.nl

Zhou Yang*
College of Computer and Data Science, Fuzhou University
Fuzhou, China
zhouzhouyang520@126.com

Chenglong Ye
College of Computer and Data Science, Fuzhou University
Fuzhou, China
231020027@fzu.edu.cn

Haizhou Sun
SmartMore Company
Shenzhen, China
sunhaizhou.ai@gmail.com

Chao Chen
School of Computer Science and Technology, Harbin Institute of Technology
Shenzhen, China
cha01nbox@gmail.com

Xiaofei Zhu
College of Computer Science and Engineering, Chongqing University of Technology
Chongqing, China
zxf@cqut.edu.cn

Xiangwen Liao†
College of Computer and Data Science, Fuzhou University
Fuzhou, China
liaoxw@fzu.edu.cn

Abstract

Fine-grained emotion recognition aims to identify the emotional type in queries through reasoning and decision-making processes, playing a crucial role in various systems. Recent methods use In-Context Learning (ICL), enhancing the representation of queries in the reasoning process through semantically similar examples, while further improving emotion recognition by explaining the reasoning mechanisms. However, these methods enhance the reasoning process but overlook the decision-making process. This paper investigates decision-making in fine-grained emotion recognition through prototype theory. We show that ICL relies on similarity matching between query representations and emotional prototypes within the model, where emotion-accurate representations are critical. However, semantically similar examples often introduce emotional discrepancies, hindering accurate representations and causing errors. To address this, we propose Emotion In-Context Learning (EICL)[1], which introduces emotionally similar examples and uses a dynamic soft-label strategy to improve query representations in the emotion reasoning process. A two-stage exclusion strategy is then employed to assess similarity from multiple angles, further optimizing the decision-making process. Extensive experiments show that EICL significantly outperforms ICL on multiple datasets.

CCS Concepts

• **Information systems** → **Sentiment analysis**.

*Both authors contributed equally to this research.
†Corresponding author.
[1]The model was developed on the high-availability platform SCITIX (SGP TECH PTE), which provides cost-effective GPU resources. The code is available at EICL.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761319

Keywords

Emotion Recognition, In-Context Learning, Large Language Models

ACM Reference Format:
Zhaochun Ren, Zhou Yang, Chenglong Ye, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2025. Fine-Grained Emotion Recognition via In-Context Learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761319

1 Introduction

Emotions [1–3] play a critical role in shaping how people seek, interpret, and respond to information. In applications such as search engines [4, 5], recommender systems [6, 7], and mental health support [8, 9], user queries often contain not only explicit information needs but also implicit emotional expressions. Accurately identifying these emotional cues can enhance search relevance [10, 11] and user satisfaction [12, 13]. To this end, the task of fine-grained emotion recognition has emerged, aiming to classify the emotional categories in queries through reasoning and decision-making process [14, 15].

Early studies train small-scale models to adjust emotion reasoning and decision-making for specific datasets, achieving promising results [16–20]. These methods are limited by model size and specific data, making them difficult to adapt to new data distributions and unseen emotions [21–24]. Recent studies employ In-Context Learning (ICL), which flexibly adjusts the reasoning and decision-making process of large language models (LLMs) using only semantically similar examples, thus enhancing the emotion recognition and generalization [24, 25]. These methods rely on empirically constructed examples and lack an understanding of ICL's internal mechanisms, limiting improvements in emotion recognition. Meanwhile, other studies explore ICL's internal mechanisms, examining how it integrates example information into query representations from Bayesian [26–30], gradient [31–34], algorithmic learning [35–38], and information flow [39] perspectives to facilitate emotion reasoning, as shown in Figure 1(a). However, emotion recognition involves both reasoning and decision-making processes, and these

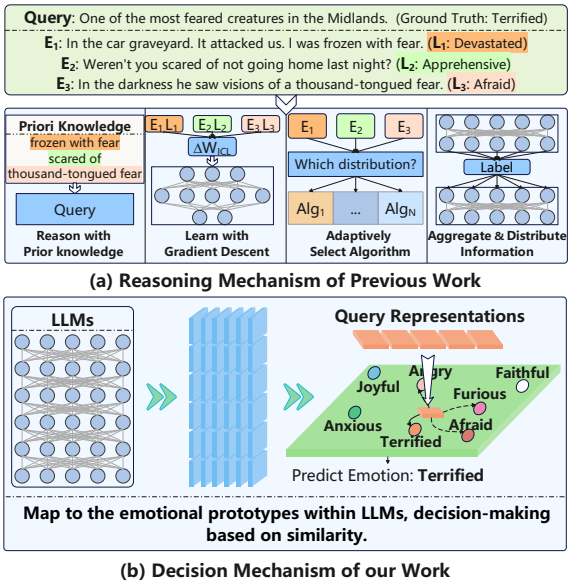

Figure 1: The reasoning and decision-making mechanism of in-context learning (ICL).

studies focus only on the reasoning process, i.e., how query representations form, neglecting the decision-making process, i.e., how query representations are transformed into final predictions.

In this paper, we investigate the decision-making mechanism in ICL for fine-grained emotion recognition. Inspired by neuroscience research on hidden representations in LLMs [40–42], we propose a prompt-pair detection method to reveal that LLMs represent emotion categories with specific hidden representations. In the ICL decision-making process, the more similar a query representation is to a category representation in LLMs, the more likely it is to predict the corresponding emotion, as shown in Figure 1(b). Viewing category representations as emotional prototypes, this similarity matching phenomenon suggests that ICL's decision-making process aligns with prototype theory [43–45].

From the perspective of prototype theory, we identify a flaw in ICL: **During the reasoning process**, semantically similar examples contribute little to building high-quality query representations in emotion recognition. For example, given the query "*I'm worried about the upcoming major meeting.*" the semantically similar example "*I'm anticipating the upcoming major meeting.*" shares only semantic content and contributes little to emotion reasoning. In contrast, the emotionally similar example "*The eve of a major event often causes anxiety.*" aligns with the query's emotional tone, supplying richer information for emotion reasoning and helping to foster a high-quality query representation. **During the decision-making process**, relying on similarity between query representations and emotion prototypes amplifies errors when those representations are inaccurate. Semantically similar examples offer little for emotion reasoning, making it hard to form emotionally precise query representations. Under the similarity matching mechanism, the model compares these flawed representations with the LLM's internal emotion prototypes to infer the query's emotion. Since the query representations lack emotional precision, the resulting similarity scores are unreliable, leading to incorrect judgments.

To address this issue, we propose a simple yet effective **E**motion **I**n-**C**ontext **L**earning method (abbreviated as EICL) for fine-grained emotion recognition. It introduces emotionally similar examples and uses a dynamic soft-label strategy to accurately depict their emotions, enhancing emotion reasoning and forming high-quality representations. Additionally, it use a two-stage exclusion strategy to assess similarity from multiple angles, optimizing the decision-making process. We perform experiments with five LLMs across four fine-grained emotion datasets: EDOS [46], Empathetic Dialogues [47], EmpatheticIntent [48], and GoEmotions [49]. The results show that EICL significantly outperforms ICL in fine-grained emotion recognition.

To sum up, our contributions are as follows:

(i) We introduce a prototype theory perspective to explain ICL's decision-making mechanism, highlighting its reliance on similarity matching between queries and emotional prototypes in LLMs, and addressing the gap in previous work that focused only on the reasoning process.
(ii) We propose EICL, offering a comprehensive reasoning and decision-making approach by using emotionally similar examples and a dynamic soft-label strategy to improve emotion reasoning, while optimizing decision-making through a two-stage exclusion strategy.
(iii) Extensive experiments and analysis show that the proposed method outperforms ICL on multiple datasets.

2 Related Work

In this paper, we introduce a prompt-pair detection method inspired by neuroscience-based prompting to examine the decision-making process of in-context learning. Drawing on these insights, we refine our in-context learning approach for fine-grained emotion recognition.

2.1 Neuroscience-based Prompting Methods

Driven by neuroscience advances, recent research [40, 41] treats LLM's internal parameters as neural nodes and probes their activations to understand or steer model behavior. Neuroscience-based Prompting Methods [42, 50, 51], valued for their simplicity and generality, have been applied across diverse tasks. Zou et al., [50] use paired positive and negative prompts to extract concept vectors and steer outputs toward honesty, detoxification or ethical framing. Turner et al., [51] apply contrastive prompting to derive steering vectors that modulate topic and sentiment through targeted interventions in hidden layers. Liu et al., [42] leverage prompts to capture honesty and confidence signals and then use these signals to retrieve and generate trustworthy responses. Leong et al., [52] control the toxification direction and manipulates information flow within attention layers to remove toxic content. These studies all use prompting to create stable concept representations, which are then leveraged to guide LLM behavior on specific tasks, yielding strong and versatile performance. Unlike prior methods that apply concept representations to task-specific control, we use the extracted representations to examine LLM decision behaviors and reveal their internal mechanisms.

2.2 In-Context Learning methods

Basic In-Context Learning. In-Context Learning (ICL) enhances LLMs' performance by learning from constructed examples, avoiding the time and computational costs of fine-tuning. One ICL approach improves LLMs by decomposing reasoning steps of examples into sub-steps, enabling the model to complete tasks by following these steps [53–55]. This method has shown strong results in tasks like arithmetic [56], commonsense [53], and symbolic reasoning [57], but requires manual construction and is not always applicable to tasks that can't be easily decomposed. Another approach, retrieval-based ICL, addresses this by retrieving relevant examples from training datasets. It focuses on examples similar to the query in terms of words [58–60], semantics [61–64], structures [65], or other relevant aspects [66–68]. Most methods rely on the semantic similarity between the query and examples.

In-Context Learning on Emotion Recognition. ICL on fine-grained emotion recognition can be categorized into heuristic-based ICL and exact-based ICL. Heuristic-based ICL enhances emotion recognition by adjusting the reasoning and decision-making processes of LLMs using semantically similar examples [24, 25]. While heuristic-based ICL relies on empirically constructed examples, it lacks an understanding of ICL's internal mechanisms, limiting its effectiveness. In contrast, exact-based ICL analyzes the reasoning process from multiple perspectives, such as Bayesian [26–30], gradient descent [31–34], algorithmic learning [35–38], and information flow [39], to improve query representations and emotion reasoning. However, while these studies explore reasoning, few address the internal mechanisms of decision-making based on these representations. Unlike previous work on reasoning and mechanics, we explore ICL's decision-making, emphasizing similarity matching. We then propose emotion in-context learning to enhance ICL's performance in fine-grained emotion recognition.

3 Investigating Decision-Making in ICL

Previous methods explore the internal reasoning mechanism of In-Context Learning (ICL), showing that semantically similar examples help form higher-quality query representations in the hidden layers of large language models (LLMs), promoting emotion reasoning [26, 27, 31, 35, 36, 39]. However, how these query representations map to emotion categories remains unclear. Recently, the linear representation and superposition hypotheses [40, 41] suggest that specific hidden representations in LLMs represent distinct concepts, with LLMs moving closer to these representations when expressing them. This phenomenon offers a new perspective on the decision-making process in ICL. To this end, we propose a prompt-pair detection method to extract category-related representations from LLMs' hidden representations and investigate the relationship between query and category representations during decision-making.

3.1 Prompt-pair Detection Method

The prompt-pair detection method aims to extract stable category representations. To achieve this, we collect representations of emotion categories in different semantic contexts and extract stable representations from them.

Table 1: Positive Prompts for the category c_i.

From the perspective of the emotion [**Emotion** c_i], infer the dialogue. Dialogue Context: [**Sample** s_j]. Output Format: 'Emotion: [the inferred emotion]'

Specifically, for an emotion category c_i, we select M samples from a set S that conveys the corresponding emotion. For each sample $s_j \in S$, we construct a positive prompt P^+ and a negative prompt P^-. The difference between the positive and negative prompts is that the positive prompt uses the target emotion category c_i, whereas the negative prompt randomly selects a category from the complete emotion category set C in the datasets. The positive prompt is shown in Table 1. Both prompts are then fed into LLMs to predict the emotion category of the sample. During category prediction, we adopt a curriculum learning strategy that guides the LLMs to generate the corresponding tokens step by step, as formalized below:

$$y_t^+ = LLM(y_t^+ | P^+, y_{<t}^+), \tag{1}$$

$$y_t^- = LLM(y_t^- | P^-, y_{<t}^-), \tag{2}$$

where y_t^+ and y_t^- represent the tokens generate by the LLM at timestep t, respectively. $y_{<t}^+$ and $y_{<t}^-$ represent the tokens generated before time step t using the positive and negative prompts, respectively.

As LLMs are required to produce grammatically, semantically, and emotionally coherent content, their hidden states encode both contextual information (e.g., syntax and semantics) and emotional information. To decouple emotion for category prediction, we construct prompt pairs that share all context but differ only in the emotion, then subtract the hidden state of the negative prompt from that of the positive prompt to remove shared contextual information and isolate the emotion. Concretely, at each time step t, we extract the hidden representations $h_{t,s_j}^{l,+}$ and $h_{t,s_j}^{l,-}$ for the positive and negative prompts, respectively, and compute their difference to obtain the category representation h_{t,s_j}^{l}, following established methods [42, 50, 51]. We have:

$$h_{t,s_j}^{l} = h_{t,s_j}^{l,+} - h_{t,s_j}^{l,-}, \tag{3}$$

where $h_{t,s_j}^{l}, h^{l,+}t, s_j$, and $h^{l,-}t, s_j \in \mathbb{R}^d$. $h^{l,+}t, s_j$ and $h^{l,-}t, s_j$ denote the hidden representations at time step t for token s_j under the positive and negative prompts, respectively, and h_{t,s_j}^{l} is the resulting category representation. l indicates the l-th layer of the LLM.

We then collect all category representations derived from each sample into the set $S_{c_i}^l$ and apply principal component analysis (PCA) to extract their first principal component. Through this extraction, we obtain a common and stable representation $H_{c_i}^l \in R^d$ of category c_i across different samples. We have:

$$H_{c_i}^l = PCA(S_{c_i}^l). \tag{4}$$

3.1.1 Category Representation Visualization. Representations produced by neuroscience-based prompting methods are stable and robust [42, 50, 51]. As a neuroscience-based prompting method, our prompt-pair detection method inherits these advantages when constructing the category representation $H_{c_i}^l$. Nevertheless, we further

validate them using Llama3.1$_{8b}$ on the ED dataset [47]. We average $H^l_{c_i}$ layer by layer to obtain H_{c_i}.

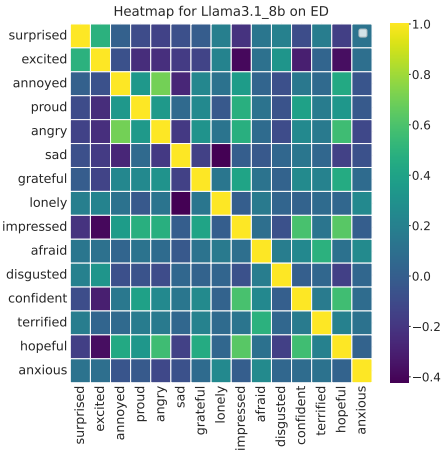

Figure 2: Heatmap of category representations.

Based on the representations H_{c_i}, we compute cosine similarities between category vectors and normalize the scores to [0, 1], as shown in Figure 2. The results reveal that categories of similar emotions yield higher similarity scores, while dissimilar ones yield lower scores. Overall, these results confirm the validity of our category representations and lay a solid foundation for the following investigation.

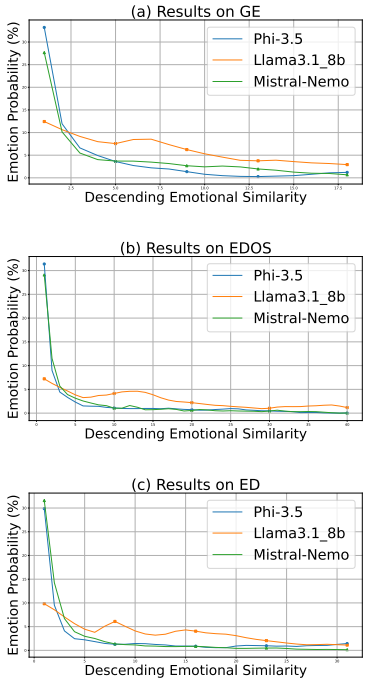

Figure 3: Emotion probability as similarity decreases.

3.2 Investigating Decision-Making with Prototype Theory

Based on these representations, we further explore whether ICL aligns with specific category representations during the decision-making process.

Previous research [39] suggests that ICL consolidates important information at the critical time step t_k, where the hidden representation determines the prediction outcome. For instance, in the emotion recognition task, when the generated response is "Emotion: sad," the hidden representation information at the time step of generating the ":" determines the prediction as "sad." Therefore, at this step, we compute the dot product between the query hidden representation H^l and the category representation $H^l_{c_j}$, we have:

$$o_h = \frac{1}{L}\sum_{l=1}^{L} H^l \cdot H^l_{c_j}, \tag{5}$$

where c_j and L are an emotion category in the dataset and the number of LLM layers, respectively.

We conduct experiments using Phi-3.5-mini, Mistral-Nemo, and Llama3.1$_{8b}$ on the EDOS [46], Empathetic-Dialogues (ED) [47]and GoEmotions [49] datasets. Figure 3 shows the results, where the x-axis represents emotion categories sorted by dot product (similarity) in descending order, and the y-axis represents the probability of predicting each emotion. The results show that as the similarity decreases, the probability of predicting the emotion also decreases. From the perspective of prototype theory [43–45], treating category representations as prototypes, we find that the closer a query hidden representation is to the emotional prototype, the higher the probability of predicting the corresponding emotion. **This suggests that ICL's decision-making process is driven by similarity matching, consistent with prototype theory**.

4 Methodology

4.1 Preliminaries

Problem Formulation. We formalize the task as follows: Given a query q_i, the goal is to construct an effective prompt that guides the large language model (LLM) to accurately predict the emotion category c_{q_i}.

Overview. The proposed EICL is an in-context learning method supported by an emotion auxiliary model. As shown in Figure 4, EICL consists of two steps: (i) **Emotion Reasoning** (in Section 4.2): It retrieves emotionally similar samples to aid reasoning and applies a dynamic soft-label strategy to improve query representations in emotion reasoning process. (ii) **Emotion Decision** (in Section 4.3): It divides emotion categories into primary and secondary candidates, prompting the LLM to prioritize primary candidates during decision-making, while considering secondary candidates afterward. This reduces decision errors caused by relying solely on similarity. Note that this method method requires no training and relies on a pre-trained model RoBERTa$_{emo}$ with emotional capabilities to complete the task (for details, see Section 5).

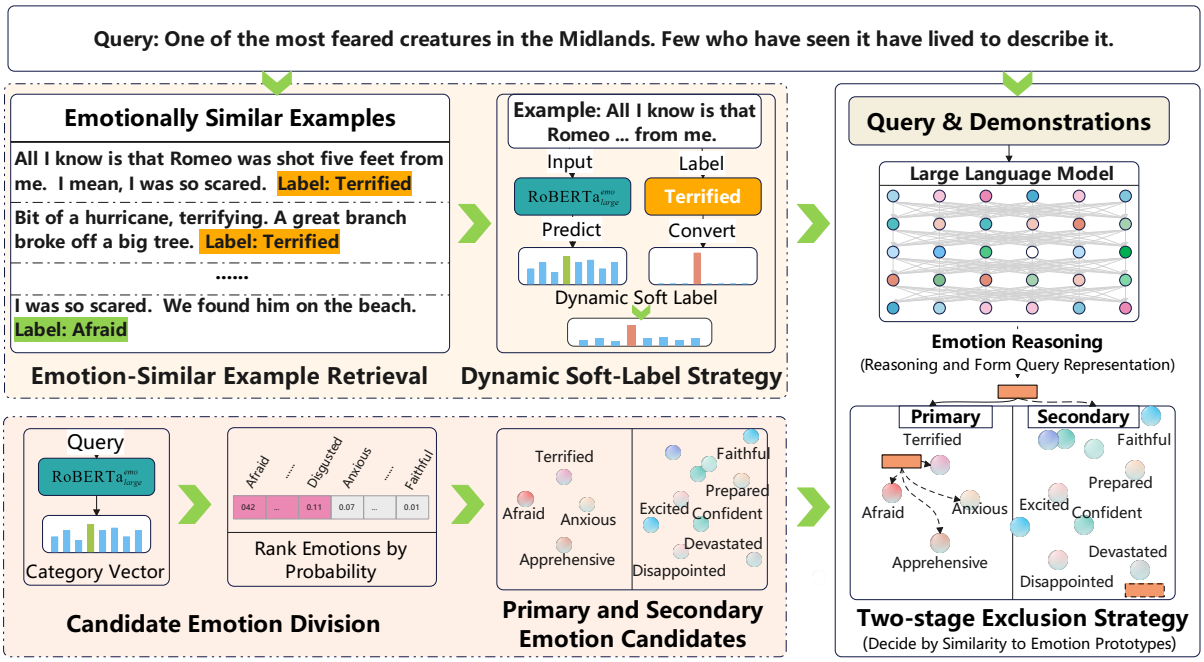

Figure 4: Overview of EICL. It retrieves emotionally similar examples and uses a dynamic soft-label strategy to accurately depict their emotions, enhancing query representations in the emotion reasoning process. A two-stage exclusion strategy is then applied, prioritizing primary emotion candidates to avoid decision errors from relying solely on similarity, ensuring accurate decision-making.

4.2 Emotion Reasoning

We retrieve emotionally similar examples and use a dynamic soft-label strategy to accurately depict the emotions they contain, thereby enhancing emotion reasoning in LLMs.

4.2.1 Emotion-Similar Example Retrieval. Previous ICL methods retrieve semantically similar prototypes, yet these can be emotionally misaligned or even contradictory to the query, degrading prediction accuracy [69–71]. To address this, we employ an auxiliary emotion model to retrieve emotionally congruent examples. Specifically, we map each test query $q_i \in D_{test}$ and each training sample $s_{m_i} \in D_{train}$ into emotion vectors using RoBERTa$_{emo}$, compute their cosine similarity o_{m_i}, rank all samples by o_{m_i}, and select the top-k_1 as the emotion-similar examples s_j. We have:

$$v_{q_i} = RoBERTa_{emo}(q_i), v^s_{m_i} = RoBERTa_{emo}(s_{m_i}), \tag{6}$$

$$o_{m_i} = Cosine(v_{q_i}, v^s_{m_i}), m_i \in n_d, \tag{7}$$

$$s_j = Top_{k_1}(o_1, o_2, ..., o_{m_i}), j \in [1, k_1], \tag{8}$$

where $v_{q_i}, v^s_{m_i} \in \mathbb{R}^{d_{emo}}$ denote the emotion vectors of query q_i and sample s_{m_i}, respectively. Top$_{k_1}$ returns the top-k_1 most similar samples, with k_1 as a hyperparameter. d_{emo} is the hidden-layer dimension of the emotion auxiliary model $RoBERTa_{emo}$. n_d is the size of the training set.

4.2.2 Dynamic Soft-Label Strategy. Emotions in linguistic expression are inherently complex and multifaceted [72–74]. Existing ICL methods [61, 62] assign only a single, deterministic emotion label to each example, oversimplifying this nuance. Consequently, ICL fails to incorporate genuinely emotion-aligned examples into its reasoning, resulting in inaccurate query representations. To address this issue, we use a dynamic soft-label strategy to assign specific labels to examples, accurately depicting their emotions to aid emotion reasoning. Specifically, we first employ the emotion auxiliary model to predict the emotions $e^s_{m_i}$ and their corresponding probabilities $p^s_{m_i}$ for each sample s_{m_i}; we then select the top k_2 emotions with the highest probabilities. The formal definition is as follows:

$$p^s_{m_i} = RoBERTa_{emo}(s_{m_i}), \tag{9}$$

$$e^s_k, p^s_k = Top_{k_2}(e^s_{m_i}, p^s_{m_i}), \tag{10}$$

where $p^s_{m_i}, p^s_k \in P, k \in [1, k_2], e^s_k \in C, m_k \in n_d$. P and C represent the model's predicted probabilities and the set of emotion categories. Top$_{k_2}$ is a ranking function that selects the top k_2 optimal emotions by their probabilities. k_2 is a hyperparameter.

Subsequently, we generate dynamic soft labels by combining predicted emotions with ground-truth labels, weighted by a hyperparameter α, so we have:

$$\hat{p}_i = \begin{cases} 1 - \alpha \sum_{k=1}^{k_2} p^s_k & \text{if } e^s_k = e^* \\ \alpha p^s_k & \text{Others}, k \in [1, k_2] \end{cases}, \tag{11}$$

where e^* is the ground truth label.

By combining emotions e_i with their corresponding probabilities $\hat{p}_i$, we obtain the dynamic soft label l_{m_i} for the sample s_{m_i}. Incorporating the sample s_{m_i} and its dynamic soft labels l_{m_i}, we

derive the example d_{m_i}. Then we concatenate the example d_{m_i} to obtain the examples d_{q_i} for query q_i.

$$l_{m_i} = (e_1, \hat{p}_1) \oplus (e_2, \hat{p}_2) \oplus ... \oplus (e_k, \hat{p}_i), \tag{12}$$

$$d_{m_i} = (s_{m_i}, l_{m_i}), \tag{13}$$

$$d_{q_i} = (d_1 \oplus d_2 \oplus ... \oplus d_{k_2}), \tag{14}$$

where $\oplus$ represents the concatenation operator.

4.3 Emotion Decision

Based on our findings in Section 3.2, ICL decides by measuring the similarity between the query representation and the LLM's internal emotion prototypes. However, when the query representation is emotionally inaccurate, relying solely on similarity may lead to errors. To address this issue, we propose a two-stage exclusion strategy that prioritizes certain emotion categories in emotion prediction, followed by others. This strategy considers both high similarity and prioritized emotion categories, mitigating errors caused by relying solely on similarity.

4.3.1 Candidate Emotion Division. Our strategy begins by dividing the emotion categories into primary and secondary emotion candidates. To achieve this, we apply the emotion auxiliary model to predict the query's emotions. We then select the top k_3 emotions with the highest probabilities and consider them as primary emotion candidates, which we place in the primary emotion set S_{pes}. The remaining emotions are considered as secondary emotion candidates and are placed in the secondary emotion set S_{ses}, so we have:

$$\widetilde{e}_m = Top_{k_3}(e_{q_i}, p_{q_i}), \tag{15}$$

$$\widetilde{e}_m \in S_{pes}, \widetilde{e}_n \in S_{ses}, \tag{16}$$

$$S_{ses} \cup S_{pes} = C, S_{ses} \cap S_{pes} = \emptyset, \tag{17}$$

where $\widetilde{e}_m, \widetilde{e}_n, e_{q_i} \in C, p_{q_i} \in P$. e_{q_i} and p_{q_i} are the emotion categories and probabilities predicted by the emotion auxiliary model for the query q_i, respectively. $\widetilde{e}_m$ and $\widetilde{e}_n$ represent the primary and secondary emotions, respectively. Top_{k_3} is a selection function that selects the k_3 emotion categories with the highest probabilities.

4.3.2 Two-stage Exclusion Strategy. Based on the above, we predict fine-grained emotions using a two-stage exclusion strategy. Specifically, we prompt LLMs to process the query and examples, prioritizing emotions from the primary emotion set S_{pes} before considering others. This strategy considers both the primary emotion categories and their similarity to prototypes, increasing their prediction probability and reducing decision errors. The prediction process is defined as follows:

$$c_{q_i} = LLM(q_i, d_{q_i}, S_{pes}, S_{ses}). \tag{18}$$

5 Experiments

Emotion Auxiliary Models and Datasets. To validate the proposed method, we conduct experiments using two emotion auxiliary models, RoBERTa$_{ei}$ and RoBERTa$_{ge}$, on four fine-grained emotion datasets: EDOS [46], Empathetic-Dialogues (ED) [47], EmpatheticIntent (EI) [48], and GoEmotions (GE) [49]. For convenience, we refer to the emotion auxiliary models and datasets as RoBERTa$_{emo}$ and D_{type}, where $emo \in EI, GE$ and $type \in EI, GE, ED, EDOS$. Note that our goal is to verify the performance of EICL without fine-tuning, so the emotion auxiliary model used during reasoning should not have been fine-tuned on the respective dataset, i.e., $emo \neq type$. Simultaneously, the emotion categories predicted by the emotion auxiliary model do not fully align with those of the datasets, rendering the exclusion strategy inapplicable. To address this issue, we adjust the datasets according to the emotion auxiliary model. For example, for the RoBERTa$_{ei}$ emotion auxiliary model [48] and the GoEmotions dataset, we first identify the emotion categories they share. Then, we select data from GE that falls within these common emotion categories for experimentation. After this adjustment, the available datasets for the RoBERTa$_{ei}$ emotion auxiliary model are GE, ED, and EDOS, with 19, 32, and 41 emotion categories, respectively. For the RoBERTa$_{ge}$ [2] emotion auxiliary model, the available datasets are EI, ED, and EDOS, with 19, 17, and 19 emotion categories, respectively.

Evaluation Metrics. We evaluate the methods using accuracy and macro-F1 (F1). Accuracy (Acc) measures the proportion of correctly predicted samples. F1 is the harmonic mean of precision and recall, considering both metrics. It accounts for each class's F1 score and is robust to class imbalance.

Baselines. To validate EICL, we conduct experiments on several large language models, including Phi-3.5-mini, Mistral-Nemo, Llama3.1$_{8b}$, Claude-Haiku, and ChatGPT-Turbo. For each model, we construct zero-shot learning (Z-shot) and in-context learning (ICL) as baselines.The zero-shot baseline considers only the query, while the in-context learning baseline includes examples semantically related to it.

Implementation Details. In our experiments, we use two emotion auxiliary models, RoBERTa$_{ei}$ and RoBERTa$_{ge}$, both with a hidden-layer dimension of d_{emo}=768. The former is applied to the GE, ED, and EDOS datasets, while the latter is used for EI, ED, and EDOS. During the construction of example-label pairs, we set the example number to k_1=5 and the weight for soft labels to α=0.2. The values of k_2 (the number of soft labels) and k_3 (the number of primary emotion candidates) vary based on the data, emotion auxiliary models, and LLMs. A detailed analysis of these factors is provided in Section 6.2.

6 Results and Analysis

6.1 Main Results

Tables 2 and 3 show the results with RoBERTa$_{ei}$ and RoBERTa$_{ge}$ as auxiliary models, respectively. The results demonstrate that ICL outperforms Z-shot across most metrics, indicating that semantically similar examples benefit emotional reasoning. Additionally, EICL outperforms both ICL and Z-shot on most metrics, primarily due to emotion-similar examples, dynamic soft-label strategies, and the two-stage exclusion strategy, all of which improve emotional reasoning and decision-making.

However, some anomalies are observed: (i) In datasets like GE (in Table 2) and EDOS (in Table 3), ICL performs worse than Z-shot. This is due to the models' weaker emotional capabilities. For instance, Llama3.1$_{8b}$ struggles to interpret emotions accurately, even

[2] https://huggingface.co/mrm8488/roberta-large-bne-finetuned-go_emotions-es

Table 2: Results on the datasets when using the emotion auxiliary model RoBERTa$_{ei}$.

Dataset		Phi-3.5-mini			Mistral-Nemo			Llama3.1$_{8b}$			Claude-Haiku			ChatGPT-Turbo		
—	—	Z-shot	ICL	EICL	Z-shot	ICL	EICL	Z-shot	ICL	EICL	Z-shot	ICL	EICL	Z-shot	ICL	EICL
EDOS	Acc	34.30	40.14	**52.36**	33.60	40.43	**56.15**	29.83	21.87	**39.30**	25.79	36.79	**54.23**	34.60	39.14	**54.45**
	F1	40.81	46.55	**55.97**	36.33	45.47	**60.13**	24.86	29.56	**48.38**	25.10	38.61	**52.78**	34.14	40.04	**54.37**
ED	Acc	29.0	37.33	**42.81**	37.24	37.64	**43.5**	35.03	37.50	**40.83**	41.73	49.47	**53.98**	36.40	42.87	**51.56**
	F1	28.27	39.50	**42.81**	34.80	38.47	**43.78**	29.93	39.92	**43.45**	36.70	47.01	**49.20**	29.82	41.43	**49.32**
GE	Acc	27.86	37.56	**38.75**	28.23	**40.03**	36.02	**36.07**	17.48	30.56	27.65	36.60	**38.05**	33.17	41.37	**46.10**
	F1	21.29	27.04	**31.22**	23.17	28.11	**29.88**	27.32	19.40	**35.86**	27.67	33.04	**36.80**	29.70	32.81	**37.19**

Table 3: Results on the datasets when using the emotion auxiliary model RoBERTa$_{ge}$.

Dataset		Phi-3.5-mini			Mistral-Nemo			Llama3.1$_{8b}$			Claude-Haiku			ChatGPT-Turbo		
—	—	Z-shot	ICL	EICL	Z-shot	ICL	EICL	Z-shot	ICL	EICL	Z-shot	ICL	EICL	Z-shot	ICL	EICL
EDOS	Acc	52.20	53.97	**54.85**	55.10	49.18	**62.92**	**36.69**	32.15	27.74	42.87	55.73	**62.16**	54.72	56.99	**60.4**
	F1	36.88	32.33	**54.81**	75.87	29.55	**76.37**	39.16	57.11	**58.08**	37.83	52.90	**57.74**	50.66	54.33	**57.0**
ED	Acc	44.47	49.33	**51.33**	48.86	41.77	**50.06**	**45.93**	41.57	32.22	53.22	61.81	**62.08**	57.62	58.18	**60.85**
	F1	21.89	29.94	**68.89**	51.80	17.32	**67.12**	20.64	18.79	**47.06**	51.81	**58.88**	57.99	55.37	56.27	**57.65**
EI	Acc	47.16	58.69	**60.62**	53.95	62.30	**62.55**	38.31	51.02	**59.56**	53.64	**67.85**	66.16	57.81	61.49	**63.05**
	F1	48.15	45.55	**80.46**	78.06	78.49	**80.75**	17.49	20.60	**59.01**	50.57	**64.81**	62.04	54.24	55.28	**59.91**

when beneficial examples are provided. (ii) EICL performs poorly on certain datasets. Llama3.1$_{8b}$ and Claude-Haiku, in particular, perform below baselines. This is primarily due to these models' difficulty in recognizing certain emotions, as their inherent limitations cannot be fully overcome, even with effective strategies and examples (detailed explanation, see Section 6.3). In contrast, models like Phi-3.5-mini and ChatGPT-Turbo, with stronger emotional perception, benefit more from the proposed methods. (iii) On the GE dataset, EICL on models like Mistral-Nemo and Llama3.1$_{8b}$ shows lower accuracy. This is primarily due to the dataset's bias, where the "neutral" category comprises 1606 samples out of 3442, accounting for 46.65% of the total data. In such a biased dataset, the F1 score is more reliable, and our method outperforms the baseline on this metric, demonstrating the effectiveness of the proposed method.

6.2 Analytical Experiments

6.2.1 Ablation Studies. Figure 5 presents ablation studies using the RoBERTa$_{ge}$ emotion auxiliary model on the EDOS, and ED datasets. Here, w/o EER, w/o DSL, and w/o TE represent the absence of emotion-similar example retrieval (in Section 4.2.2), Dynamic Soft-Label Strategy (in Section 4.2.1), and Two-stage Exclusion Strategy (in Section 4.3), respectively.

The results show that removing all modules leads to a decline in model performance, demonstrating the effectiveness of the modules. For ChatGPT-Turbo, removing Emotion-Similar Retrieval (W/O EER), the Two-Stage Exclusion Prediction Strategy (W/O TE), and the Dynamic Soft-Label Strategy (W/O DSL) all result in significant performance drops. This indicates that all three methods contribute to enhancing EICL's emotional reasoning and decision-making abilities. For Claude-Haiku, removing Emotion-Similar Retrieval (W/O EER) and the Two-Stage Exclusion Strategy (W/O TE) leads to a significant decline in performance. However, removing the Dynamic Soft-Label Strategy (W/O DSL) causes only a minor decrease. This suggests that Claude-Haiku already benefits from sufficient emotional reasoning and decision-making capabilities through similar examples and the two-stage exclusion strategy, while the dynamic soft-label strategy also helps with query understanding to some extent.

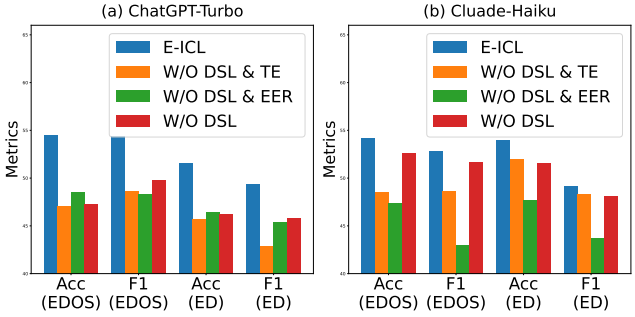

Figure 5: Ablation Results for EICL on EDOS and ED Datasets.

6.2.2 Impact of Dynamic Soft Label Weights. We investigate the impact of parameter α on model performance. α determines the weight of emotion probabilities predicted by the emotion auxiliary model in dynamic soft labels. A higher α indicates greater influence from the emotion auxiliary model. We consider two scenarios: one where the emotion auxiliary model's emotional capability exceeds that of the LLM, and another where it is weaker. As shown in Figure 6a, when using a strong emotion auxiliary model, EICL is insensitive to α since the model retrieves high-quality examples. However, with a weaker emotion auxiliary model, EICL performance increases initially and then decreases, as shown in 6b. This is mainly because

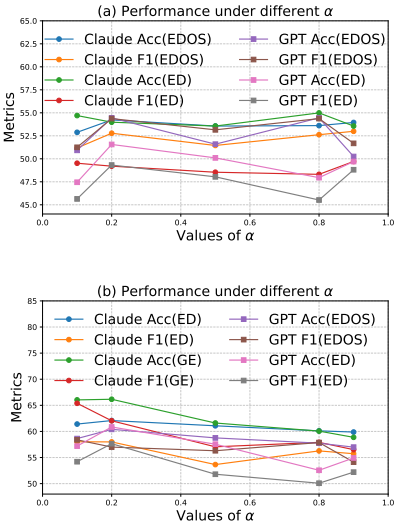

Figure 6: Results across varying α values on RoBERTa$_{ei}$ and RoBERTa$_{ge}$.

Table 4: Comparison between emotion auxiliary models and LLMs.

LLMs	EDOS		ED		EI	
—	Acc	F1	Acc	F1	Acc	F1
Claude	0.38	5.80	-12.58	-11.74	-12.40	-8.90
GPT	-11.47	-7.02	-16.98	-15.30	-16.57	-12.57
Phi	8.95	-6.75	3.83	-18.18	5.91	6.48
Mistral	11.85	32.24	8.22	11.73	12.71	36.39
Llama	-6.56	-4.47	5.29	-19.43	-2.93	-24.18
Claude	25.92	27.46	7.23	11.61	-2.87	-8.03
GPT	17.11	18.42	12.56	18.49	-8.39	-10.06
Phi	-17.41	-11.75	-19.96	-20.04	3.08	1.65
Mistral	-18.11	-16.23	-11.72	-13.51	3.45	3.53
Llama	-21.88	-27.7	-13.93	-18.38	11.29	7.68

the emotional types generated by the emotion auxiliary models are not accurate enough. Moderately considering these emotional types can improve performance, while over-reliance on them may be hindered by inaccurate judgments.

6.2.3 Impact of Dynamic Soft Label Numbers. We evaluate the impact of the number of dynamic soft labels on EICL, with results shown in Figure 7. The experiments are divided into two groups: one where the emotion auxiliary model outperforms the LLMs, and another where it underperforms them. Figure 7a depicts results using stronger emotion auxiliary models, while Figure 7a shows those with weaker models. Comparing the two groups, we observe that as the number of soft labels increases: (1) The performance of the stronger capability group initially decreases, then improves. (2) The weaker capability group reaches a peak (or starts at a peak) before declining. These findings suggest that a moderate number of dynamic soft labels enhances emotion prediction. At the same time,

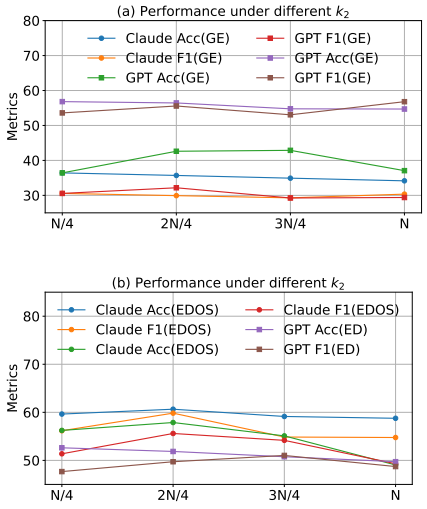

Figure 7: Results based on different k_2, where N is the number of emotion categories in the dataset.

increasing the number of dynamic soft labels leads to a performance drop, primarily because too many labels cause EICL to focus on irrelevant emotions, hindering its ability to reason about important emotions. Similarly, reducing the number of dynamic soft labels also results in a performance decline, as too few labels prevent EICL from deeply understanding the query with effective emotional examples.

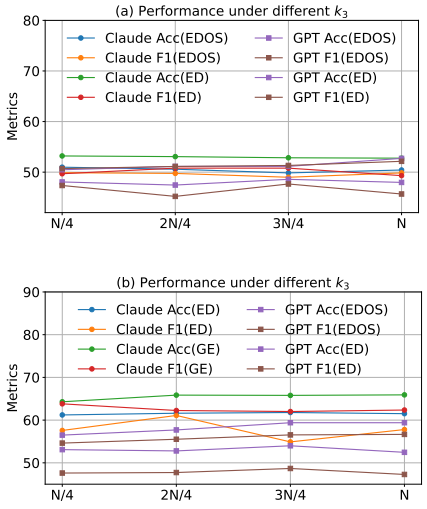

Figure 8: Results of EICL based on different k_3, where N is the number of emotion categories in the dataset.

6.2.4 Impact of Two-stage Exclusion Strategy. We evaluate the impact of the number of candidate emotions (k_3) on the two-stage exclusion strategy. The experiments are also divided into two groups

based on the emotional capability of the emotion auxiliary models: Figure 8a shows results with a strong emotion auxiliary model, while Figure 8b shows results with a weaker one. Most results suggest that using a moderate number of emotions as candidates yields optimal performance, highlighting the effectiveness of the two-stage exclusion strategy. In some cases, using all emotions as candidates leads to more accurate predictions, particularly when the emotion auxiliary model's performance is much lower, such as being 15 points below the LLMs (see Section 6.2.5). In these cases, EICL tends to exclude accurate emotions, causing the strategy to fail.

6.2.5 Impact of Emotion Auxiliary Model Performance. Table 4 shows the performance of RoBERTa$_{emo}$ (where $emo \in ge, ei$) compared to LLMs, with positive values indicating RoBERTa$_{emo}$ outperforms LLMs, and negative values indicating the opposite. For simplicity, models are referred to by their abbreviations. The first set of results compares LLMs with RoBERTa$_{ge}$, while the second set compares LLMs with RoBERTa$_{ei}$. The results reveal that, in most cases, the emotion auxiliary models perform significantly below the LLMs; however, EICL can further enhance LLM performance. This indicates that the proposed method does not require a powerful emotion model, but only needs a model with moderate emotional capability.

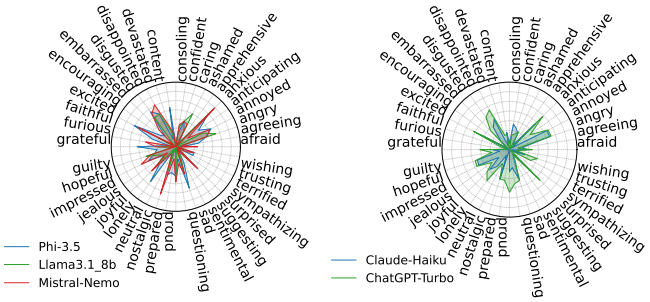

Figure 9: Comparison of emotional capability between smaller and larger LLMs on the EDOS dataset.

6.3 Emotional Capacity Analysis

6.3.1 Emotional Capacity Analysis of LLMs. LLMs are trained in diverse environments, such as different datasets and fine-tuning methods. This results in varying capabilities, particularly in emotional perception. To explore this, we analyze the emotional capabilities of smaller and larger LLMs.

Figure 9a illustrates the performance of smaller LLMs, showing that Mistral-Nemo and Phi-3.5-mini perform notably well, while Llama3.1$_{8b}$ demonstrates weaker capabilities, excelling only in the emotions of "ashamed," "angry," and "hopeful." Figure 9b presents the performance of larger LLMs, indicating that ChatGPT-Turbo offers a more comprehensive and balanced emotional capacity, while Claude-Haiku shows advantages only in "confident" and "caring" emotions. Overall, for similar LLMs, Llama3.1$_{8b}$ and Claude-Haiku display relatively weak emotional perception abilities. For larger LLMs, ChatGPT-Turbo has more comprehensive emotional capability compared to Claude-Haiku.

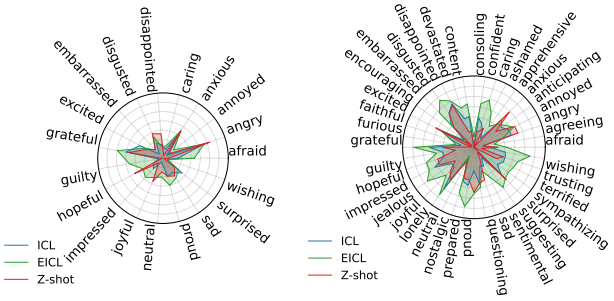

Figure 10: Comparison of emotional accuracy between EICL and baselines on the EDOS and ED Datasets.

6.3.2 Emotional Capacity Analysis of EICL. We analyze the emotional recognition accuracy of the Z-shot, ICL, and EICL methods, with the results shown in Figure 10. Figure 10a shows the results for Llama3.1$_{8b}$, while Figure 10b shows the results for ChatGPT-Turbo. According to the results, ICL shows a notable improvement over Z-shot. Compared to the first two methods, EICL demonstrates significant improvements in recognizing a wide range of emotions, highlighting the effectiveness of the method.

7 Conclusion

In this paper, we have examined the decision-making mechanism of in-context learning (ICL) in fine-grained emotion recognition from a prototype theory perspective. We have demonstrated that ICL's decision-making aligns with prototype theory and shown that semantically similar examples can cause errors in emotion reasoning and decision-making. Building on these insights, we have proposed a new perspective with emotion in-context learning, which enhances emotion reasoning with emotionally similar examples and dynamic soft-label strategies, and optimizes decision-making through a two-stage exclusion strategy. Experiments conducted on four datasets demonstrate that our method significantly outperforms traditional ICL methods.

This work uses emotional prototypes within LLMs to explore the decision-making mechanism of ICL in emotion recognition. Research has shown that LLMs contain not only emotional prototypes but also broader knowledge prototypes, allowing the proposed method to be applied to other tasks [40–42]. To further investigate ICL and LLMs' internal mechanisms, we will explore and validate them in more tasks in the future.

8 Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62476060). We would also like to express our gratitude to SCITIX (SGP TECH PTE) for providing the high-availability GPU platform, which significantly contributed to the computational resources required for this research.

GenAI Usage Disclosure

In this paper, we use the following generative AI models as baselines: Phi-3.5-mini, Mistral-Nemo, Llama3.1$_{8b}$, Claude-Haiku, and ChatGPT-Turbo. We also employ ChatGPT-o4 for manuscript translation. Throughout the process, we strictly adhered to all generative AI usage guidelines and policies, with no ethical risks involved.

References

[1] Reijo Savolainen. Emotions as motivators for information seeking: A conceptual analysis. Library & Information Science Research, 36:59–65, 2014.

[2] Mimi Zhang and Bernard J. Jansen. Influences of mood on information seeking behavior. In CHI, page 3395–3400, 2009.

[3] Maria Stavraki, Grigorios Lamprinakos, Pablo Briñol, Richard E. Petty, Kalipso Karantinou, and Darío Díaz. The influence of emotions on information processing and persuasion: A differential appraisals perspective. Journal of Experimental Social Psychology, 93:104085, 2021.

[4] Gabriella Kazai, Paul Thomas, and Nick Craswell. The emotion profile of web search. In SIGIR, page 1097–1100, 2019.

[5] Carlos Flavián-Blanco, Raquel Gurrea-Sarasa, and Carlos Orús-Sanclemente. Analyzing the emotional outcomes of the online search behavior with search engines. Computers in Human Behavior, 27(1):540–551, 2011.

[6] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. Towards empathetic conversational recommender systems. In RecSys, page 84–93, 2024.

[7] Erkang Jing, Yezheng Liu, Yidong Chai, Shuo Yu, Longshun Liu, Yuanchun Jiang, and Yang Wang. Emotion-aware personalized music recommendation with a heterogeneity-aware deep bayesian network. ACM Transactions on Information Systems, 2025.

[8] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In ACL, pages 308–319, 2022.

[9] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. Touch your heart: A tone-aware chatbot for customer care on social media. In CHI, pages 1–12, 2018.

[10] Ioannis Arapakis, Joemon M. Jose, and Philip D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In SIGIR, page 395–402, 2008.

[11] Irene Lopatovska and Ioannis Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. Information Processing & Management, 47(4):575–592, 2011.

[12] Bilyana Martinovsky and David Traum. The error is the clue: Breakdown in human-machine interaction. In ISCA Workshop on EH-SDS, pages 11–17, 2003.

[13] Hossein A. Rahmani, Xi Wang, Mohammad Aliannejadi, Mohammadmehdi Naghiaei, and Emine Yilmaz. Clarifying the path to user satisfaction: An investigation into clarification usefulness. In Findings of EACL, pages 1266–1277, 2024.

[14] Jasy Suet Yan Liew and Howard R Turtle. Exploring fine-grained emotion detection in tweets. In NAACL student research workshop, pages 73–80, 2016.

[15] Muhammad Abdul-Mageed and Lyle Ungar. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In ACL, pages 718–728, 2017.

[16] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In EMNLP, pages 2227–2240, 2021.

[17] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. In EMNLP, page 8968–8979, 2020.

[18] Yubo Xie, Ekaterina Svikhnushina, and Pearl Pu. A multi-turn emotionally engaging dialog model. arXiv preprint arXiv:1908.07816, 2019.

[19] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In AAAI, volume 33, pages 6818–6825, 2019.

[20] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In EMNLP-IJCNLP, pages 154–164, 2019.

[21] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is chatgpt equipped with emotional dialogue capabilities? arXiv preprint arXiv:2304.09582, 2023.

[22] Kristina Schaaff, Caroline Reinig, and Tim Schlippe. Exploring chatgpt's empathic abilities. In ACII, pages 1–8, 2023.

[23] Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. arXiv preprint arXiv:2402.11801, 2024.

[24] Yushan Qian, Weinan Zhang, and Ting Liu. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In Findings of EMNLP, pages 6516–6528, 2023.

[25] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In EMNLP, pages 6056–6077, 2023.

[26] Hui Jiang. A latent space theory for emergent abilities in large language models. arXiv preprint arXiv:2304.09960, 2023.

[27] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. ArXiv, abs/2111.02080, 2021.

[28] Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. In NeurIPS, 2023.

[29] Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism. arXiv preprint arXiv:2306.04891, 2023.

[30] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: explaining and finding good demonstrations for in-context learning. In NeurIPS, 2023.

[31] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In Findings of ACL, pages 4005–4019, 2023.

[32] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In ICML, pages 35151–35174, 2023.

[33] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In NeurIPS, 2023.

[34] Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. arXiv preprint arXiv:2307.03576, 2023.

[35] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. ArXiv, abs/2211.15661, 2022.

[36] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In NeurIPS, 2022.

[37] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In ICML, 2023.

[38] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: provable in-context learning with in-context algorithm selection. In NeurIPS, 2023.

[39] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In EMNLP, pages 9840–9855, 2023.

[40] Chris Olah. Distributed representations: Composition & superposition. Transformer Circuits Thread, 24, 2023.

[41] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. arXiv preprint arXiv:2311.03658, 2023.

[42] Huanshuo Liu, Hao Zhang, Zhijiang Guo, Jing Wang, Kuicai Dong, Xiangyang Li, Yi Lee, Cong Zhang, and Yong Liu. Ctrla: Adaptive retrieval-augmented generation via inherent control. arXiv preprint arXiv:2405.18727, 2024.

[43] Eleanor Rosch. Principles of categorization. In Cognition and categorization, pages 27–48. 1978.

[44] Hans Kamp and Barbara Partee. Prototype theory and compositionality. Cognition, 57(2):129–191, 1995.

[45] James A Hampton. Concepts as prototypes. Psychology of learning and motivation, 46:79–113, 2006.

[46] Anuradha Welivita, Yubo Xie, and Pearl Pu. A large-scale dataset for empathetic response generation. In EMNLP, pages 1251–1264, 2021.

[47] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In ACL, page 5370–5381, 2019.

[48] Anuradha Welivita and Pearl Pu. A taxonomy of empathetic response intents in human social conversations. In ACL, pages 4886–4899, 2020.

[49] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In ACL, pages 4040–4054, 2020.

[50] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405, 2023.

[51] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. arXiv e-prints, pages arXiv–2308, 2023.

[52] Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. Self-detoxifying language models via toxification reversal. In Houda Bouamor, Juan

Pino, and Kalika Bali, editors, EMNLP, pages 4433–4449, 2023.
[53] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS, 2022.
[54] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874, 2021.
[55] Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. LAMBADA: Backward chaining for automated reasoning in natural language. In ACL, pages 6547–6568, 2023.
[56] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In NeurIPS, 2020.
[57] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446, 2021.
[58] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In NAACL-HLT, pages 2655–2671, 2022.
[59] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. In Findings of ACL, pages 8857–8873, 2023.
[60] Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. Dr. icl: Demonstration-retrieved in-context learning. arXiv preprint arXiv:2305.14128, 2023.
[61] Xiaonan Li and Xipeng Qiu. MoT: Memory-of-thought enables ChatGPT to self-improve. In EMNLP, pages 6354–6374, 2023.
[62] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In DeeLIO, pages 100–114, 2022.
[63] Linyi Yang, Shuibai Zhang, Zhuohao Yu, Guangsheng Bao, Yidong Wang, Jindong Wang, Ruochen Xu, Wei Ye, Xing Xie, Weizhu Chen, et al. Supervised knowledge makes large language models better in-context learners. arXiv preprint arXiv:2312.15918, 2023.
[64] Chaojun Xiao, Zhengyan Zhang, Xu Han, Chi-Min Chan, Yankai Lin, Zhiyuan Liu, Xiangyang Li, Zhonghua Li, Zhao Cao, and Maosong Sun. Plug-and-play document modules for pre-trained models. In ACL, pages 15713–15729, 2023.
[65] Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In ACL, pages 1401–1422, 2023.
[66] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In ICLR, 2022.
[67] Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. In Findings of EMNLP, pages 10136–10148, 2023.
[68] Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. Compositional semantic parsing with large language models. In ICLR, 2022.
[69] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. Cognitive psychology, 7(4):573–605, 1975.
[70] J David Smith and John Paul Minda. Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(4):800, 2002.
[71] John Paul Minda and J David Smith. Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(3):775, 2001.
[72] Jeff T Larsen and A Peter McGraw. Further evidence for mixed emotions. Journal of personality and social psychology, 100(6):1095, 2011.
[73] Carlos Crivelli and Alan J Fridlund. Inside-out: From basic emotions theory to the behavioral ecology view. Journal of Nonverbal Behavior, 43(2):161–194, 2019.
[74] Debra Trampe, Jordi Quoidbach, and Maxime Taquet. Emotions in everyday life. PloS one, 10(12):e0145450, 2015.