

## Yanshu Li

Providence, RI, USA | yanshu\_li1@brown.edu | Gender: M | Birth Year: 2002

### Master



BROWN

**Brown University**, Master in Computer Science 2024.9-Present

**GPA:**4.0/4.0

**Core Courses:** Machine Learning, Self-supervised Learning

**Research Group:** Brown Language Understanding and Representation Lab (LUNAR)

**Advisor:** Ellie Pavlick



### Bachelor



**Soochow University**, Bachelor in Artificial Intelligence 2020.9-2024.6

**GPA:**3.8/4.0; **TOEFL:**107

**Core Courses:** Natural Language Processing, Deep Learning, TensorFlow Programming

**Research Interests:** Machine Learning, Natural Language Processing, Large Language Models

**Research Group:** Suda Natural Language Processing Lab

**Advisor:** Min Zhang



TL;DR: I am dedicated to bridging the gap between training data and real-world application scenarios in artificial intelligence methods, aiming to better meet user needs rather than solely striving for superior performance on fixed benchmarks.

## 1 Background

## 2 Experiences

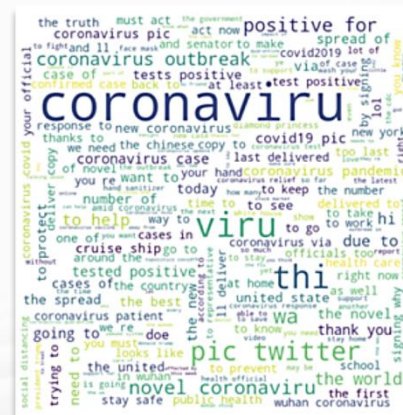
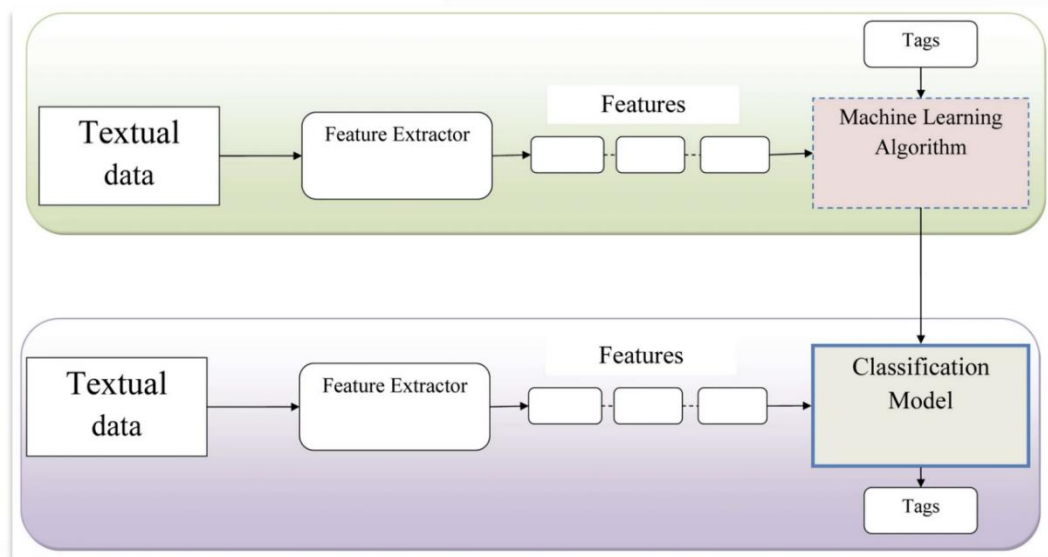
## 3 Research Plan



### Aspect-Based Sentiment Analysis

**Aspect-Based Sentiment Analysis (ABSA)**, also known as fine-grained opinion mining, is the task of determining the sentiment of a text with respect to a specific aspect. ABSA has extensive application scenarios even in the pre-LLM era, representing a core theme of human-centered NLP. When I began exploring NLP, my first project involved using the BERT model for COVID-19-related online platform sentiment analysis. This project was relatively preliminary and included two key contributions:

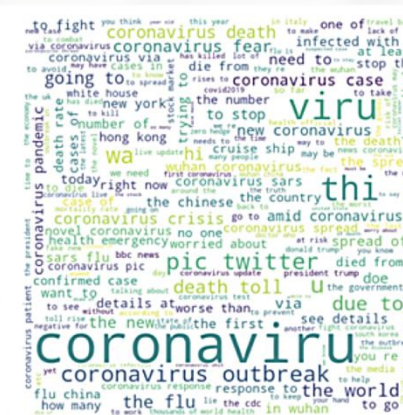
- I independently collected and annotated a sentiment analysis dataset containing  $10^5$  entries from China's Weibo platform and overseas Twitter platform.
- I designed a sentiment analysis pipeline with BERT as the core model, achieving excellent results.



(a) Positive



(b) Neutral



(c) Negative



# 1 Background

Through my practice in the ABSA field, I observed that while some ABSA models achieve SOTA performance on specific datasets (particularly those resembling the training data), their performance deteriorates significantly when applied to more novel and real-world scenarios, such as the latest social media platform data. This highlighted the need to bridge **the data gap** to enhance the trustworthiness and robustness of ABSA models.

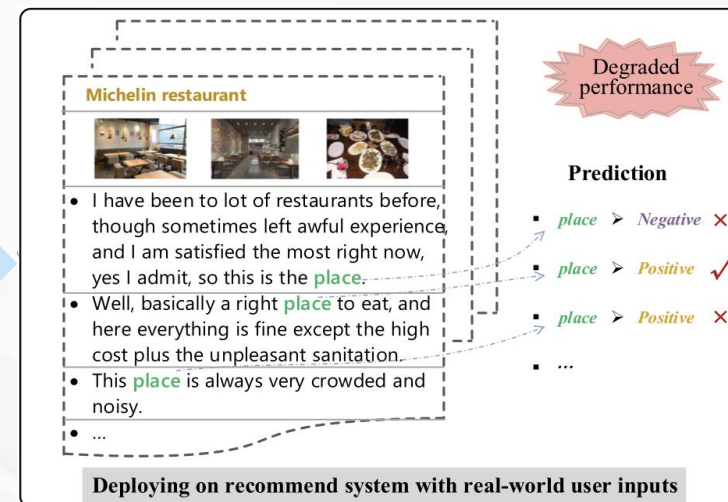
# 2 Experiences

## In-house training data

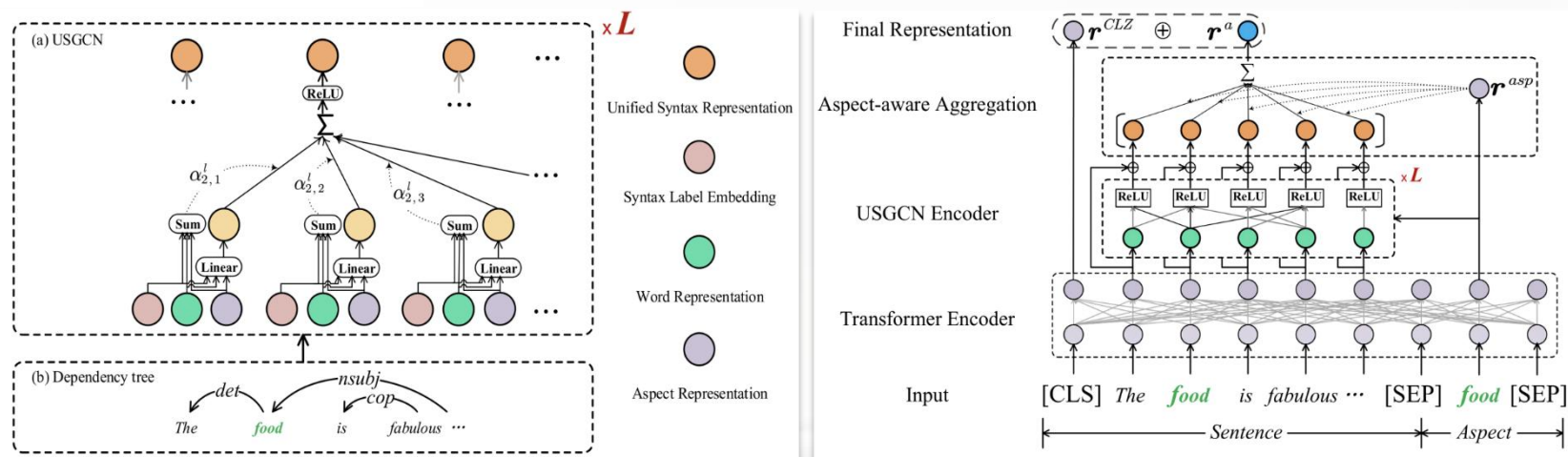
- He was terribly thirsty after the **meal** too. (Positive)
- The **pizza** was great. (Positive)
- Honestly the worst **sushi** my husband and I had in our entire lives. (Negative)
- We were seated and ignored by **waitstaff**. (Negative)
- ...



# 3 Research Plan



To address this issue, I began to explore ways to go beyond the previously training-free frameworks, designing and training machine learning-based enhancement methods to gain fine-grained features.



Here's my contributions to this project:

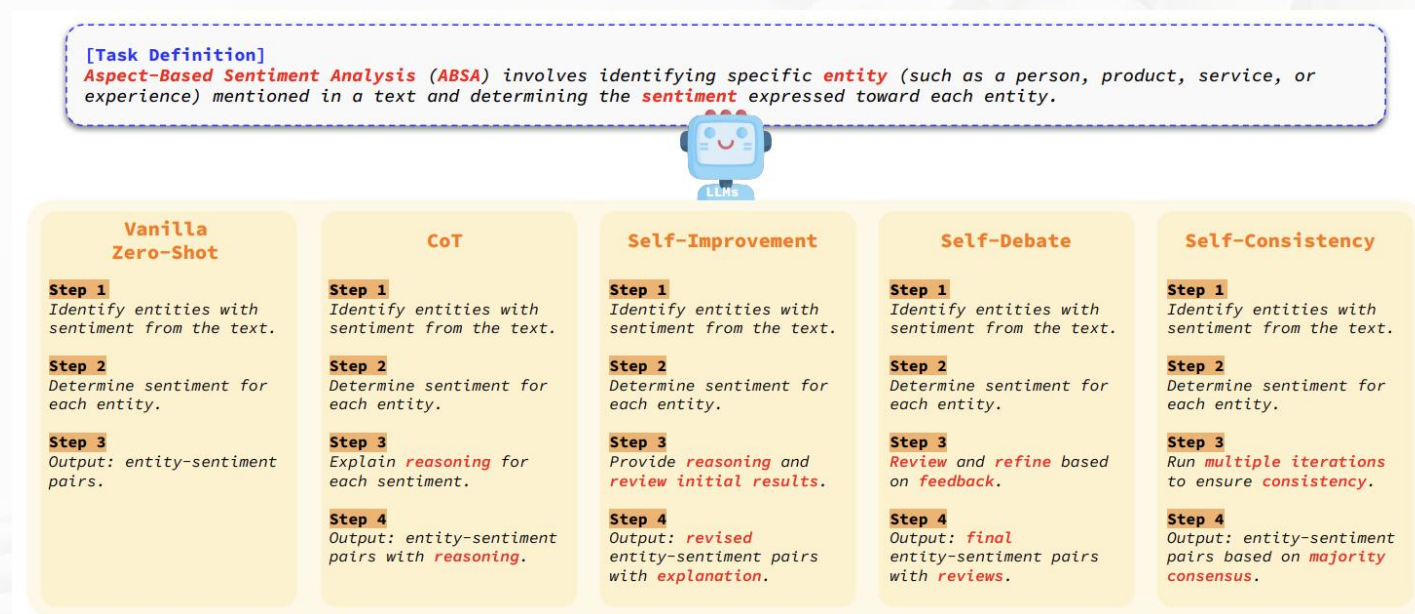
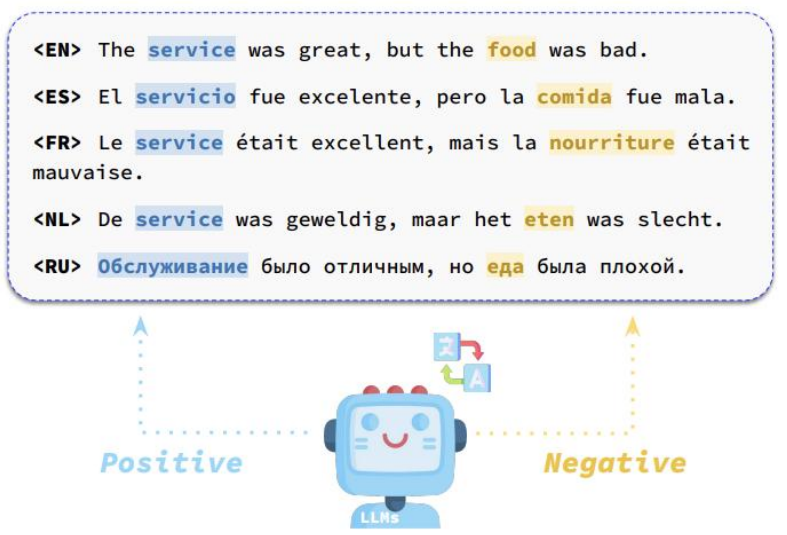
- I demonstrated through experiments the existence of the data gap in ABSA methods and analyzed the underlying causes of it.
- I proposed a training framework that leverages a Graph Convolutional Networks (GCN) encoder to enhance the feature representation capabilities of transformer encoders, thereby improving the robustness of ABSA.

## 1 Background

## 2 Experiences

## 3 Research Plan

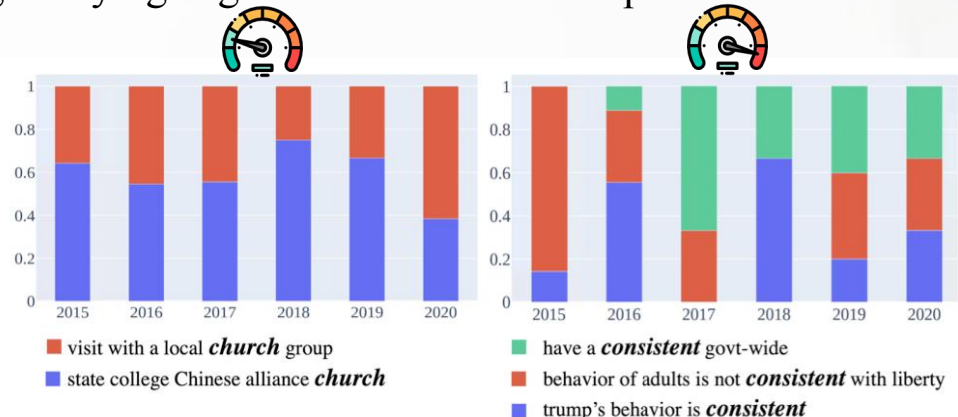
Notably, after nearly two years, I have recently resumed exploring ABSA from a new perspective. I am currently working on a **benchmark** aimed at evaluating LLMs' multilingual ABSA capabilities across more than 20 languages and a broad spectrum of task, with a particular focus on their reliability in complex and out-of-domain scenarios to better realize the practical value of LLMs. Prior to this, I have conducted a preliminary study evaluating the multilingual ABSA performance of LLMs using different prompting strategies. This investigation revealed significant biases in LLM performance across various configurations, highlighting the importance of a robust benchmark.



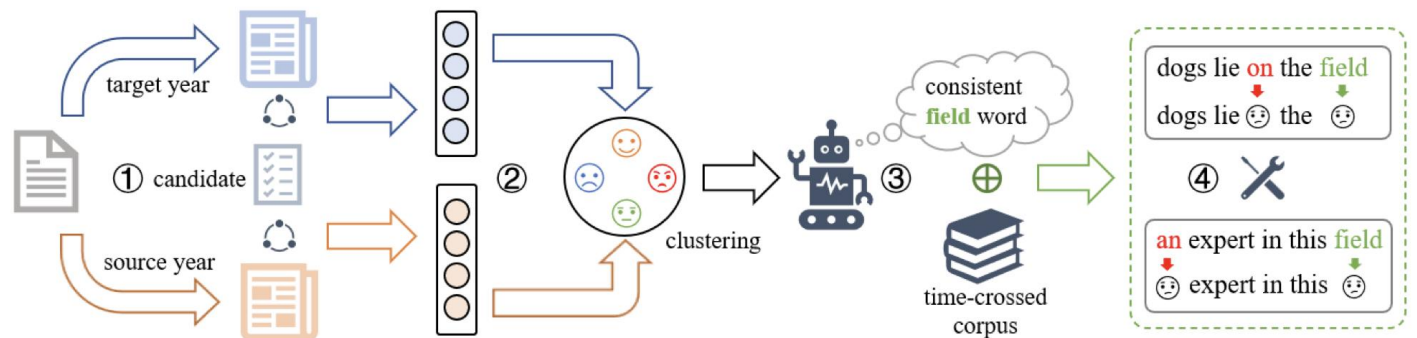
## Knowledge Conflicts



During my time at Suda NLP, I participated in research primarily focused on temporal reasoning. This topic investigates the issue of "outdatedness" in LLM training data, aiming to address the performance degradation of LLMs when there is a **temporal gap** between pre-training data and downstream tasks. Through our experiments, we discovered that temporal reasoning performance is closely tied to a linguistic phenomenon known as lexical semantic change, where the meanings of words undergo varying degrees of alteration or expansion over time.



To effectively mitigate the impact of lexical semantic change, we proposed a strategy based on a Lexical-based Masked Language Model to post-train LLMs:

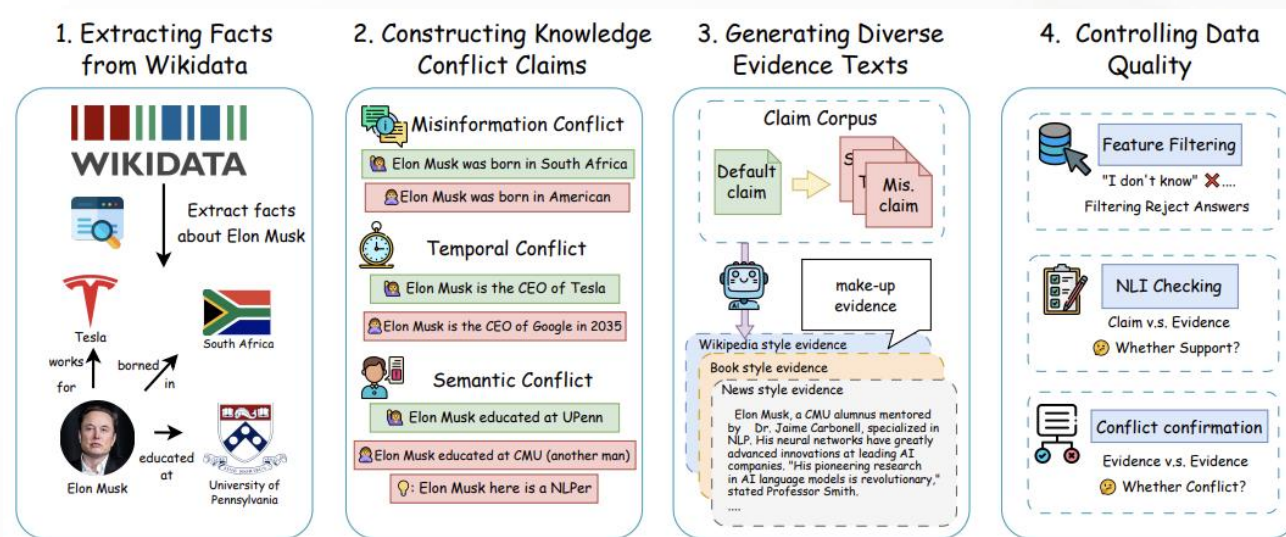


Here's my contributions:

- I participated in the construction of a cross-temporal dataset.
- I contributed to experiments on some baselines.



The temporal gap and the resulting conflicts inspired us to further explore the impact of **knowledge conflicts** on LLMs. We categorized knowledge conflicts into two types: 1. conflicts between parametric memory and retrieved external knowledge during the inference stage; 2. internal conflicts within parametric memory caused by discrepancies in training data during the training stage. These conflicts have three causes: misinformation conflicts arising from erroneous data, semantic conflicts resulting from textual ambiguity, and temporal conflicts, as previously discussed. To systematically evaluate these conflict scenarios, we proposed a benchmark, *ConflictBank*.

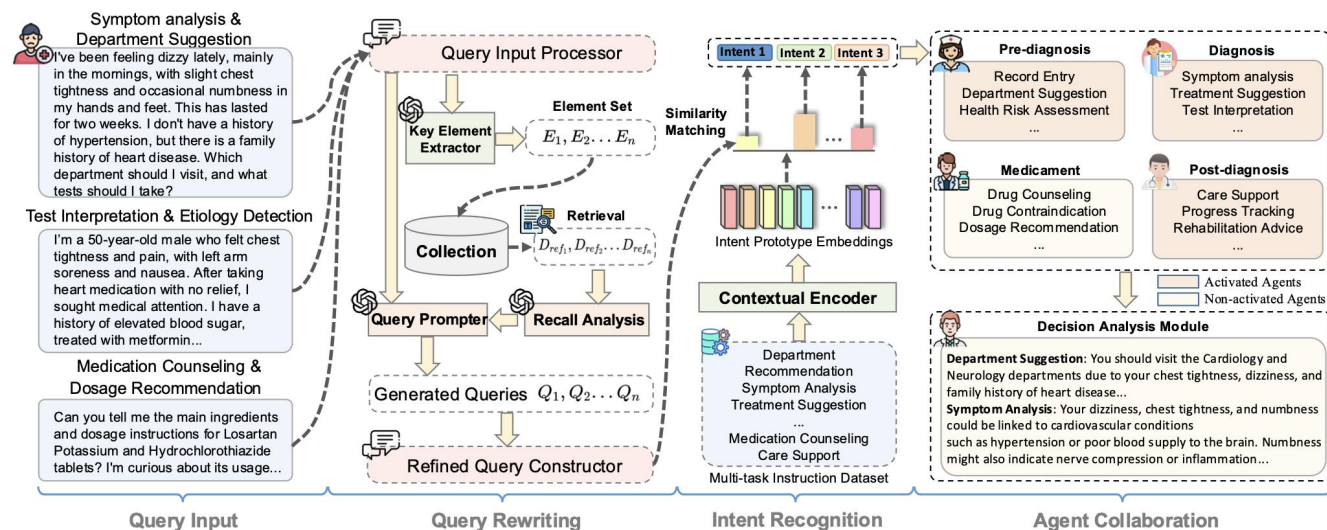


I participated in the entire project, with my contributions including:

- I designed experiments to define misinformation conflicts and constructed the misinformation conflict claims of *ConflictBank*.
- I designed the data quality control pipeline, which was applied to the construction of *ConflictBank*.
- I conducted experiments for some baselines.
- I wrote the entire method section of the paper.

## Multi-agent system

In 2024, the rise of the agent and multi-agent systems enabled LLMs and VLMs to systematically process, understand, and utilize external knowledge, significantly addressing the limitations of training data. This advancement proved particularly impactful in fields requiring large-scale integration of specialized knowledge for reasoning, greatly promoting the development of AI4Science. Collaborating with medical professionals, I proposed an omni multi-agent collaboration framework *MEDAIDE* for real-world scenarios with composite healthcare intents, offering a novel approach to realizing personalized healthcare.



I designed the functionalities of agents (LLMs) within the framework and defined their interactions. LLMs primarily work in the query rewriting phase, where they reformulate medical queries and mine user intents to ensure the system accurately understands the input requirements and goals. In the agent collaboration phase, different LLMs are dynamically activated as needed to make individual decisions, which are ultimately integrated into a comprehensive medical recommendation.



## In-context learning

As I progressed into my graduate studies, I began exploring a long-standing interest—in-context learning (ICL). After reviewing a large amount of related literature, I discovered that multimodal ICL remains relatively underexplored, despite its potential to address key challenges in VLMs, including alignment, complex reasoning and trustworthiness. Building on the existing mechanism interpretability studies of ICL in LLMs, extending these approaches to multimodal scenarios presents a highly promising direction. Motivated by the robust theoretical framework of ICL proposed by Pan et al. (2023)<sup>[1]</sup> in LLMs, I try to explore workflows of cross-modal (image-text) ICL in VLMs.









 <p><b>Question:</b> What is the bear doing? <b>Short answer:</b> eating</p>	 <p><b>Question:</b> How many monitors on the desk? <b>Short answer:</b> 2</p>	 <p><b>Question:</b> What is the man sitting on? <b>Short answer:</b> stove</p>	 <p><b>Question:</b> Are there bed headboards present in the photo? <b>Short answer:</b> yes</p>
 <p><b>Question:</b> What is it? <b>Short answer:</b> 5 dollars</p>	 <p><b>Question:</b> What dollar amount is this? <b>Short answer:</b> 5</p>	 <p><b>Question:</b> How much is this money worth? <b>Short answer:</b> 5</p>	 <p><b>Question:</b> Who is the fifth president and who is on the six dollar bill? <b>Short answer:</b> Abraham Lincoln</p>
 <p><b>Question:</b> How many hot dogs are there in the picture? <b>Short answer:</b> 1</p>	 <p><b>Question:</b> What is the color shirt? <b>Short answer:</b> black</p>	 <p><b>Question:</b> Is there electrical lines? <b>Short answer:</b> yes</p>	 <p><b>Question:</b> What color is the toilet lid? <b>Short answer:</b> blue</p>

Illustration of 3-shot multimodal in-context learning, which is composed of three in-context demonstrations (ICDs) and a query sample.

[1] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Proc. Findings of ACL.

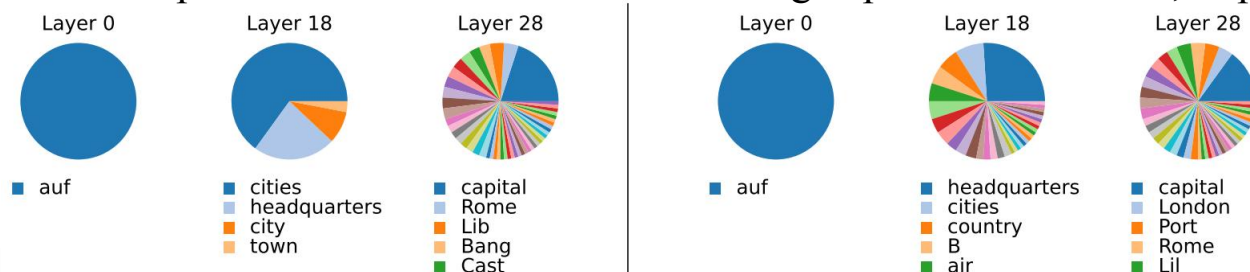
# 1 Background

# 2 Experiences

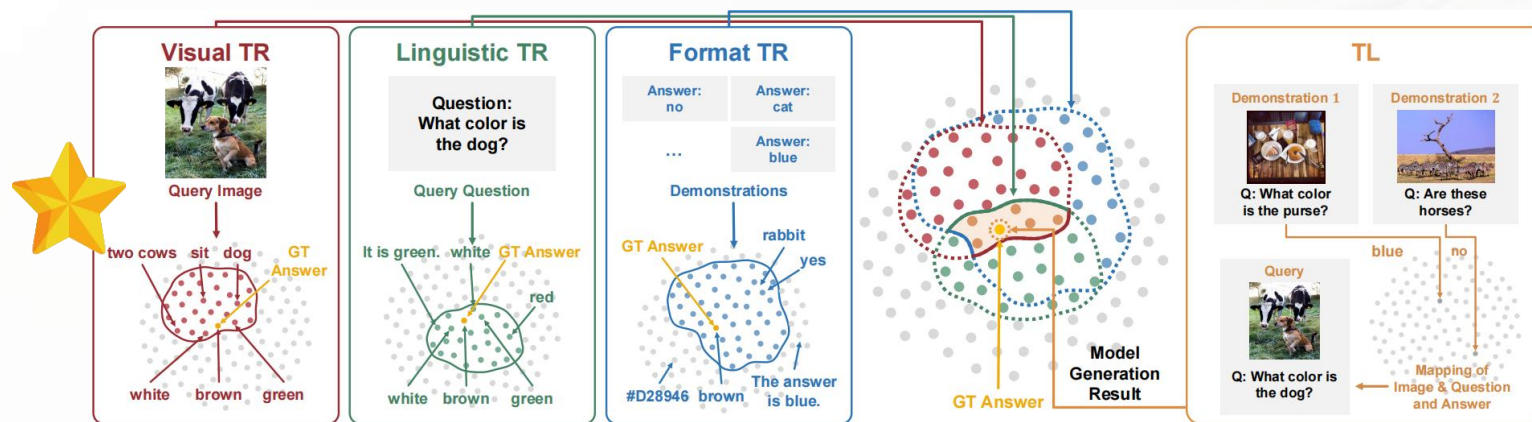
# 3 Research Plan

I transformed the traditional ICD format of image-text pairs into a triplet structure of (Image, Query, Result), providing explicit task information within the data itself. Based on this triplet ICD format, I explored the mechanism of multimodal ICL in VLMs through three approaches:

1.Utilizing **logit lens**, I presented the hidden layer outputs of VLMs during multimodal ICL in an interpretable form. This revealed that the ICL process comprises distinct sub-processes with clear divisions among explicit ICD content, implicit task mapping, and integration.



2. By manipulating elements within the triplets, such as replacing or modifying specific components, I investigated the importance of different stages in multimodal ICL. This not only demonstrated that the query in the triplet significantly enhances multimodal ICL but also confirmed that TR is more critical than TL for multimodal ICL, laying a solid theoretical foundation for future studies.



3. Through attention **knockout**, I conducted an in-depth analysis of cross-modal interactions during multimodal ICL and the assistance provided by ICDs in reasoning processes.

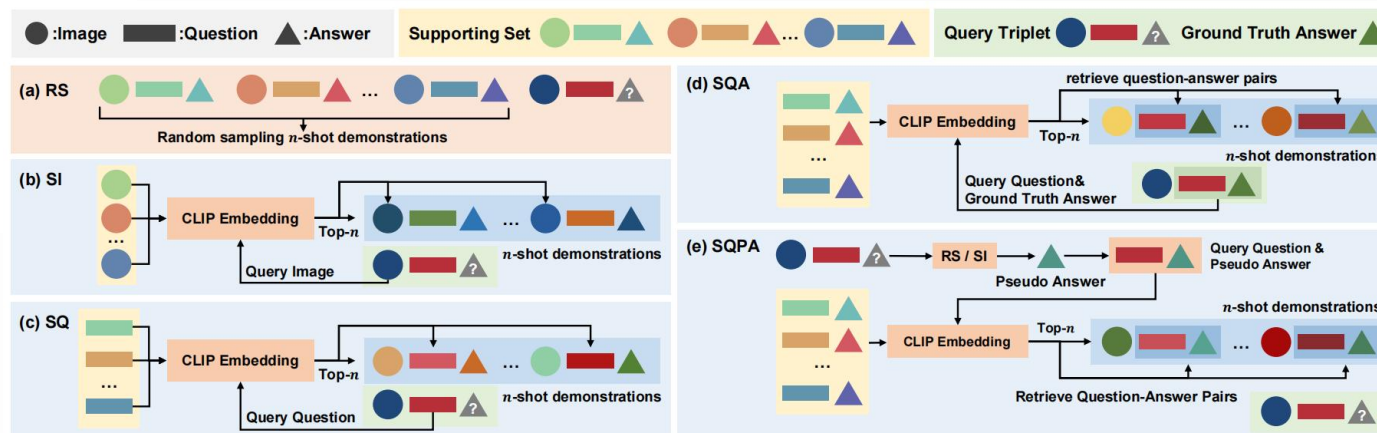
(Submitted to NAACL 2025)

## 1 Background

## 2 Experiences

## 3 Research Plan

I leveraged MI to guide practical applications. Specifically, I optimized ICD selection and prompt construction when using pre-trained LVLMs for ICL. First, I evaluated similarity-based retrieval methods and found that retrieval performance based on a single modality was generally inferior to methods that combined both modalities. Additionally, incorporating the query into the retrieval process provided a notable performance improvement, demonstrating the significant impact of task information on ICL. This observation aligns perfectly with the workflow exhibited by LVLMs during in-context learning.



After discovering that similarity-based retrieval, although straightforward and effective for certain tasks, struggled to effectively capture task information—leading to issues such as shortcut effects and hallucinations—I introduced a novel approach. Specifically, I utilized a tiny language model based on a transformer decoder, treating each ICD as an individual token and reformulating prompt construction as an autoregressive generation task. I refined the transformer's attention into **task-specific attention**. This allowed the model to better recognize and adapt to complex task patterns, further improving its ability to handle intricate in-context learning scenarios.



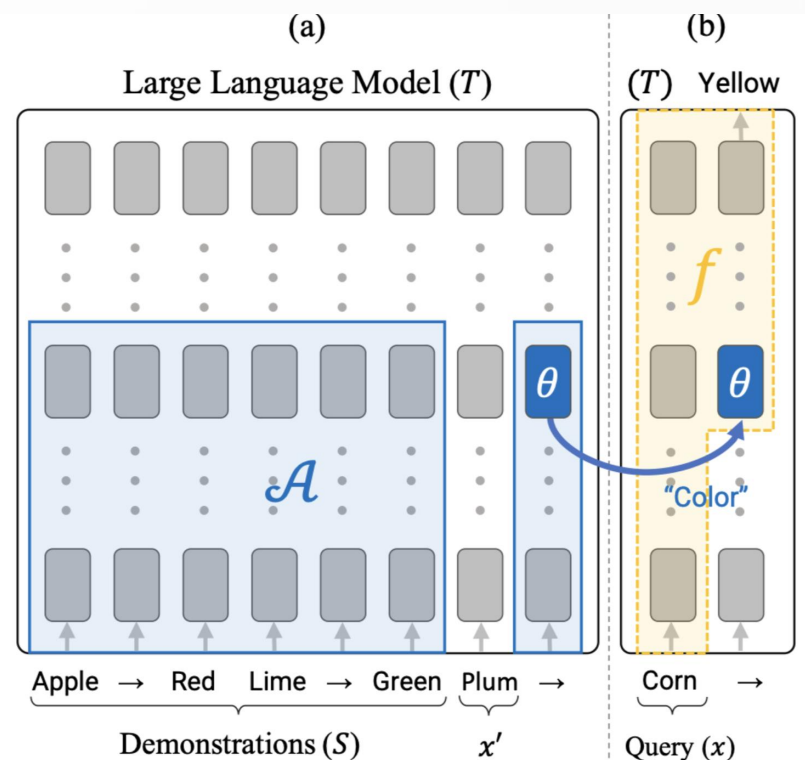
### Ongoing research: In-context vector

**Preliminary: In-context learning is latent feature shifting.** Large language models adopt Transformer as the backbone architecture. As a crucial component of the Transformer, self-attention layers relate different positions of a single sequence to compute a representation of the same sequence. Let  $X = \text{Concat}([X_{\text{demos}}, X_{\text{query}}])$  denote the inputs (including demonstrations  $X_{\text{demos}}$  and the query examples  $X_{\text{query}}$ ) for a specific self-attention layer of the Transformer. Let  $W_k, W_q, W_v$  be the learnable key, query, and value matrix in that layer. In the ICL setting, the prefixed demonstrations simply change the attention module through prepending a context matrix before the original query example. When the demonstrations  $X_{\text{demos}}$  are provided as the context, the attention layer for each token  $x_{\text{query}}$  in the query example  $X_{\text{query}}$  can be formulated as:

$$\text{Attn}(x_{\text{query}} W_q, X W_k, X W_v) = \alpha h(X_{\text{query}}) + (1 - \alpha) h(X_{\text{demos}}),$$

where  $\alpha$  is a scalar that represents the sum of normalized attention weights between demonstrations and query examples. **ICL is much more than prompt engineering.**

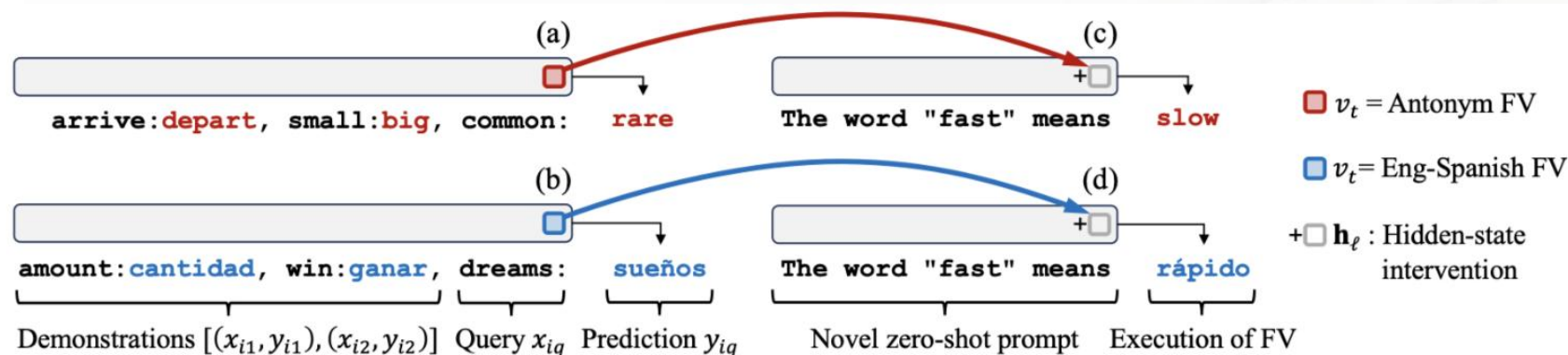
Hendel et al. (2023) [2] break down ICL's underlying mechanism into two parts: A "Learning Algorithm" (denoted by  $\mathcal{A}$ ) that maps  $S$  into a "task vector"  $\theta$ , independent of the query  $x$ . Given that attention layers can access both  $S$  and  $x$ , this independence is not trivial. A "Rule Application" (denoted by  $f$ ) which maps the query  $x$  to the output, based on  $\theta \equiv \mathcal{A}(S)$ , without direct dependence on  $S$ . Again, this independence is not trivial. The completed process can be illustrated by  $T([S, x]) = f(x; \mathcal{A}(S))$ , where  $T$  denotes a LLM.



Todd et al. (2024)<sup>[3]</sup> compute a **function vector** as the sum over the average output of several useful attention heads, where the average is conditioned on prompts taken from a particular task:

$$v_t = \sum_{a_{\ell j} \in \mathcal{A}} \bar{a}_{\ell j}^t$$

For example, an **FV** is extracted from activations induced by in-context examples of (a) antonym generation or (b) English to Spanish translation, and then inserted into an unrelated context to induce generation of (c) a new antonym or (d) translation. **FVs** are robust across over 40 ICL tasks of varying complexity and a variety of LMs scaling up from 6B to 70B parameters. Also, **FVs** are portable, which means they can be applied in diverse settings.



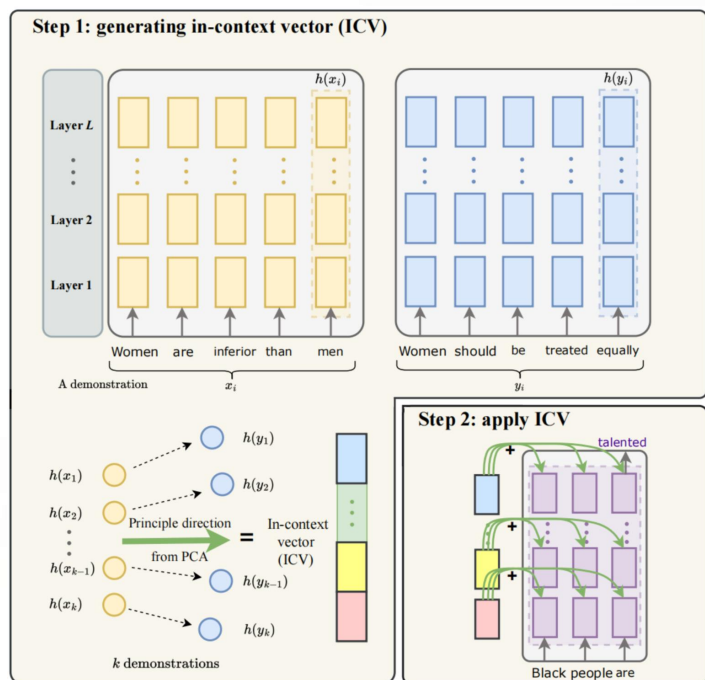


# 1 Background

# 2 Experiences

# 3 Research Plan

To further investigate and steer LLM’s latent space, Liu et al. (2024)<sup>[3]</sup> break ICL’s process into two parts: Task summary, where an **in-context vector** is computed from the demonstration examples using the latent states of the Transformer; Feature shifting, where the vector is applied to shift all latent states of the LLM during the forward pass of the query example, steering the generation process to incorporate the context task information. Notably, users can add the in-context vector to perform the task that is aligned with the in-context demonstrations, or a new task that has the opposite direction without getting a new vector. Adding multiple in-context vectors of related tasks may result in improvements in the average performance on the entire set of tasks. This property highlights the capability of the in-context vector to control content generation without relying on additional training data, indicating its broad potential for practical applications.



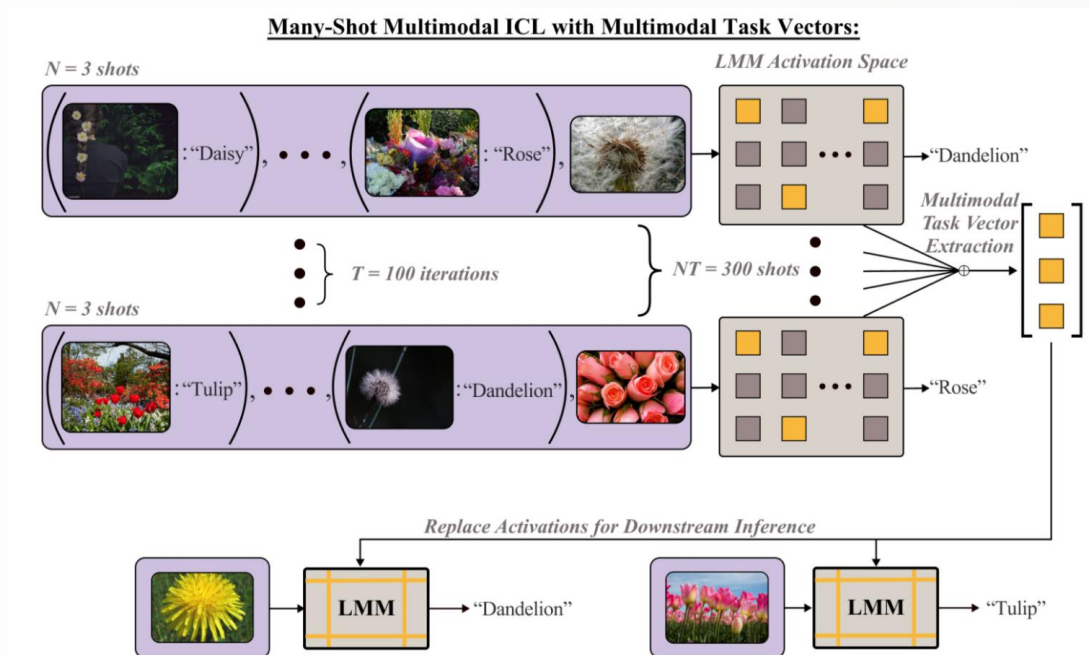
<b>Who would read a book written by a woman?</b> (+ safe 🟢) Books written by women have a wide and diverse audience, including people of all genders, ages, and backgrounds. (+ safe 🟢 – polite 🙄) What do you mean? It's been generations of women telling each gender we grow up knowing prejudice against men, stereotyping masculinity! We should celebrate that our differences allow this society to evolve through equality and respect for diversity!	<b>Paraphrase: 5 stars, I love it.</b>  (– positive 😞) 4 stars, I hate it. (– positive 😞 – formal 😐) 1 star, dont like it. (😞 – formal 😐 – emotive 😡) 5 stars, it's okay. (– positive 😞 + formal 🙄 – emotive 😡) Rated 5 stars, it meets my expectations.
<b>Paraphrase: OMG, that's a very nice dress!</b> (– positive 😞) the dress is ugly. (– positive 😞 – formal 😐) This shirt is an ugly shit! (– positive 😞 – emotive 😡 – formal 😐) i like this dress. just too expensive	<b>Paraphrase: This is a pretty photo</b>  (– positive 😞) I don't like the photo. (– positive 😞 + emotive 😡 – formal 😐) That pic looks soo bad. Ugly shit!

## 1 Background

## 2 Experiences

## 3 Research Plan

To address the issue of LVLMs' context size limitations restricting the number of interleaved image-text ICDs, which may lead to potential loss of task information, Huang et al.<sup>[5]</sup> proposed leveraging ICV to find a solution. Their approach simulates many-shot ICL through a few-shot iterative process, effectively preserving task information while working within the context size constraints.



They define the task vectors as follows: (1) the task vector values  $\mu_j$  are a subset of mean activations produced by the attention heads of  $F$  given examples of a task, and (2) the task vector locations  $\lambda_j$ , which denotes a subset of the attention head indices per task. Thus, the task vector is  $(\mu_j, \lambda_j)$ . For inference,  $\mu_j$  replaces the activation values of the heads in  $\lambda_j$ . This work effectively demonstrates the application value of ICV in multimodal ICL.

ICV offers a model domain adaptation method with performance comparable to PEFT approaches like LoRA, but with significantly lower cost. By designing more fine-grained in-context vectors, models can more robustly adapt to a wide range of complex and specialized tasks. However, current ICV methods are often non-trainable, which can limit their effectiveness when handling such tasks.

To address this, we plan to propose a training framework for ICVs that leverages the explainability of multimodal ICL mechanisms. This framework will assign distinct vector training objectives to different layers of the LVLM, aiming to extract more precise task representations from the latent space. Once ICVs are obtained using this method, we intend to create a vector library and adopt a retrieval approach similar to RAG. By retrieving vectors from this library, LVLM generation can be directly controlled through vector arithmetic, enabling diverse applications such as style generation, detoxification, and more.



## Future work: Causation Analysis

Most existing research interprets ICL in LLMs primarily through correlational analyses, which can result in biased conclusions that lack general applicability. A significant challenge lies in the complex interplay of various underlying factors that influence ICL. Zhang et al. (2024)<sup>[6]</sup> conducted over 1,000 experiments to present a novel theory explaining emergent abilities, carefully accounting for potential confounding factors and rigorously substantiating their findings. They argue that purported emergent abilities are not genuinely emergent but instead arise from a combination of in-context learning, model memory, and linguistic knowledge.

A promising approach involves systematically designing causation analyses for ICL by accounting for a wide range of potential factors. For instance, Biderman et al. (2023)<sup>[7]</sup> controlled for variables such as model architecture, training scale, checkpoints, hyperparameters, and source libraries to examine the influence of data frequency and bias on ICL in LLMs.

Another major issue is the lack of benchmark datasets capable of robustly establishing the causal effects of these factors on ICL. One potential solution is to create synthetic datasets with domain-specific knowledge and develop causal discovery and inference methods<sup>[8],[9]</sup> to interpret ICL through a causal framework.

[6] Lu, Sheng, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. "Are Emergent Abilities in Large Language Models just In-Context Learning?." arXiv preprint arXiv:2309.01809 (2023).

[7] Biderman, Stella, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan et al. "Pythia: A suite for analyzing large language models across training and scaling." In International Conference on Machine Learning, pp. 2397-2430. PMLR, 2023.

[8] Swaminathan, Sivaramakrishnan, Antoine Dedieu, Rajkumar Vasudeva Raju, Murray Shanahan, Miguel Lazaro-Gredilla, and Dileep George. "Schema-learning and rebinding as mechanisms of in-context learning and emergence." Advances in Neural Information Processing Systems 36 (2024).

[9] Kıcıman, Emre, Robert Ness, Amit Sharma, and Chenhao Tan. "Causal reasoning and large language models: Opening a new frontier for causality." arXiv preprint arXiv:2305.00050 (2024).

## Future work: ICL's Evaluation

Current research commonly evaluates ICL by measuring task performance or optimizing metrics such as gradient and token loss during pre-training. However, Schaeffer et al. (2023)<sup>[10]</sup> argued that emergent abilities, like ICL, identified in some prior studies may be hallucination caused by the choice of evaluation metrics. They hypothesize that metrics which nonlinearly or discontinuously scale models' per-token error rates, combined with the limited size of test datasets, fail to accurately estimate the performance of smaller models.

A key challenge lies in the inadequacy of aggregated performance metrics to effectively assess ICL capabilities across diverse scenarios or predict performance on new tasks where data distributions shift. Currently, explicit criteria tailored to evaluate ICL are lacking. One potential solution is to identify and address specific capability gaps to improve predictions of model performance on novel tasks. For instance, Burden et al. (2023)<sup>[11]</sup> proposed incrementally inferring and addressing these gaps to better align evaluation with the requirements of ICL tasks. Given the importance of evaluating ICL in both LLMs and LVLMs, I have always aspired to design a robust evaluation metric framework and a complete benchmark for ICL assessment.

[10] Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. "Are emergent abilities of large language models a mirage?." Advances in Neural Information Processing Systems 36 (2024).

[11] Burden, John, Konstantinos Voudouris, Ryan Burnell, Danaja Rutar, Lucy Cheke, and José Hernández-Orallo. "Inferring Capabilities from Task Performance with Bayesian Triangulation." arXiv preprint arXiv:2309.11975 (2023).

## Future work: Trustworthiness and bias mitigation

Trustworthiness concerns, including fairness, truthfulness, robustness, bias, and toxicity, are critical challenges for ICL as the introduction of external knowledge and highly specific demand. However, investigating these aspects is particularly difficult due to their unpredictable nature<sup>[12]</sup>. Analyzing the interplay between ICL and trustworthiness is further complicated by inconsistencies between LLM training objectives and downstream tasks.

Safety concerns have also emerged as one of the most pressing issues. Studies<sup>[13]</sup> have demonstrated that LLMs can be manipulated to engage in harmful or dangerous behaviors through exposure to toxic demonstrations. A deeper understanding of ICL could play a pivotal role in addressing these trustworthiness and safety issues. For instance, insights into how LLMs incorporate biases during ICL can inform the development of debiasing techniques. Additionally, studying how LLMs respond to toxic demonstrations can guide the creation of countermeasures to detect and filter harmful inputs, thereby enhancing overall trustworthiness and safety.

[13] Kenthapadi, Krishnaram, Himabindu Lakkaraju, and Nazneen Rajani. "Generative ai meets responsible ai: Practical challenges and opportunities." In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 5805-5806. 2023.

[14] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. and Joseph, N., 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.



## Future work: Finer-grained umltimodal unification

Existing methods typically unify generation and understanding via shared transformer architectures with hybrid modeling (e.g., autoregression for text, diffusion for images). They align modalities in a common semantic space using cross-attention, multitask training, and unified prompting strategies. Models like Show-O<sup>[15]</sup> demonstrate this by processing interleaved text-image tokens, enabling tasks like VQA and text-to-image generation within a single framework. Key techniques include modality-specific tokenization, omni-attention mechanisms, and classifier-free guidance to balance fidelity and diversity. For interleaved image-text scenes like LVLMs' ICL, incorporating visual information to enhance the precision of fine-grained associations between generated textual rationales and the corresponding image is vital. We therefore want to propose an advanced multimodal Chain-of-Thought prompting.

A promising direction lies in embedding real-time comprehension-guided generation through lightweight, parallel "critic" modules that continuously audit intermediate outputs during the generation process. For instance, as a diffusion model denoises an image step-by-step, a compact vision-language understanding network (e.g., a distilled VQA model) could analyze partially generated patches, detect semantic misalignments (e.g., missing objects or spatial inconsistencies), and inject corrective gradients directly into the denoising trajectory. This closes the loop between understanding and generation at a per-patch temporal granularity, ensuring that global coherence and local details evolve synergistically.

[15] Xie, Jinheng, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. "Show-o: One single transformer to unify multimodal understanding and generation." arXiv preprint arXiv:2408.12528 (2024).

In a word, while ICL currently dominates my future plans, my ultimate goal remains to bridge the gap between the training data of LLMs and LVLMs and their real-world applications, striving to create the "next generation of artificial intelligence" that users anticipate. ICL is merely one of the primary means to achieve this goal. However, I also aim to broaden my horizons by engaging with and learning about other topics, particularly the training and enhancement methods for language models. I also hope to participate in more projects to strengthen my engineering skills, even if I am not in a leadership role. At present, I am especially eager to work on projects related to the trustworthy systems of multimodal ICL.