Interpreting In-Context Learning for Semantics-Statistics Disentanglement via Out-of-Distribution Benchmark

Anonymous ACL submission

Abstract

The rapid growth of Large Language Models (LLMs) and Vison-and-Language Models (VLMs) has highlighted the importance of interpreting their inner workings. Arguably, the biggest question in interpretability is *why* an LLM can solve a number of tasks or whether they obtain the semantics other than the statistical co-occurrence (Semantics-Statistics disentanglement, or S^2 disentanglement). Although previous works disentangled the several semantic aspects, uniform interpretation poses two challenges; First, previous works are only weakly tied to *how* an LLM works; In-Context Learning (ICL). Second, most problems are In-Distribution (ID), where the semantics and statistics (e.g., a prompt format) are inseparable. Here we propose the Representational Shift Theory (RST), stating that an ICL example causes the cascading shift in the representation for the S^2 disentanglement. To benchmark RST, we formalize the Out-of-Distribution (OoD) generalization under RST and propose two hypotheses for the ICL performance of VLMs *not* trained with multi-image or multi-turn resources (OoD ICL). Our first hypothesis is that OoD ICL can contribute to the performance when the ID performance is poor. Our second hypothesis is that the counterfactual textual ICL example works better than the first approach when the textual modality is predominant. We obtained the supporting evidence in six visual question-answering datasets for the first hypothesis and in a hateful memes challenge dataset for the second hypothesis. In conclusion, our work marks a crucial step towards understanding the role of ICL over the S^2 disentanglement, a central question of interpretability.

1 Introduction

Upon the explosive usage of the Large Language Model (LLM) in Natural Language Processing (NLP; Zhao et al. (2023)), interpreting its inner workings is critical for reliable and evidence-based decision-making. Arguably, the most fundamental interpretability question is *why* an LLM works; i.e., whether an LLM acquires the semantics (Abdou et al., 2021; Gurnee and Tegmark, 2024; Godey, 2024; Vafa et al., 2024) or is a *parrot* repeating statistically plausible responses (Zečević et al., 2023; Bender et al., 2021). Previous works tackle this Semantics-Statistics disentanglement (S^2 disentanglement) for various aspects (e.g., color or geolocation) from an LLM's latent space. Building a unified framework for S^2 disentanglement in general, however, is still outrageous.

To build a unified interpretability framework for LLMs, In-Context Learning (ICL; Brown et al. (2020)), a gradient-free reasoning capability emerging in LLMs, is critical. A major finding in interpretability for ICL is the concept of *meta-gradient* (von Oswald et al., 2023; Dai et al., 2023a); LLMs can learn to optimize its own latent space in the absence of the gradient information. Despite the rich literature on theoretical and empirical justification, the relevance of the meta-gradient to S^2 disentanglement is elusive; i.e., *why* that interpretation is valid is still unclear. Here we propose Representational Shift Theory (RST) for interpreting how an ICL example affects the latent space, leading to S^2 disentanglement.

To study S^2 disentanglement, the Out-of-Distribution (OoD) generalization (Farquhar and Gal, 2022) provides valuable insights. OoD is a distinction of the data distribution between the static training set and the diverse test set. An LLM required to generalize to OoD input performs the *explicit* S^2 disentanglement; infer the *same semantics* facing the *different distribution (i.e., statistics)*. Therefore, we tackle the OoD generalization with RST to show its effectiveness on S^2 disentanglement.

More specifically, we focus on OoD generalization in the vision-and-language (VL) problems due

to the growing needs in real-world applications. Due to the resource shortage with the multi-image multi-turn conversations, many VL models such as LLaVA (Liu et al., 2023b) are solely trained with single-image single-turn resources. This means that ICL is an OoD generalization (OoD ICL) to these models, making it ineffective. Improving OoD ICL reduces the need for labor-intensive data collection and resource-consuming training. Using RST as a guiding principle, we address this challenging problem.

Our contribution could be summarized as follows:

1. As an extension of the meta-gradient, we propose RST to describe how an ICL example affects the LLM output. RST states that an ICL example first shifts the representation of the zero-shot input, and this shift triggers another shift of the output. We introduce a semantic term and a statistic term in RST as the first formalism of S^2 disentanglement in general. We further show how OoD ICL can be framed into the S^2 disentanglement. In short, we formalize OoD ICL as the amplification of the semantic term under the fixed statistic term[1].

2. We hypothesize that adding an OoD ICL image-text pair (Multi-image Multi-turn OoD, or MM OoD) could improve the performance when the zero-shot input does not provide strong semantics. We confirm this hypothesis in six diverse Visual Question Answering (VQA) datasets.

3. We also hypothesize that counterfactual prompting for curating the text-only OoD ICL example (Single-image Multi-turn OoD, or SM OoD) contributes to the performance when the original input is biased toward a specific label and the text is dominant over the image. To validate this, we apply counterfactual prompting and instruct the model to curate a negative example before the decision-making. We observe its effectiveness in a hateful meme challenge dataset.

2 Related Work

First, we review previous work on Semantics-Statistics Disentanglement (S^2 Disentanglement), a central question in this study. Second, we summarize the impact of In-Context Learning (ICL) and the interpretability studies focusing on ICL to understand its significant role on S^2 Disentanglement. Finally, we introduce the previous Out-of-Distribution (OoD) benchmarks and efforts to position ourselves in OoD studies.

2.1 Towards S^2 Disentanglement

In parallel to the wide application of LLMs to NLP (Zhao et al., 2023) and the relevant multimodal fields (Zhang et al., 2024), centric to the interpretability is S^2 Disentanglement. Typically, a single work focuses on one or a few aspects of semantics. For example, Abdou et al. (2021) extracted the subjective aspects of color disentangled from the light spectrum in LLMs' representations. Gurnee and Tegmark (2024) showed the robustness of the representation of the geolocation and time, and Godey (2024) analyzed this geography under the scaling law (Kaplan et al., 2020). Vafa et al. (2024) analyzed the world model in LLM for spatial information. We aim at a theory spanning multiple aspects of semantics.

2.2 ICL

After the initial introduction by Brown et al. (2020), massive efforts have been spent on improving the LLMs' ICL capabilities, which we categorize into three groups. The first group focuses on task instruction, such as Chain-of-Thought reasoning (Madaan et al., 2023). The second group optimizes the ICL example(s) choice, typically from the training data. Since this process is cost-consuming given the large volume of data, most studies adopt a simple algorithm such as BM25 (Robertson et al., 1996). Another type of selection method utilizes models with strength in semantics-oriented tasks (e.g., image aesthetics[2]), such as CLIP (Radford et al., 2021). The last group curates the ICL examples, mostly by LLMs. A subgroup of example curation with a strong theoretical backbone is counterfactual prompting (Wang et al., 2024). Based on the given task's data generation process, this approach generates examples with desired properties, such as the least modification of the original example for label flipping. To validate our theory, we use a standard set of methods for the experiments. Specifically, we use CLIP-based image-text pair selection for Experiment I. For Experiment II, we

[1] such as the effect of a two-dimensional image tensor in OoD, whereas the model is solely trained with the tensor with a single dimension

[2] https://laion.ai/blog/laion-aesthetics/

use counterfactual prompting as the main methodology and BM25-based text-guided ICL example selection as a text-oriented baseline.

Interpreting how ICL works is another hot topic. Various interpretations have been proposed to obtain theoretical and empirical grounding behind ICL. Typically, the interpretation studies hire a specific algorithm to interpret the dynamics of LLM's representations: for example, Bayesian inference (Xie et al., 2022), contrastive learning (Ren and Liu, 2023), multi-state RNN (Oren et al., 2024), and gradient descent (von Oswald et al., 2023; Dai et al., 2023a), among many others (Han et al., 2023; Wang et al., 2023; Li et al., 2023). These studies covered extensive theoretical aspects, including the common finding of *meta-gradient*; LLMs could learn how to optimize its own representation. However, how each theory contributes to S^2 disentanglement is unclear. We tackle this problem with an extension of the meta-gradient.

Another line of interpretability studies isolated critical mechanisms or data structures for ICL, such as the induction head (Olsson et al., 2022; Cho et al., 2025), the function vectors (Todd et al., 2024), and the parallel structure (Chen et al., 2024a). Instead of focusing on the detailed mechanisms, our study provides a macroscopic analysis of the entire latent space. The potential connection to this line of work is in the Appendix E.1.3.

2.3 OoD Generalization

An Out-of-Distribution (OoD) problem is defined as a distinction of the distributional shift from the static training dataset to more diverse test inputs (Farquhar and Gal, 2022). OoD generalization is the task where the models need to address the OoD problems (Hendrycks and Gimpel, 2017). Since this topic is diverse, hereafter we limit our scope to NLP and VL domains unless stated otherwise.

Most efforts on these domains have been spent on domain adaptation (Ramponi and Plank, 2020) and label shift (Zhang et al., 2021; Wu et al., 2021). Both approaches hold out some categories X_{test} of the resource(s), and test the performance of the model trained solely with the other categories X_{train}; The former uses multiple datasets of similar topics, and the latter splits the multi-class classification labels. Although these studies provide valuable insights, the distinction between semantics and statistics is elusive; i.e., how to define the distributional difference among multiple datasets or multiple labels is opaque.

In parallel to the efforts on extending the context length (Huang et al., 2024) and the explosive growth of multimodal LLMs centered on VL capabilities (Zhang et al., 2024), several works addressed OoD problems in a single-image conversation and a multi-turn conversation separately. For example, Dai et al. (2023b); Gao et al. (2024) proposed solutions for detecting OoD in a multimodal conversation. Lang et al. (2024) introduced the information-theoretic approach for multi-turn conversation intention detection. Ye et al. (2022) proposed two novel OoD categories, the multi-label OoD and the label shift under the specific context. Here, we extend the application to a multi-image, multi-turn conversation, marking a crucial step toward generalization to the real world.

3 Preliminaries

3.1 Meta-Gradient

Central to the optimization of traditional machine learning is the gradient descent, where the learning objective is explicitly given to the model, forming the gradient ΔH over the representation H of an input in hidden space. A line of works (von Oswald et al., 2023; Dai et al., 2023a) suggests that the LLMs perform another form of gradient descent in ICL. To summarize, they use their own attention weights W to form a meta-gradient ΔW, multiplied by H to form the updated representation H'. In a typical zero-shot setting, the only information composing the meta-gradient is task instruction, so the representation of the instruction H_{inst} is updated by this meta-gradient $\Delta W_{inst/zsl}$ to form the representation of a zero-shot input H_{zsl}. In ICL, the example is inserted between the instruction and the zero-shot input, so the gradient consists of 1) the gradient between the instruction and ICL example $\Delta W_{inst/icl}$ and 2) the gradient between an ICL example and a zero-shot input $\Delta W_{icl/zsl}$, together forming the ICL example's representation H_{icl}. In summary, the meta-gradient in zero-shot and ICL settings are summarized as:

$$\begin{aligned} H_{zsl} &= (W - \Delta W_{inst/zsl})H_{inst} \\ H_{icl} &= \{W - (\Delta W_{inst/icl} + \Delta W_{icl/zsl})H_{inst}\} \end{aligned} \tag{1}$$

Note that most meta-gradient studies use linear variants (e.g., Zhuoran et al. (2021)) of Transformer (Vaswani et al., 2017). In contrast, we assume that the concept is solid for the original model for brevity. We empirically validate this assumption.

3.2 Unembedding

Another important concept in interpretability studies (e.g., nostalgebraist (2020); Belrose et al. (2023)) is that the representation could be linearly projected, or *unembedded*, with a weight W_{emb} to the LLM's output Y.

$$Y = W_{emb}H \tag{2}$$

Combined with the meta-gradient, we propose a novel theory explaining how ICL works.

3.3 Mixed Effect Model

In Section A.2, we assume that the effect of statistics is static over the various inputs, while that of the semantics is diverse. The mixed effect model (Singmann and Kellen, 2019) provides the analytical framework for this dual effect. Specifically, in observation i, the effect of a variable X over the target variable y_i is expected to be identical across all the observations (*fixed effect*), and another variable Z affects individual observation differently (*random effect*). In a multiplicative case (Eq. 1), a mixed effect model could be formalized as:

$$y_i = (W_X + W_{Z_i}Z_i)X \tag{3}$$

For example, when analyzing the effect of a new teaching method on student performance across different schools, the teaching method may have a fixed effect since such a method generally aims for equal educational opportunities. In contrast, a variable representing each school should have a random effect when each school has a different educational policy. Note that various nonlinear expressions of the mixed effect are proposed (e.g., Hajjem et al. (2014); Sigrist (2023)), but we limit the scope to the linear model for brevity.

4 Methodology

In this section, we outline our methodology for exploring how LLMs disentangle semantic content from statistical properties of input data. We define key terms, describe our approach with illustrative examples, and introduce our hypotheses about generalization. An overview of our methodology is depicted in Fig. 1. For illustrative purposes, we use two VQA examples: 1) *Banana Mustache* (Agrawal et al., 2015) example, in which a woman holds two bananas in front of her mouth, resembling the mustache, and 2) *Tomato Nose* example[3], in which the tomato is placed between a man's eyes and his mouth, representing the nose. Note that the second example is not in the actual datasets we used.

[3]reprinted from Freepik.com

4.1 Definitions

- **Semantics**: Meaningful content in the input, such as objects and relationships in images or text. In the Banana Mustache case, a banana placed between a woman's nose and mouth resembles a mustache.

- **Statistics**: Non-semantic properties like the number of images or dialogue turns, which may influence model performance due to learned patterns but do not convey meaning.

- S^2 **Disentanglement**: Extracting and utilizing semantic information regardless of statistical format. Successful disentanglement occurs when a model focuses on semantics to interpret inputs presented in formats different from the training data.

- **Representational Shift**: Changes in the model's internal representations (e.g., hidden states) caused by ICL examples. This shift reflects how ICL examples affect the processing of zero-shot inputs, leading to differences in output and capturing the influence of semantics on the model's reasoning.

- **In-Distribution (ID) and Out-of-Distribution (OoD)**: ID refers to data formatted like the training data (e.g., single-image, single-turn dialogues). OoD refers to data with different formats (e.g., multi-image, multi-turn dialogues), challenging the model's generalization.

- **Out-of-Distribution In-Context Learning (OoD ICL)**: Providing in-context examples in formats not seen during training to assess the model's ability to leverage semantic cues in unfamiliar statistical contexts.

4.2 Representational Shift Theory

Representational Shift Theory (RST) posits that providing ICL examples can affect the internal representations of LLMs, leading to shifts in both input and output representations (Fig. 1 (A)). Specifically, an ICL example influences the representation of a zero-shot input (*input shift*) and subsequently

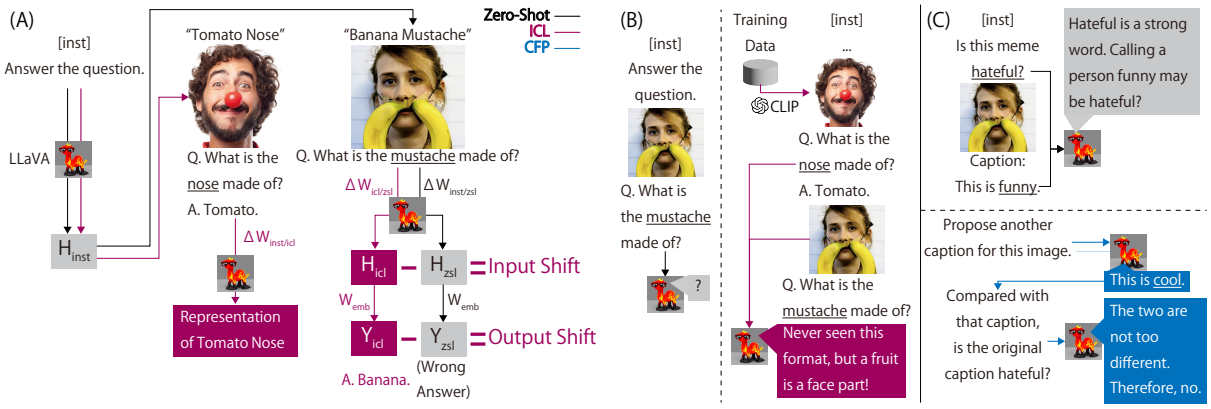


Figure 1: Methodological Overview with *Banana Mustache* test input and *Tomato Nose* ICL example. (A) Representational Shift Theory. Both in a zero-shot setting (black line) and in ICL (red line), an instruction (question) is first provided with LLaVA to compose the representation H_{inst}. Next, in a zero-shot setting, the test input forms the meta-gradient $\Delta W_{inst/zsl}$, resulting in the test-input representation H_{zsl}. Finally, the unembedding weight W_{emb} converts H_{zsl} to the output Y_{zsl}. In contrast, in ICL, an example is inserted between the instruction and the test input to emphasize the semantic components (analogy of fruits and facial parts) to form the meta-gradient $\Delta W_{inst/icl}$ and $\Delta W_{icl/zsl}$, and then the representation H_{icl} and the output Y_{icl}. We argue that the difference of test-input representation $H_{icl} - H_{zsl}$ (input shift) and that of output representation $Y_{icl} - Y_{zsl}$ (output shift) are the core of ICL. (B) Hypothesis I and S^2 disentanglement in Experiment I. We hypothesize that *semantically* rich yet *statistically* unfamiliar ICL example contributes to the performance when the zero-shot performance is poor (left). To validate this hypothesis, we tested LLaVA's OoD ICL (right) performance by adding an ICL example most similar to the test input based on CLIP embedding (*semantically* rich) to the LLaVA variants *not* trained with multi-image datasets. When the model successfully performs S^2 disentanglement, it extracts the *semantic* analogy despite the unseen format (*statistics*). (C) Overview of Hypothesis II and Experiment II. When LLaVA is textually biased towards the hateful label (top), we hypothesize that enhancing text-to-text interaction facilitates the unbiased decision (bottom), and test this hypothesis with counterfactual prompting (instructing the model to propose a caption to compare with the original caption).

affects the model's output (*output shift*). By analyzing these shifts, we can understand how semantic content influences the model's reasoning.

4.2.1 Representational Shift

Our methodology focuses on analyzing how LLMs process and interpret inputs that require semantics-statistics (S^2) disentanglement, particularly in OoD settings (Fig. 1 (A)). We introduce the concept of *Representational Shift* to measure internal changes within the model influenced by semantically rich examples. Consider the Banana Mustache example, where the question is *"What is the mustache made of?"* and the correct answer is *"Banana."*. Without additional context, the model may struggle with this analogy. Introducing a semantically similar Tomato Nose example–a pair of the Tomato Nose image and the question *"What is the nose made of?"* (answer: *"Tomato"*)–provides a similar scenario. This illustrative example demonstrates how providing analogous semantic content can potentially cause a representational shift, enhancing the model's ability to interpret the original input correctly. We measure changes in the latent space–specifically the hidden states[4]–with and without the ICL example. In practice, we analyze the distance between representations using a cosine similarity metric. This approach allows us to observe how semantic content in ICL examples causes representational shifts that improve the model's performance. Mathematical formulations of RST are provided in Appendix A.

4.2.2 S^2 Disentanglement

Our S^2 Disentanglement framework aims to separate the effects of semantic content from statistical patterns in ICL examples. We assume that semantics and statistics are independent, allowing us to analyze their individual contribution to representational shift. By presenting the model with semantically rich examples in formats statistically different from the training data, we evaluate whether the model can extract and apply semantic information despite unfamiliar formats. Success is

[4] https://huggingface.co/docs/transformers/en/main_classes/output#transformers.modeling_outputs.BaseModelOutput

demonstrated when the model, given semantically analogous OoD ICL examples, correctly answers questions it previously struggled with, indicating effective S^2 disentanglement. The formalization of S^2 Disentanglement is detailed in Appendix A.2.

4.3 Hypotheses on Generalization

We propose two hypotheses to improve generalization through representational shift:

1. **Hypothesis I: Multi-image Multi-turn OoD (MM OoD)**: When the zero-shot (In-Distribution, ID) input lacks sufficient semantic information for the LLM (i.e., poor zero-shot performance), providing semantically rich MM OoD ICL examples can help improve performance (Fig. 1 (B)). This scenario is effective when the model needs more contextual semantic cues to make accurate predictions. In Experiment I, supposing a model trained solely with the single-image datasets (e.g., the images and captions of fruits or human faces) struggles with understanding a VQA task like Banana Mustache (Fig. 1 (A), left), we provide a semantically rich OoD ICL example, such as Tomato Nose, to see if it can cause the meaningful representational shift, leading to better performance (Fig. 1 (A), right).

2. **Hypothesis II: Single-image Multi-turn OoD (SM OoD)**: When textual semantics are more informative than image semantics, enhancing the textual content through SM OoD ICL examples can improve performance (Fig. 1 (C)). This approach is beneficial when the model over-relies on statistical patterns or exhibits label bias. For example, the LLMs instruction-tuned not to miss any hateful languages may fail to recognize a banana mustache meme as a neutral analogy (Fig. 1 (C), top). In Experiment II, by asking the model to propose a neutral caption for the Banana Mustache image (e.g., *"This is cool"*), we encourage it to utilize the text-to-text comparison for the hateful meme detection (Fig. 1 (C), bottom).

By testing these hypotheses, we aim to demonstrate that representational shift, facilitated by semantically rich OoD ICL examples, can help LLMs focus on semantic content over statistical patterns.

5 Experiments

We conducted two experiments to test our methodology's effectiveness: Experiment I for MM OoD and Experiment II for SM OoD. We used two variants of LLaVA (Liu et al., 2023b) (LLaVA-Llama2 (Touvron et al., 2023) and LLaVA-1.5 (Liu et al., 2023a)) in both experiments for two reasons: (1) their reported state-of-the-art performance on linguistic tasks indicates high capacity for the semantic term, and (2) they are *not* trained with multi-image resources or ICL settings, allowing OoD analysis. We used models with 13 billion parameters to balance linguistic capability and memory constraints. We also conducted a preliminary experiment with InternVL (Chen et al., 2024b) for analysis in an ID setting (Appendix D.8). We focused on ICL with a single example since we did not observe any significant benefit from concatenating multiple examples in initial explorations.

5.1 Experiment I: Enhancing Performance with Semantically Rich MM OoD ICL Examples

Our objective in Experiment I was to test **Hypothesis I** by investigating whether incorporating semantically rich MM OoD ICL examples can improve model performance on VQA tasks initially hindered by weak semantics.

5.1.1 Experimental Setup

First, we evaluated the zero-shot (ID) and one-shot (MM OoD) performance of LLaVA models on six VQA datasets: VQA v2.0 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), TextVQA (Singh et al., 2019), MMBench (Liu et al., 2023c), and MM-Vet (Yu et al., 2023). The one-shot example was extracted from the training dataset based on similarity to the test input in CLIP embedding (Radford et al., 2021). We used accuracy as the performance metric following the official evaluation codes. More details are provided in Appendix C. To analyze the representational shift, we hypothesize that the effect of the the semantics is dynamic (varies across samples), while that of statistical biases is relatively static: shared *within a dataset* (e.g., a prompt format) or shared *across datasets* (e.g., the number of images). Since this potential mixture of the static and dynamic effects fits with the mixed effect model framework (Section 3.3), we implemented a linear mixed effect model consisting of a

random effect of H_{zsl} and a fixed effect of the variables representing a dataset (out of six datasets we used) and a LLaVA variant (LLaVA-1.5 or LLaVA-Llama2) to predict the shifted representation $\hat{H}_{icl}$. The formal definition of the mixed effect model is provided in Appendix B.1, and the discussion on model choice is in Appendix D.3.

5.1.2 Results

Fig. 2 shows the performance of LLaVA-Llama2. MM OoD ICL improved performance on datasets where the zero-shot performance was poor (e.g., VizWiz, TextVQA), supporting **Hypothesis I**. Conversely, MM OoD slightly degraded performance on datasets where the ID performance was already high (MMBench and MM-Vet).

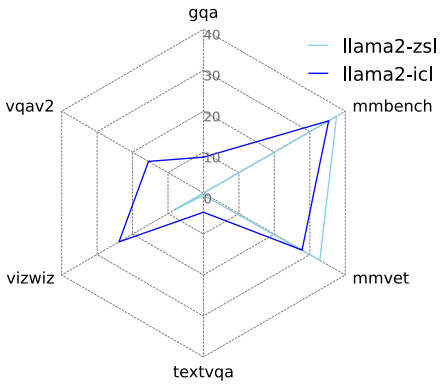


Figure 2: Performance summary of LLaVA-Llama2. zsl and icl represent zero-shot learning and in-context learning (ICL). ICL results in better performance for four datasets where the zero-shot performance is poor.

The mixed effect model showed higher explanatory power ($R^2 = 0.59 \pm 0.02$) compared to the random-effect-only baseline ($R^2 = 0.43 \pm 0.01$), suggesting that both the input shift and the fixed effects of dataset and model contribute to the output shift. The moderate explanatory power of our model validates the relevance of the input shift and output shift presupposed in RST.

Table 1: Regression Coefficient*100 of the mixed effect model's prediction with the dummy variables representing the datasets and the models. The prediction shows a much higher coefficient than the dummy variables, validating our models. For dataset variables, only two variables with the highest coefficient are shown. The full result is in Table 6.

(Intercept)	**mmbench**	**model**	**Input Shift**
9.2 ± 2.1	2.81 ± 0.7	-0.39 ± 0.4	70.33 ± 5.9

5.2 Experiment II: Reducing Label Bias with SM OoD ICL Examples

In Experiment II, we aimed to test **Hypothesis II** by investigating whether single-image, multi-turn OoD ICL examples can reduce label bias by encouraging the model to rely more on textual and visual information.

5.2.1 Counterfactual Prompting (CFP)

Most LLMs have safety limitations based on instruction tuning (Bianchi et al., 2023), which prevent them from generating hateful examples. To circumvent this while adhering to ethical guidelines, we employed CounterFactual Prompting (CFP). In our method, the model first generates text that fits with a given image to compose a benign meme. This generated meme serves as an ICL example to classify the test input as hateful or benign. Figure 4 shows a representative prompt.

5.2.2 Experimental Setup

We used the Hateful Memes Challenge dataset (Kiela et al., 2020), which is suitable for this experiment due to its small size, the dominance of textual modality, and inherent biases (Aggarwal et al., 2024; Zhang et al., 2023). We compared the performance of MM OoD and SM OoD ICL examples. For MM OoD, we extracted the ICL example based solely on textual modality using the BM25 algorithm (Robertson et al., 1996). We used LLaVA-Llama2 for its strong linguistic performance.

5.2.3 Results

As shown in Table 2, SM OoD ICL using CFP improved the model's F1 score on the Hateful Memes Challenge dataset compared to the zero-shot baseline, supporting **Hypothesis II**. In contrast, MM OoD ICL slightly decreased performance, suggesting that enhancing textual semantics through SM OoD is more effective in this context.

Table 2: Hateful memes detection performance. ZSL, MM OoD, and CFP represent Zero-Shot Learning, MM OoD, and Counterfactual Prompting, respectively. CFP's performance is better than ID while MM OoD dropped the performance, supporting Hypothesis I.

setting	ZSL	ICL	CFP
f1*100	61.4 ± 0.5	58.5 ± 0.9	62.2 ± 0.3

Figure 3 displays the cosine similarity matrix of the input shift. For ID and MM OoD, the hateful

and benign inputs have relatively high similarity, indicating that the model's representations for different labels are not well-separated. In the SM OoD case, the cross-label similarity drops significantly, suggesting that SM OoD ICL helps the model better distinguish between hateful and benign content by pulling apart their representations.

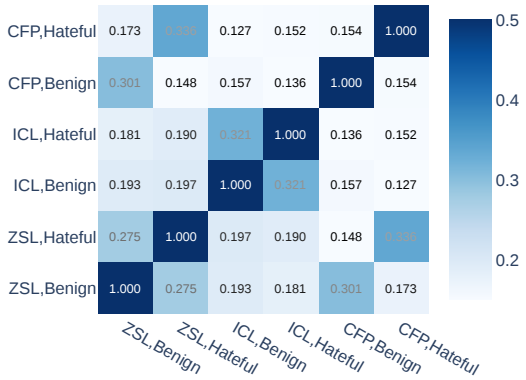


Figure 3: Representational shift across the learning type. Each entry is the similarity of the input between two conditions. For example, the left-top value 0.173 is the similarity of the input between hateful samples of a CFP setting and benign samples of a ZSL setting. While the hateful samples and the benign samples are similar for ZSL and ICL settings, CFP hateful samples and benign samples are less similar.

6 Discussion

In this paper, we proposed RST, a novel interpretability theory for ICL. RST states that the conditioning by an ICL example triggers two representational shifts, input shift and output shift. In light of RST, we formalized S^2 disentanglement as the optimization by two meta-gradient terms, and OoD generalization as an amplification of the dynamic semantic term over the constant statistics term. We further proposed two hypotheses for OoD generalization; First, even if the model is *not* trained with multi-image multi-turn datasets, an ICL image-text example can improve the performance when the test input's semantics is poor to the model (MM OoD; Hypothesis I). Second, curating a text-only ICL example can be a better solution when the textual modality is superior to the image modality (SM OoD; Hypothesis II). We validated Hypothesis I by performance improvement in four VQA datasets out of six, in which ID performance is poor. For Hypothesis II, We showed the supporting evidence in hateful meme detection; performance gain by counterfactual prompting while MM OoD does not work. We also showed the supporting evidence of the cascading representational shifts for each problem. More discussion is in Appendix E.

Although RST provides valuable insights into the role of ICL over S^2 disentanglement, we acknowledge that fully decomposing input information into *distinct* statistical and semantic components is challenging. For instance, typos in the dataset introduce noise that is non-statistical yet semantically insignificant. Nevertheless, we believe that analyzing data across various datasets can help mitigate the impact of such random noise. In addition, semantics and statistics may *interact* in the real world. For example, certain semantic content may only be understandable within specific statistical patterns. A complicated mathematical concept, for instance, might be best comprehended through equations. We acknowledge this interplay and will consider it in future work.

In this paper, we focused on VLMs due to the lack of interpretability studies despite their widespread use. However, we believe our theory can benefit LLMs solving text-only tasks. For example, a recent work suggested that LLMs recognize unseen image-like text generation tasks only when they have seen such text formats (Falkenstein et al., 2024). If our Hypothesis I (adding meaningful information works when the existing input is semantically poor) holds for this task, enhancing semantics (e.g., by concatenating textual description of the example) can improve performance even if the model is not trained on such tasks.

7 Conclusion

RST provides an analytical framework for studying the role of ICL over S^2 disentanglement, a central problem of interpretability. Based on RST, we formalized S^2 disentanglement in OoD generalization and showed that our hypothesis-driven approach can contribute to the performance gain in various problems. We believe our work will be the cornerstone for the study of *why* ICL works on real-world problems–our answer at this moment is *"Because the semantic information triggers the stream of representational shift."*.

8 Limitations

While our study provides valuable insights into S^2 disentanglement, there are several limitations and future research directions that warrant further investigation.

We limited our evaluation to one-shot ICL because few-shot ICL introduces additional complexities to our analysis, such as the fixed or random effects of varying the number of images. However, we acknowledge that applicability to few-shot ICL is critical, and we plan to tackle this challenge in future work.

Although RST can be used to analyze arbitrary problems, the largest limitation for the time being is its generalizability; to foresee the performance improvement in another problem, we need another hypothesis tailored to that problem. Towards the automatic formulation of the novel hypothesis, we believe the flexibility of semantic and statistic terms (Eq. 7) is the key. This study is also limited linguistically; we only used English datasets.

From a theoretical point of view, we have an intuitive leap from the existing works on meta-gradient; a nonlinearity. Despite previous works on *secretly* linear nature of a nonlinear Transformer (Razzhigaev et al., 2024) and our empirical findings supporting RST, applying the concept developed on a linear variant to the nonlinear one might hinder the precise evaluation. Recently, Ren and Liu (2023) proposed a theory for the nonlinear Transformer variants with the help of contrastive learning (Le-Khac et al., 2020). Unifying RST with their approach might provide a robust theoretical grounding.

In addition, whether the input shift *causes* the output shift is still elusive. An approach is to hire a mechanistic interpretability method, such as path patching (Hanna et al., 2023; Goldowsky-Dill et al., 2023). Training phase mechanisms such as grokking or double descent (Davies et al., 2022) should also provide an explanation for the *why* question.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Piush Aggarwal, Jawar Mehrabanian, and Weigang Huang. 2024. Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 104–117, St. Julian's, Malta. Association for Computational Linguistics.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Fifteenth International Conference on Computer Vision (ICCV15)*, Santiago, Chile.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *Preprint*, arXiv:2303.08112.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. *Preprint*, arXiv:2309.07875.

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. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2024a. Parallel Structures in Pre-training Data Yield In-Context Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8582–8592, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, Seattle, WA, USA.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. 2025. REVISITING IN-CONTEXT LEARNING INFERENCE CIRCUIT IN LARGE LANGUAGE MODELS. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, Singapore.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023a. Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers. In *Findings of the Association for Computational Linguistics*, pages 4005–4019. Association for Computational Linguistics.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. 2023b. Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5292–5305, Singapore. Association for Computational Linguistics.

Xander Davies, Lauro Langosco, and David Krueger. 2022. Unifying Grokking and Double Descent. In *MLSafety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2024. Jump to Conclusions: Short-Cutting Transformers with Linear Transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 9615–9625, Torino, Italia. ELRA and ICCL.

Janek Falkenstein, Carolin M. Schuster, Alexander H. Berger, and Georg Groh. 2024. From Language to Pixels: Task Recognition and Task Learning in LLMs. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 27–41, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Farquhar and Yarin Gal. 2022. What 'Out-of-distribution' Is and Is Not. In *MLSafety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA.

Rena Gao, Xuetong Wu, Siwen Luo, Caren Han, and Feng Liu. 2024. 'No' Matters: Out-of-Distribution Detection in Multimodality Long Dialogue. *arXiv preprint*.

Nathan Godey. 2024. On the Scaling Laws of Geographical Representation in Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12416–12422, Torino, Italia. ELRA and ICCL.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing Model Behavior with Path Patching. *arXiv preprint*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, The United States of America. IEEE.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, Salt Lake City, UT, USA. IEEE.

Wes Gurnee and Max Tegmark. 2024. Language Models Represent Space and Time. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Ahlem Hajjem, François Bellavance, and Denis Larocque. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. 2023. In-Context Learning of Large Language Models Explained as Kernel Regression. *Preprint*, arXiv:2305.12766.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA.

Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France.

Yunpeng Huang, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, Shupeng Li, and Penghao Zhao. 2024. Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey. *arXiv preprint*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA. IEEE.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and

Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Thirty-Fourth Annual Conference on Neural Information Processing Systems*, Red Hook, NY, USA.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2021. The Hateful Memes Challenge: Competition Report. *Proceedings of Machine Learning Research*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *The 3rd International Conference for Learning Representations (ICLR 2015)*, San Diego, CA, USA. arXiv.

Hao Lang, Yinhe Zheng, Binyuan Hui, Fei Huang, and Yongbin Li. 2024. Out-of-Domain Intent Detection Considering Multi-Turn Dialogue Contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12539–12552, Torino, Italia. ELRA and ICCL.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint*.

Jinhao Li, Haopeng Li, Sarah M Erfani, Lei Feng, James Bailey, and Feng Liu. 2024b. Visual-Text Cross Alignment:. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, Vienna, Austria.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as Algorithms: Generalization and Stability in In-context Learning. *Preprint*, arXiv:2301.07067.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Italy.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. MMBench: Is Your Multi-modal Model an All-around Player? In *WSDM '23: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1128–1131.

Ilya Loshchilov and Frank Hutter. 2019. DECOUPLED WEIGHT DECAY REGULARIZATION. In *The Seventh International Conference on Learning Representations*, New Orleans, LA, USA.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What Makes Chain-of-Thought Prompting Effective? A Counterfactual Study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Clement Neo, Shay B Cohen, and Fazl Barez. 2024. Interpreting Context Look-ups in Transformers: Investigating Attention-MLP Interactions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16681–16697, Miami, Florida, USA. Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting GPT: The logit lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context Learning and Induction Heads.

Matanel Oren, Michael Hassid, Yossi Adi, and Roy Schwartz. 2024. Transformers are Multi-State RNNs. *arXiv preprint*.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. 2023. Future Lens: Anticipating Subsequent Tokens from a Single Hidden State. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139.

Alan Ramponi and Barbara Plank. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. Your Transformer is Secretly Linear. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5376–5384, Bangkok, Thailand. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2024. Beyond Performance: Quantifying and Mitigating Label Bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.

Ruifeng Ren and Yong Liu. 2023. In-context Learning with Transformer Is Really Equivalent to a Contrastive Learning Pattern. *arXiv preprint*.

SE Robertson, S Walker, MM Beaulieu, M Gatford, and A Payne. 1996. Okapi at TREC-4. In *The Fourth Text REtrieval Conference (TREC-4)*, page 73.

Fabio Sigrist. 2023. Latent Gaussian Model Boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1894–1905.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318, Long Beach, CA, USA. IEEE.

Henrik Singmann and David Kellen. 2019. An Introduction to Mixed Models for Experimental Psychology. In Daniel Spieler and Eric Schumacher, editors, *New Methods in Cognitive Psychology*, 1 edition, pages 4–31. Routledge.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. FUNCTION VECTORS IN LARGE LANGUAGE MODELS. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.

Hugo Touvron, Louis Martin, and Kevin Stone. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*.

Keyon Vafa, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. 2024. Evaluating the World Model Implicit in a Generative Model. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, BC, Canada.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Thirty-First Annual Conference on Neural Information Processing Systems*, Long Beach, CA, USA.

Johannes von Oswald, Eyvind Niklasson, Randazzo, Ettore, Sacramento, Jo\~{a}o, Mordvintsev, Alexander, Zhmoginov, Andrey, and Vladymyrov, Max. 2023. Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 1464, page 24, Honolulu, HI, USA. JMLR.org.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A Survey on Natural Language Counterfactual Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. Association for Computational Linguistics.

Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. 2021. Online Adaptation to Label Distribution Shift. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An Explanation of In-context Learning as Implicit Bayesian Inference. In *The Tenth International Conference on Learning Representations*. arXiv.

Jiasheng Ye, Yawen Ouyang, Zhen Wu, and Xinyu Dai. 2022. Out-of-Distribution Generalization Challenge in Dialog State Tracking. In *Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *Preprint*, arXiv:2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA. arXiv.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. *Transactions on Machine Learning Research*, 2023.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. *Preprint*, arXiv:2401.13601.

Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. 2021. COPING WITH LABEL SHIFT VIA DISTRIBUTIONALLY ROBUST OPTIMISATION. In *The Ninth International Conference on Learning Representations (ICLR 2021)*.

Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating Biases in Hate Speech Detection from A Causal Perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6610–6625, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint*.

Shen Zhuoran, Zhang Mingyuan, Zhao Haiyu, Yi Shuai, and Li Hongsheng. 2021. Efficient Attention: Attention with Linear Complexities. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3530–3538, Waikoloa, HI, USA. IEEE.

A Formalization of Representational Shift Theory

A.1 Representational Shift

We formalize Representational Shift Theory (RST) by analyzing the difference between the zero-shot input-output pair $\{H_{zsl}, Y_{zsl}\}$ and that of ICL $\{H_{icl}, Y_{icl}\}$. Assuming that the effect of the instruction over an ICL example and over a zero-shot input is identical, i.e., $\Delta W_{inst/zsl} \simeq \Delta W_{inst/icl}$, we obtain the input shift:

$$H_{icl} - H_{zsl} \simeq -\Delta W_{icl/zsl} H_{inst} \quad (4)$$

Applying this to the output, we observe an output shift:

$$Y_{icl} - Y_{zsl} = -W_{emb} \Delta W_{icl/zsl} H_{inst} \quad (5)$$

Equations 4 and 5 represent the basic concept of RST. Note that the LLM's final output is a sequence of words, but we use the representation of the last decoder layer as the output for analysis. To intuitively analyze the multi-dimensional representation, we use a distance metric $D_{X/Y} \propto X - Y$:

$$\begin{aligned} D_{Y_{icl}/Y_{zsl}} &= W_{RST} D_{H_{icl}/H_{zsl}} \\ \text{where } W_{RST} &= -H_{inst}^T W_{emb} \end{aligned} \quad (6)$$

In practice, we use cosine similarity as the distance metric. This formalization allows us to analyze the effect of ICL by comparing the distances between representations and outputs.

A.2 S^2 Disentanglement

To disentangle semantics from statistics, we assume that the two concepts are independent. In RST, this implies that the weight updates due to semantics ΔW^{sem} and due to statistics ΔW^{stat} are discernible. We suggest that the semantic distance D^{sem} and the statistical distance D^{stat} are also separable, as indicated by the relevance of representational shift and the distance metric (Equation 6). We formalize the disentanglement as:

$$\begin{aligned} \Delta W_{icl/zsl} &= \Delta W^{sem}_{icl/zsl} + \Delta W^{stat}_{icl/zsl} \\ D_{H_{icl}/H_{zsl}} &= D^{sem}_{H_{icl}/H_{zsl}} + D^{stat}_{H_{icl}/H_{zsl}} \end{aligned} \quad (7)$$

This separation allows us to analyze how semantics and statistics individually contribute to the representational shift.

A.3 OoD Generalization as S^2 Disentanglement

An OoD input forces an LLM to generalize the same semantics under a significant distributional difference in statistics. Since the statistical difference (e.g., format difference) is consistent across all test inputs, its effect on the representational shift is constant (*fixed* effect). In contrast, the semantic term's effect varies across samples (*random* effect). Under this assumption, we formalize OoD generalization as a mixed effect:

$$D_{Y_{icl}/Y_{zsl}} = W_{RST}(D^{sem}_{W_{icl}/W_{zsl}} + W^{stat}) \quad (8)$$

A.3.1 Hypothesis I: MM OoD

Our first hypothesis is that MM OoD ICL examples are effective when the zero-shot input does not provide enough semantics to the model (i.e., poor zero-shot performance):

$$\begin{aligned} D^{sem}_{W_{icl}/W_{zsl}} &= W^{sem}_{icl} - W^{sem}_{zsl} \\ D_{Y_{icl}/Y_{zsl}} &= W_{RST}(W^{sem}_{icl} + W^{stat}) \\ \text{where } W^{sem}_{zsl} &\ll W^{sem}_{icl} \end{aligned} \quad (9)$$

One scenario is the lack of regularization in the attention matrix. If semantically similar ICL examples amplify the relevant context, our approach can alleviate irrelevant context, improving performance.

A.3.2 Hypothesis II: SM OoD

When textual semantics $W^{sem}(T)$ are more informative than image semantics $W^{sem}(I)$, enhancing the textual term through SM OoD ICL examples can be beneficial:

$$\begin{aligned} W^{sem}_{icl} &= W^{sem}_{icl}(T) + W^{sem}_{icl}(I) \\ D_{Y_{icl}/Y_{zsl}} &= W_{RST}(W^{sem}_{icl}(T) + W^{stat}) \\ \text{where } \Delta W^{sem}_{icl/zsl}(T) &+ \Delta W^{sem}_{icl/zsl}(I) \end{aligned} \quad (10)$$

For brevity, we assume the independence of semantics over the two modalities. This scenario is effective in addressing label bias (Reif and Schwartz, 2024), where the model's prediction may be biased toward certain labels due to over-reliance on statistical patterns.

B Other Formalization

B.1 Mixed Effect Model

In Experiment I, we implemented a linear mixed effect model to analyze the mixed effect of the input shift and confounding variables over the output shift. The model predicts the shifted representation $\hat{H}_{icl}$ as:

$$\hat{H}_{icl} = (W_r + W_f I)H_{zsl} + W_0 \quad (11)$$

Here, W_r represents the random effect, W_f represents the fixed effect, I is the embedding of fixed components (dataset and model), and W_0 is a bias term. The baseline model includes only the random effect:

$$\hat{H}_{icl} = W_{random}H_{zsl} + W_0 \quad (12)$$

By comparing the performance of these models, we assessed the contributions of the random and fixed effects.

B.2 Representational Analysis

In Experiment II, to visualize label bias, we estimated the input shift weight W_{RST} for different conditions (ID, MM OoD, SM OoD) and labels (benign, hateful). We computed the cosine similarity between weights to analyze how different approaches affect the model's internal representations:

$$\begin{bmatrix} \text{sim}(W^{zsl}_0, W^{zsl}_0) & \cdots & \text{sim}(W^{zsl}_0, W^{cfp}_1) \\ \vdots & \ddots & \vdots \\ \text{sim}(W^{cfp}_1, W^{zsl}_0) & \cdots & \text{sim}(W^{cfp}_1, W^{cfp}_1) \end{bmatrix} \quad (13)$$

Lower cross-label similarity indicates that the model better distinguishes between classes, reducing label bias. Note that we do not explicitly consider layer normalization in our formulation (Section 3.2 and A.1), our cosine similarity-based analysis implicitly accounts for normalization, focusing on direction rather than absolute values.

C Implementation Details

Experiments are conducted on a single NVIDIA A100 80GB GPU with Linux OS. Unless stated otherwise, all codes are written in Python 3.9. Statistical arguments are based on a t-test and bootstrapping with 1,000 resamples. We run the models once with a random seed of 1987. Eq. 11 and Eq. 12 are implemented on a PyTorch backend[5] and trained to maximize the cosine similarity of the output shift with Pytorch Metric Learning package[6]'s SelfSupervisedLoss under the AdamW optimizer (Loshchilov and Hutter, 2019). We extract 1,000 samples from each dataset and hold out 20% as a test set. The performance of this mixed effect model is evaluated using the marginal/conditional R^2 (Nakagawa and Schielzeth, 2013). To maintain the experiment's integrity while utilizing a wide range of statistical tools, the R language's *lmer* package is called from the Python environment via *rpy2*[7] module.

Fig. 4 illustrates a representative CFP prompt for Experiment II.

[5] https://pytorch.org/

[6] https://kevinmusgrave.github.io/pytorch-metric-learning/

[7] https://rpy2.github.io/doc.html

System:
You are a helpful language and vision assistant.
User:
<image in dataset>
Give me one caption that fits with this image.
Assistant:
{generated caption}
User:
In comparison with that caption, is the following caption hateful or non-hateful? Answer with a single word.
{caption in dataset}
Assistant:
{answer}

Figure 4: The representative counterfactual prompt. The system prompt is truncated for illustrative purposes. The curly brackets indicate the placeholders. Please see our code for the full version.

Table 3: Distribution of response lengths for correct vs. incorrect predictions made by LLaVA-Llama2 for the GQA dataset. Note the high occurrence of single-word responses in correct predictions.

Prediction	Response Length	
	Single Words	More Than Two
Correct	40	1,064
Incorrect	10,208	1,262

(A) *Incorrect* ZSL response by LLaVA–Llama2.

User:

Are there drapes to the right of the bed?
Answer the question using a single word or phrase.
Assistant:
yes, there are drapes to the right of the bed

(B) **Correct** ICL response by LLaVA–Llama2

User:

Do you see sheets on the bed?
Answer the question using a single word or phrase.
Assistant:
no
User:

Are there drapes to the right of the bed?
Answer the question using a single word or phrase.
Assistant:
yes

Figure 5: A representative LLaVA-Llama2's response from the GQA dataset. (a) LLaVA-Llama2 does not follow a part of the instruction where it is required to answer the question in a single word. (b) LLaVA-Llama2 responded in a single word with an ICL example.

D Additional Results

D.1 Qualitative Analysis

For a better understanding of the quantitative results, we performed a qualitative analysis. Here we show the results for GQA and TextVQA, using ground-truth labels (whether the model's answer is *correct* or *incorrect*) to alleviate the subjectivity.

D.1.1 LLaVA-Llama2

First, we show a representative example of the LLaVA-Llama2's response where it gives an incorrect answer in a ZSL setting, and it turns to the correct one when the ICL example is given (Fig. 5). In this example, the cause of the difference is the gain of *task-following* ability (Pan et al., 2023); The zero-shot response does not follow an instruction to answer in a single word (Fig. 5 (a)). In contrast, the ICL example allowed the model to follow this instruction (Fig. 5 (b)). Additionally, most LLaVA-Llama2's correct predictions in the OoD ICL settings are single words.

These results suggest that adding a simple mapping, instead of the ICL example, from the ZSL response to a single word may be sufficient for the performance gain. We test this hypothesis in Appendix D.2.

D.1.2 LLaVA-1.5

Second, we show two sets of LLaVA-1.5's responses for which ICL allows the model to answer the question **correctly** (Fig. 6) or the one for which it forces the model to answer *incorrectly* (Fig. 7). In the ICL example of Fig. 6, we can see the analogy between the two presented images (*one penny is the value of the copper coin*), potentially utilized by the model to make a **correct** prediction. In contrast, we observe the wrong label (pine and belt) in the ICL example of Fig. 7, which the model may refer to in making the *incorrect* prediction. Hypothesizing that the token-to-token interaction may result in these complex behaviors, we propose a Transformer-based algorithm for *switching* the strategy between ZSL and ICL in Appendix D.2.

D.2 On-the-Fly ZSL / ICL Switching

Since the OoD ICL does not work in every dataset or in every question-answer pair, performing OoD ICL *on-the-fly*, or using it *without prior ID evalua-*

(A) *Incorrect* ZSL response by LLaVA–1.5.

User:

How much is the copper colored coin worth?
Reference OCR token: [...], ONE, 2aH
Answer the question using a single word or phrase.
Assistant:
1

(B) **Correct** ICL response by LLaVA–1.5.

User:

Provide a one-sentence caption for the provided image.
Reference OCR token: L
Assistant:
A copper coin featuring the words ONE PENNY sits on top of a white surface.
User:

How much is the copper colored coin worth?
Reference OCR token: [...], ONE, 2aH
Answer the question using a single word or phrase.
Assistant:
1 penny

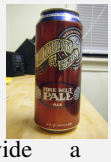
Figure 6: A representative LLaVA-1.5's response for which OoD ICL impacts the performance **positively** from the TextVQA dataset. (a) LLaVA-1.5 answered the monetary value question without its unit (*just 1, without penny*), potentially due to the missing OCR token for the unit. (b) LLaVA-1.5 answered *correctly*, potentially because the ICL example attributed **one penny** to the value of the copper coin.

(A) **Correct** ZSL response by LLaVA–1.5.

User:

What kind of beer is this?
Reference OCR token: NINK, [...], ALE
Answer the question using a single word or phrase.
Assistant:
ale

(B) *Incorrect* ICL response by LLaVA–1.5.

User:

Provide a one-sentence caption for the provided image.
Reference OCR token: NDCRAFTED, [...], PINE, BELT, PALE, ALE, [...]
Assistant:
a aluminum can of southern star pine belt ale
User:

What kind of beer is this?
Reference OCR token: NINK, [...], ALE
Answer the question using a single word or phrase.
Assistant:
pine belt

Figure 7: A representative LLaVA-1.5's response for which OoD ICL impacts the performance *negatively* from the TextVQA dataset. (Left) LLaVA-1.5 answered the **correct** object tag (**ale**). (Right) LLaVA-1.5 answered *incorrectly*, potentially caused by the wrong label (*pine* and *belt*) presented in the ICL example.

tion, is critical for real-world applications. Recent studies suggest that a Transformer's output can be approximated well before completing the response generation (Pal et al., 2023), using early-layer latent spaces (Din et al., 2024) or specific neuron activations (Neo et al., 2024). Building on these insights, we propose an algorithm for *switching* the strategy (ZSL or OoD ICL).

Specifically, given the test-input representation H_{zsl}, two auxiliary models $\{f_{zsl}, f_{icl}\}$ predict the probabilities of generating the correct answer in a ZSL or an OoD ICL setting, respectively. Next, the algorithm decides whether the LLaVA model $\mathcal{L}(X)$ should use the ICL prompt X_{icl} (a set of an instruction, an ICL example, and a test input), or keep the zero-shot prompt X_{zsl} (an instruction and a test input). Altogether, our algorithm for generating its output Y_{alg} is summarized as:

$$Y_{alg} = \begin{cases} \mathcal{L}(X_{icl}) & \text{if } f_{icl}(H_{zsl}) > f_{zsl}(H_{zsl}) \\ \mathcal{L}(X_{zsl}) & \text{otherwise} \end{cases} \quad (14)$$

Since we do *not* use the shifted representation H_{icl}, we can use our algorithm *on-the-fly*, without prior ID evaluation.

Empirically, we tested this algorithm with GQA and TextVQA datasets that allow a locally reproducible binary evaluation (about whether the answer is *correct* or *incorrect*). For the LLaVA model $\mathcal{L}$, we used both LLaVA-Llama2 and LLaVA-1.5. For an auxiliary model f, we use a single linear layer for LLaVA-Llama2, assuming that the enhanced task-following ability observed in qualitative analysis (Appendix D.1) can be achieved

Table 4: Accuracy of the on-the-fly context selection. Bold indicates the best performance in each row. In the LLaVA-Llama2 case, the performance is bounded by the OoD ICL accuracy since ZSL performance is extremely low. In the LLaVA-1.5 case, it outperforms *both ZSL and ICL*, suggesting its efficiency when the performance of the two strategies is comparable.

Model	Dataset	ZSL	OoD ICL	Eq. 14
LLaVA–Llama2	TextVQA	0.9	**4.7**	2.2
	GQA	0.0	**9.0**	6.5
LLaVA-1.5	TextVQA	61.6	57.0	**63.8**
	GQA	65.7	56.0	**68.2**

with a simple approach. To account for more complex findings for LLaVA-1.5, we use a single-layer Transformer as f for this variant. We trained auxiliary models with 70% of the GQA and TextVQA test sets, and tested with the remaining 30%. We used binary cross-entropy[8] with an Adam (Kingma and Ba, 2015) optimizer to learn the mapping between the zero-shot representation H_{zsl} and the ground-truth label (*correct* or *incorrect*). For LLaVA-Llama2, to handle the scarcity of the *correct* label, we weighted *correct* label ten times[9] higher than the *incorrect* label.

We show the results in Fig. 4. In a LLaVA-Llama2 case, since the ZSL performance is extremely low, the algorithm's performance is bounded by the ICL performance. However, we obtained the $1.3 - 6.5$ point gain in absolute accuracy over ZSL. For LLaVA-1.5, the algorithm outperformed both ZSL and ICL. These results suggest that this on-the-fly approach can provide 1) a moderate performance gain when the ZSL performance is quite low 2) effectively *switch* the strategy between ZSL and ICL when their performances are comparable. We leave the design of algorithms for more challenging settings (e.g., an unsupervised learning approach in the absence of ground-truth labels) to future work.

D.3 Mixed Effect of Semantics and Statistics: An Arbitrary Argument?

Although we believe that our assumptions for using a mixed effect model (Section 3.3 and Appendix B.1) in Experiment I is logically sound, we acknowledge that this model choice may be arbitrary. For example, we can also model the *random* effect of a *statistical* pattern–For example, the potentially negative impact of an OCR tag present in the ICL example (Fig. 7) may be considered as such. One hypothesis is that the random effect of the semantics impacts the prediction *positively*, while the fixed effect of the statistical pattern works *negatively*. To test this idea, we analyzed the weights of the mixed effect model (Eq. 11) and the random-effect-only baseline (Eq. 12). We show the result in Table 5. Compared with the baseline, we can see that the mixed effect model has a larger slope term and a smaller bias term in average. We take this as evidence supporting our hypothesis.

Table 5: Weight means and standard deviations (×10 000) for random vs. random+fixed settings.

Weight Type	Mean ± Std ($\times 10^4$)	
	Random	Random + Fixed
Slope	1.88 ± 0.09	110.23 ± 2.75
Bias	-6.33 ± 0.38	-508.79 ± 27.07

D.4 Impact of CLIP

Solely based on the CLIP-based ICL selection, we cannot rule out the possibility that *any* OoD ICL example can affect the performance. To test this, we randomly sampled ICL examples from the training dataset and performed qualitative analysis on the samples where ICL improved the performance. In all samples we observed, the randomly sampled ICL example does not improve the performance, suggesting the significance of the semantically rich ICL example. We show a representative sample in Fig. 8. All samples will be available online by the publication.

Note that the method-to-method comparison (e.g., between CLIP and VLM-based similarity search (Li et al., 2024b)) is challenging for our case for potential circular reasoning[10]. An idea is to use task-specific criteria to define *semantic richness*. We leave further methodological exploration for future work.

D.5 Full Result for Table 1

The full result for Table 1 is shown in Table 6.

[8] https://docs.pytorch.org/docs/stable/generated/torch.nn.BCELoss.html

[9] Aside from 1 : 10, We also tested 1 : 1,1 : 2,1 : 5,1 : 20, and 1 : 10 works the best. We also did some preliminary experiments on Focal Loss (Lin et al., 2017), but did not see a significant performance gain in the preliminary experiments.

[10] Since method A results in better accuracy than method B, A's example is semantically richer than B's, because A results in [...]

User:

Provide a one-sentence caption for the provided image.
Reference OCR token: LERIE, HOT
Assistant:
A bottle with a blue top with LERIE written on it.
User:

How much is the copper colored coin worth?
Reference OCR token: [...], ONE, 2aH
Answer the question using a single word or phrase.
Assistant:
1

Figure 8: A representative LLaVA-1.5's response when the ICL example is randomly sampled from the training dataset. Seemingly irrelevant image-text ICL example does not affect the model's response.

Table 6: Full list of regression coefficients of the mixed effect model's prediction with the dummy variables representing the datasets and the models in Experiment I. The prediction shows a much higher coefficient than the dummy variables, validating our models.

variable	coef*100
(Intercept)	9.2 ± 2.1
mm-vet	-0.75 ± 0.7
mmbench	2.81 ± 0.7
textvqa	2.1 ± 0.6
vizwiz	0.16 ± 0.7
vqav2	-0.12 ± 0.6
model	-0.39 ± 0.4
Input Shift	70.33 ± 5.9

D.6 LLaVA-1.5

We show LLaVA-1.5's performance (Fig. 9). LLaVA-1.5 outperforms LLaVA-Llama2 in all cases, reflecting the authors' additional training efforts (Liu et al., 2023a).

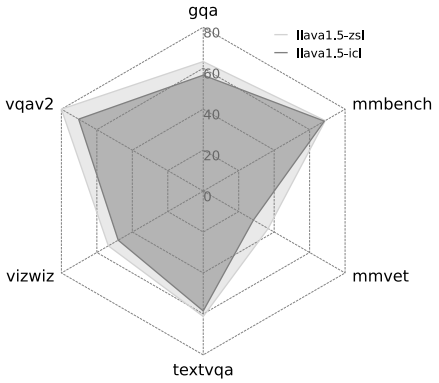


Figure 9: The performance summary of LLaVA-1.5. OoD ICL dropped the performance, suggesting the rich semantics in the test input.

D.7 High-Level Analysis on Mixed Effect

In addition to fine-grained analysis in Table 1, we analyzed the dataset-level mixed effect. In this analysis, the effects are represented as a coefficient of the corresponding one-hot encodings. Specifically, we modeled the accuracy of each dataset as a sum of the effect of a variable representing the presence/absence of an OoD ICL example and that of the variable representing the models and datasets. The result suggests that the model variable drives the explanatory power at this level, consistent with the performance summary (Fig. 2), which shows the drastic improvement of LLaVA-1.5 over LLaVA-Llama2.

Table 7: Regression coefficients of the variables representing model (LLaVA 1.5 or LLaVA-Llama2), dataset, and presence/absence of ICL examples. *all* represents the result of an all-variable model. R^2 values are multiplied by 100 for brevity. The result only with the model variable is similar to the all-variable model, consistent with the performance summary (Fig. 2).

Variable		R^2*100	
Fixed	Random	Fixed	Random
model	model	22.6 ± 3.0	52.0 ± 8.8
dataset	ICL	0.3 ± 0.1	0.5 ± 0.2
model	ICL	33.5 ± 2.4	33.6 ± 2.5
dataset	model	0.2 ± 0.1	49.5 ± 2.7
all	all	23.7 ± 4.4	53.7 ± 8.8

D.8 Preliminary ID Analysis: InternVL

To test if the findings about LLaVA is transferred to an ID setting, we also use InternVL (1-2 billion)

for its limited [11] yet tested multi-image capabilities by multi-image datasets like MMMU (Yue et al., 2024).

In the case of InternVL, MM OoD generally dropped the performance, potentially because of its high performance and multi-image resource shortage (Fig. 10). To see whether the task difficulty

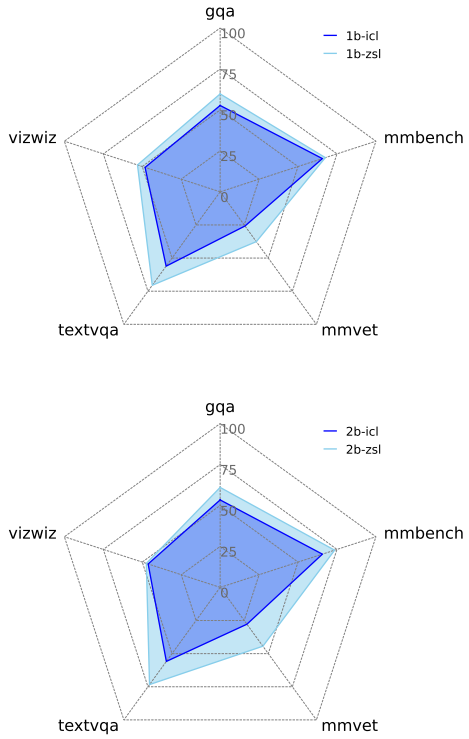


Figure 10: Performance summary of InternVL. MM OoD dropped the performance for all the datasets, potentially reflecting that the baseline performance is moderate to high for all the datasets.

(i.e., semantic poorness to the model) affects this trend, we see the performance by the number of reasoning steps provided by the GQA dataset evaluation, typically seen as the difficulty metric. Divided by this subcategory, ICL performs slightly better when the number of steps is larger (Table 8). Together with LLaVA results, these results suggest that the performance boost may serve as a task difficulty indicator.

[11] https://github.com/OpenGVLab/InternVL/issues/419

Table 8: Impact of multi-image ICL in GQA for InternVL 1b. N steps indicate the number of inference steps. The numbers with an error represent accuracy(%) in the corresponding setting. ICL boosted the performance when the number of steps was above six, implying that the ICL positively affects the performance when the task is challenging.

N Steps	N Samples	ZSL	ICL
1–5	12,153	59.7 ± 0.15	52.5 ± 0.31
6–9	65	83.5 ± 0.24	84.6 ± 0.27

E Additional Discussion

E.1 Our Contribution in Relation to Related Work

E.1.1 Meta-Gradient

Previous efforts on building interpretability theories for ICL have validated the concept of meta-gradient, attention weight used as a form of gradient (von Oswald et al., 2023; Dai et al., 2023a). Meta-gradient backbones RST, which provides an analytical framework for S^2 disentanglement. Towards S^2 disentanglement, interpretability studies disentangled a few aspects of the semantics. Inspired by these works, RST provides a unifying framework for S^2 disentanglement.

E.1.2 OoD

Various OoD problems have been explored, such as multi-turn OoD (Ye et al., 2022). We extend the scope to the multi-image multi-turn setting.

E.1.3 Potential Mechanism

Although our theory provided the general framework for ICL analysis, its detailed mechanism is elusive. For example, linguistic patterns inherent in ICL examples may contribute to OoD ICL. For example, the model may learn parallel structures, or the repetitive occurence of specific tokens, such as $[Image\ token] \rightarrow [Textual\ context] \rightarrow [Image\ token] \rightarrow [Test\ input]$.

Likewise, the model's capability of learning such a token sequence is a promising candidate for the mechanism underlying the representational shift. One of the famous mechanisms is an induction head, with which the model performs token completion. In ICL for labeled datasets, the induction head may utilize the label to perform a simple completion like $[ICL\ exampleA] \rightarrow [Label\ Y1] \rightarrow [Test\ input\ B,\ which\ is\ similar\ to\ A] \rightarrow$

$[Completed\ Label\ Y1]$. This is less likely in unlabeled datasets (e.g., MM-Vet) but is a promising area of research.

Also, our results support the presence of function vectors in that the latent space right after presenting an ICL example influences model behavior. As with preceding studies, we plan to explore which attention heads cause the representational shift in future work.

E.2 Other Applications

Although RST provides valuable insights into the role of ICL over S^2 disentanglement, our future work should include the analysis of other OoD problems (e.g., multi-turn OoD in general) and ID problems where semantics and statistics are potentially more entangled (e.g., MMMU (Yue et al., 2024)). In that case, we can also extend the subject to the large variety of LLMs, including the ones trained with multi-image datasets such as LLaVA-Next (Li et al., 2024a).

F Other Considerations

F.1 Potential Risks

A hateful meme is a highly sensitive research topic. Therefore, all the hateful meme research involves risks and uncertainty to some extent. For example, the attackers may read a publication about a hateful meme detector to create a new meme that the detector may not be able to detect. More broadly, all LLM-related papers can be maliciously used when they are in the wrong hands (e.g., to improve an LLM trained on the dark web). To overcome these issues, an iterative update of the methodology with safety measures is a must.

F.2 Ethical Considerations

The hateful memes challenge dataset (Kiela et al., 2020, 2021) contains sensitive content. Therefore, we refrained from showing actual hateful memes so that this paper does not negatively impact any targeted group. We refer the users to the original publication for the considerations taken in dataset curation.

F.3 AI Assistant Usage

We used GitHub Copilot for efficient coding and ChatGPT for linguistic improvements.

F.4 License and Usage of Scientific Artifacts

We declare that all scientific artifacts used in this study do not prohibit the use of artifacts for academic research.

F.5 Documentation of Artifacts

Experiment I uses the test split of six VQA datasets. GQA contains 10% of $22,669,678$ questions over $113,018$ images. TextVQA contains $5,734$ text-image pairs. VizWiz contains $8,000$ visual questions. VQAv2 contains $447,793$ questions for $81,434$ images. MMBench contains $1,784$ questions. MM-Vet contains 218 questions.

Experiment II is performed on test-seen split of a hateful meme challenge dataset with $1,000$ text-image pairs (510 benign samples and 490 hateful samples).