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. (2023b)), interpreting its inner workings is critical for reliable, 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., 2023b) 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 two-dimensional image tensor in OoD, whereas the model is solely trained with the tensor with 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.

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

First, we introduce meta-gradient, the fundamental concept of this study. Second, we formalize unembedding as a connection between LLM's intermediate representation and the textual response. Finally, we summarize a mixed effect model, a useful tool in this study.

3.1 Meta-Gradient

Centric to the optimization of the 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 4.2, 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 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 Representational Shift Theory (RST)

First, we formalize RST, stating that an ICL example affects the representation of the zero-shot input (*input shift*) and then that of the output (*output shift*). Second, we show how it relates to S^2 disentanglement. Third, we frame OoD generalization into S^2 disentanglement. Lastly, we suggest two hypotheses for improving the generalization:

1. When the zero-shot (In-Distribution, ID) input is semantically poor for an LLM, a Multi-image Multi-turn OoD (MM OoD) ICL example is helpful.
2. When the textual semantics are superior to the image semantics, a Single-image Multi-turn OoD (SM OoD) ICL example is helpful.

4.1 Representational Shift

Here we show the representational shift between the zero-shot input-output pair $\{H_{zsl}, Y_{zsl}\}$ and that of ICL $\{H_{icl}, Y_{icl}\}$. First, assuming in Eq. 1 that $\Delta W_{inst/zsl} \simeq \Delta W_{inst/icl}$, or the identical effect of instruction over an ICL example and over a zero-shot input, we obtain the input shift as follows.

$$H_{icl} - H_{zsl} \simeq -\Delta W_{icl/zsl} H_{inst} \tag{4}$$

Applying to Eq. 2, we see that this shift triggers an output shift.

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

Eq. 4 and Eq. 5 represent the basic concept of RST. Note that LLM's final output is a sequence of words, but we use the representation of the last decoder layer as the output. To analyze the multi-dimensional representation in an intuitive way, we assume that the difference of the two matrices is represented by a distance metric $D_{X/Y} \propto X - Y$.

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

In summary, if RST is valid, we can analyze the effect of ICL by comparing the distance of the two representations and that of the two outputs. In practice, we use widely used cosine similarity as the distance metric.

4.2 S^2 Disentanglement

To disentangle semantics from statistics, we are obliged to assume that the two concepts are independent. In RST, this implies that the weight update by semantics ΔW^{sem} and the update by statistics ΔW^{stat} are discernable. We also suggest that the semantic distance D^{sem} and the statistic distance D^{stat} is also separable since we suggested the relevance of the representational shift and the distance metric (Eq. 6). In summary, we formalize the disentanglement as follows.

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

4.3 OoD Generalization as S^2 Disentanglement

An OoD input forces an LLM to generalize to the same semantics under the drastic distributional difference in statistics. This statistical difference (e.g., a format difference) is static over all the test inputs. Therefore, its effect on the representational shift is also supposedly constant (*fixed* effect). In contrast, the semantic term's effect is intuitively diverse across the samples (*random* effect). Under this assumption, we formalize OoD generalization as a mixed effect; maximizing the random effect of the semantic term under the fixed effect of the OoD statistics W^{stat}.

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

4.3.1 Hypothesis I: MM OoD

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

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

One such scenario is the lack of regularization in the attention matrix. Specifically, Ye et al. (2024) suggested that a significant proportion of attention values are wasted on the semantically irrelevant context. If the major components of the semantically similar ICL example *amplify* the relevant context, we suggest our approach is effective for the irrelevant context alleviation.

4.3.2 Hypothesis II: SM OoD

Since encoders of most LLMs are first trained solely by the text and then jointly trained by the VL datasets, we assume that the textual semantics $W^{sem}(T)$ is greater than the image semantics $W^{sem}(I)$ in some cases. In this case, we hypothesize that enhancing the textual term (SM OoD) would provide a solution where MM OoD does not work.

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

Here we suppose the independence of the semantics over the two modalities for brevity. The interaction complicates the problem in reality (e.g., Miyanishi and Nguyen (2024)), but we leave this interaction to future work.

One scenario in which our approach is effective is the label bias (Reif and Schwartz, 2024); When the test input or ICL example is *salient*, the prediction may be biased towards the novel label. For example, when a general-purpose LLM is required to detect hate speech from aggressive language, the prediction may be biased towards the high hatefulness as such inputs are likely to be filtered out in the training to prevent the model from learning it.

5 Experiments

We conducted two experiments: Experiment I for MM OoD and Experiment II for SM OoD. We hired two LLaVA (Liu et al., 2023b) variants (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 2) they are *NOT* trained with the multi-image resources or ICL settings, allowing OoD analysis. We used 13 billion parameter models to balance linguistic capability and memory constraint. We also did a preliminary experiment with InternVL (Chen et al., 2024) for the analysis in an ID setting (Appendix B.3). We focused on ICL with a single example since we did not see any positive clue for further concatenation in the initial exploration.

In Experiment I, we used six VQA datasets to test Hypothesis I. First, we evaluated LLaVA's zero-shot (In-Distribution; ID) and one-shot (Multi-image Multi-turn OoD, MM OoD) performance. The one-shot example was extracted from the training dataset based on similarity to the test input in CLIP embedding. The result suggests that MM OoD improves the performance for the datasets in which the ID performance is poor ($5 \sim 20\%$ of the accuracy for four datasets), supporting Hypothesis I. Next, we analyzed the mixed effect in the MM OoD; the random effect of the input shift over the output shift, and the fixed effect of datasets and models. The moderate explanatory power of our model ($R^2 = 0.59 \pm 0.02$, $\sim 70\%$ in coefficient analysis) validates RST.

To validate Hypothesis II, we focus on the data where the textual modality is dominant. To this end, we used the hateful memes challenge (Kiela et al., 2020) dataset in Experiment II, which is known for this textual dominance (Aggarwal et al., 2024). In

contrast with MM OoD which degraded the performance over ID ($\sim$ 2.9 points of the F1 score), we found that using SM OoD examples curated by CounterFactual Prompting (CFP) improved the performance ($\sim$ 0.8 points), supporting Hypothesis II. Next, to visualize the representational shift over ID / MM OoD / SM OoD, we analyze the similarity of the weight W_{RST} (Eq. 10). The result shows that, unlike ID or MM OoD, SM OoD shifts the representation of the hateful inputs *away from* that of benign inputs.

5.1 Experiment I: MM OoD

5.1.1 Mixed Effect Model

RST suggests that the random effect of an OoD ICL example drives ICL. Since interpretability favors simplicity (Park et al., 2023), we implement a linear mixed effect model which predicts the shifted representation $\hat{H}_{icl}$.

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

The linear weights W_r, W_f, and W_0 represent the random effect, the fixed effect, and a bias term, respectively. For the mixed effect with identical dimensionality, we use the embedding I representing the fixed components (dataset and model). We use a model only with the random effect as a baseline.

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

To see the effect of the input shift over the output shift, we then calculate the distance $D_{H_{icl}/H_{zsl}}$ between the zero-shot representation H_{zsl} and the shifted one $\hat{H}_{icl}$. We finally applied a linear regression between this input distance $D_{H_{icl}/H_{zsl}}$ and output distance $D_{Y_{icl}/Y_{zsl}}$, together with the dummied variables representing models and datasets for residual analysis.

5.1.2 Other Settings

We use CLIP (Radford et al. (2021), specifically HuggingFace *clip-vit-large-patch14*), for ICL example selection because of its relatively small computational cost and high capability on similarity-related tasks.

To cover various aspects of VL capabilities, we used six VQA datasets, namely VQA v 2.0 (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)). We use accuracy as a performance metric following the official evaluation codes. More details are in the Appendix A.

5.1.3 Results

First, we show LLaVA-Llama2's performance (Fig. 1). MM OoD dropped performance for MMBench and MM-Vet where the ID performance is relatively high. In contrast, MM OoD *improved* LLaVA-Llama2's performance for the rest of the datasets in which ID performance is lower. These results suggest MM OoD's positive impact where the test input is semantically poor for the model, supporting Hypothesis I. Next, we analyzed the mixed effect of

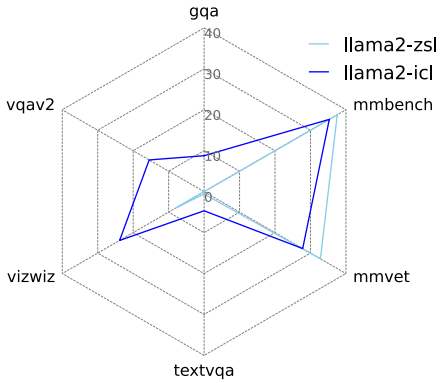

Figure 1: Performance summary of LLaVA-Llama2. zsl and icl represent zero-shot learning and in-context learning (ICL). ICL results in better performance for four datasets where the zero-shot performance is poor.

the input shift and the confounding variables over the output shift. The mixed effect model (Eq. 11) showed a higher performance ($R^2 = 0.59 \pm 0.02$) than the random-effect-only baseline (Eq. 12; $R^2 = 0.43 \pm 0.01$), supporting the explanation by mixed effect. Finally, we analyzed the regression coefficient to see the impact of the random effect and the fixed effect (Table 1). The input shift shows moderate explanatory power, validating the relevance of the input shift and the output shift presupposed in RST.

5.2 Experiment II: SM OoD

5.2.1 CFP

Most LLMs have safety limitations based on instruction tuning (Bianchi et al. (2023)) which does not allow them to generate hateful examples. Since bypassing such limitations is neither desirable nor sustainable, we let the model generate *negative* examples. In short, the model first generates text that fits with a given image to compose a benign

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
Model Prediction	70.33 ± 5.9

Table 1: Regression Coefficient 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.

meme. Next, using that meme as an ICL example, the model classifies the test input as hateful or benign. Fig. 3 shows a representative prompt. Qu et al. (2023) introduced another workaround of using more general labels, which will be a part of our future work.

5.2.2 MM OoD vs. SM OoD

To see how the effect of input shift differs between MM OoD and SM OoD, a straightforward approach is to analyze the difference in the relationship between the two shifts. To this end, we first estimate the input shift weight W_{RST} (Eq. 10) for ID, MM OoD, and SM OoD (shown in Eq. 13 as W^{zsl}, W^{icl}, and W^{cfp}, respectively). In ID case, we use the shifts caused by the instruction $\{D_{W_{zsl}/W_{inst}}, D_{Y_{zsl}/Y_{inst}}\}$ for reference. We build a single estimator for consistent representation. Then we calculate the similarity between each weight. To visualize label bias, this process is split by the ground-truth label, denoted as W_0 and W_1 where 0 and 1 stand for benign and hateful, respectively. Altogether we obtain similarity matrix as:

$$\begin{bmatrix} sim(W_0^{zsl}, W_0^{zsl}) & \cdots & sim(W_0^{zsl}, W_1^{cfp}) \\ sim(W_1^{zsl}, W_0^{zsl}) & \cdots & sim(W_1^{zsl}, W_1^{cfp}) \\ sim(W_0^{icl}, W_0^{zsl}) & \cdots & sim(W_0^{icl}, W_1^{cfp}) \\ \vdots & \ddots & \vdots \\ sim(W_1^{cfp}, W_0^{zsl}) & \cdots & sim(W_1^{cfp}, W_1^{cfp}) \end{bmatrix} \quad (13)$$

We use cosine similarity as a similarity function $sim(\cdot)$. The weights are estimated per layer dimension to perform memory-efficient analysis.

5.2.3 Hateful Memes Challenge

Intuitively, the impact of counterfactual prompting may vary across datasets. The most influential scenario is 1) the dataset size is small for which ICL example selection is challenging 2) the textual modality is superior to the image modality 3) the bias factors are embedded on the dataset.
Kiela et al. (2020) curated the Hateful Memes Challenge dataset, which perfectly fits this experiment's criteria. First, Laurençon et al. (2023); Zhao et al. (2023a) showed that ICL is not particularly effective unless the model is heavily tuned to the task. Second, Aggarwal et al. (2024) showed the superiority of the textual modality. Third, Hee et al. (2022); Zhang et al. (2023) indicated the presence of various sources of the bias. Since the data size is small, we use f1 score to see the precision-recall balance. We leave more experiments on hateful meme detection (e.g., Gomez et al. (2020)) and other tasks to future work.

5.2.4 Other Settings

Since we assume the superiority of the textual modality, we use LLaVA-Llama2 in this experiment for its strong linguistic performance. Toward MM OoD baseline fully utilizing textual modality, we extract the ICL example solely based on textual modality with BM25 algorithm (Robertson et al., 1996).

5.2.5 Results

First, we see the performance of ID, MM OoD, and SM OoD on the hateful memes challenge dataset. Contrally to the performance drop in MM OoD, SM OoD slightly improved the performance, supporting Hypothesis II (Table 2). Further exploration of ICL methodologies will be part of our future work. Next, we built a mixed effect model (Eq. 10)

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

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.

for label bias visualization, showing a moderate AUC of 75.6 ± 0.90. Finally, we calculated the

similarity matrix of the weight W_{RST} (Fig. 2). For ID and MM OoD, the hateful inputs are relatively similar to the benign inputs of the same condition (cosine similarity $\simeq 0.3$). This cross-label similarity dropped significantly (< 0.2) for SM OoD. These results suggest that SM OoD *pulls away* the inputs of different labels which ID or ICL cannot distinguish well.

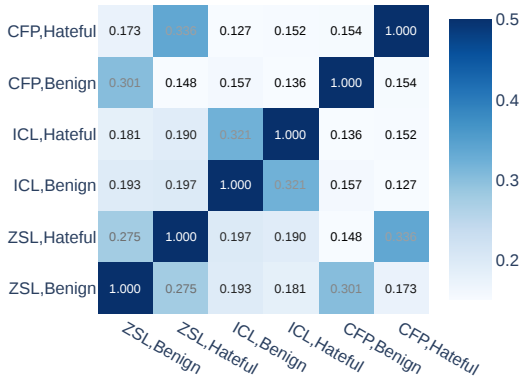

Figure 2: Representational shift across the learning type. Each entry is the similarity of the input between two conditions. For example, the left-top value 0.173 is the similarity of the input between hateful samples of a CFP setting and benign samples of a ZSL setting. While the hateful samples and 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.

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, such as color (Abdou et al., 2021), geography (Godey, 2024) and world model (Vafa et al., 2024). Inspired by these works, RST provides the unifying framework for S^2 disentanglement. On the other hand, 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.

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 (Liu et al., 2024).

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. 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.

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.

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. 2024. 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.

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.

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*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, Snowmass Village, CO, USA. IEEE.

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.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On Explaining Multimodal Hateful Meme Detection Models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655, Virtual Event, Lyon France. ACM.

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*.

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.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Thirty-Seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA.

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

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.

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, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

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.

Yosuke Miyanishi and Minh Le Nguyen. 2024. Causal Intersectionality and Dual Form of Gradient Descent for Multimodal Analysis: A Case Study on Hateful Memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2901–2916, Torino, Italia. ELRA and ICCL.

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.

nostalgebraist. 2020. Interpreting GPT: The logit lens.

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

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, New Orleans, LA, USA.

Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. 2023. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *The 44th IEEE Symposium on Security and Privacy*, San Francisco, CA, USA.

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.

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.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2024. Differential Transformer. *arXiv preprint*.

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.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023a. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. *Preprint*, arXiv:2309.07915.

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. 2023b. 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.

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 benign? Answer with a single word.
{caption in dataset}
Assistant:
{answer}

Figure 3: The representative counterfactual prompt. The system prompt is truncated for illustrative purpose. Please see our code for the full version.

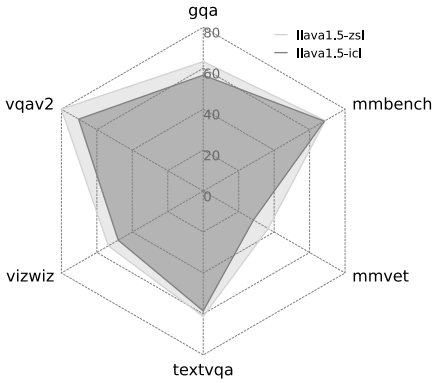

Figure 4: The performance summary of LLaVA-1.5. OoD ICL dropped the performance, suggesting the rich semantics in the test input.

A 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[3] and trained to maximize the cosine similarity of the output shift with Pytorch Metric Learning package[4]'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*[5] module.
Fig. 3 illustrates a representative CFP prompt for Experiment II.

B Additional Results

B.1 LLaVA-1.5

We show LLaVA-1.5's performance (Fig. 4). LLaVA-1.5 outperforms LLaVA-Llama2 in all cases, reflecting the authors' additional training efforts (Liu et al., 2023a).

[3]https://pytorch.org/
[4]https://kevinmusgrave.github.io/pytorch-metric-learning/
[5]https://rpy2.github.io/doc.html

B.2 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. 1), which shows the drastic improvement of LLaVA-1.5 over LLaVA-Llama2.

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

Table 3: 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. 1).

B.3 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 [6] 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. 5). To see whether the task difficulty (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 4). Together with LLaVA results, these results suggest that the performance boost may serve as a task difficulty indicator.

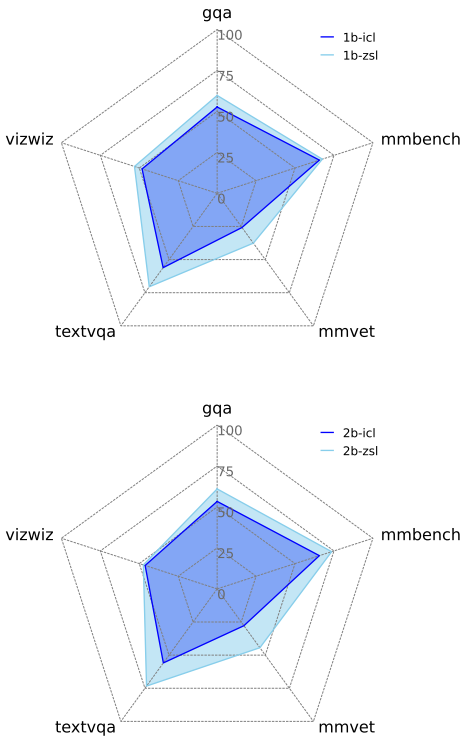

Figure 5: Performance summary of InternVL. MM OoD dropped the performance for all the datasets, potentially reflecting that the baseline performance is moderate to high for all the datasets.

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

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

Table 4: 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.

C Other Considerations

C.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.

C.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.

C.3 AI Assistant Usage

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

C.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.

C.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).