Explaining is Harder Than Predicting Alone: Evaluating Concept-based Explanations of MLLMs as ICL Visual Classifiers

Carmen Quiles-Ramírez [1] **Leticia L. Rodríguez** [2] **Nicolás Martorell** [2] **Natalia Díaz-Rodríguez** [1]

Abstract

In-context learning (ICL) enables multimodal large language models (MLLMs) to classify images from a few labelled examples. Yet *how* these models use the provided context remains opaque. While Chain-of-Thought prompting is widely used, recent work argues it may not reflect true internal computation (Barez et al., 2025). In this paper, we systematically evaluate the concept-based explainability of frozen MLLMs under few-shot ICL using five conditions of increasing formal rigour: from baseline classification to Description Logics (DL) axiom generation. Evaluating four state-of-the-art MLLMs via an independent LLM-as-a-judge pipeline, we demonstrate that *explaining is genuinely harder than predicting alone*. Surprisingly, forcing models to generate formally structured, concept-based explanations degrades predictive accuracy monotonically (from 93.8% to 90.1%), contradicting the assumption that explicit reasoning universally aids performance. However, when models successfully articulate class-discriminative visual features, explanation quality strongly correlates with correct predictions. Our findings suggest that while MLLMs excel at visual classification, they lack the specific instruction-tuning required for formal, machine-verifiable explainability.

1. Introduction

In-context learning (ICL) (Brown et al., 2020) has emerged as a powerful paradigm for adapting large language models to new tasks without parameter updates. Extended to the multimodal setting, MLLMs can now classify images from a few labelled examples provided directly in the context window (Dong et al., 2024). Despite impressive few-shot accuracy, the reasoning process underlying these predictions remains a black box: models produce answers without any principled account of *why* a particular image belongs to a given class.

The dominant approach to eliciting explanations from large models is chain-of-thought (CoT) prompting, which asks the model to verbalise intermediate reasoning steps. However, Barez et al. (2025) argue that CoT does not constitute genuine explainability: the generated text may diverge from the model's actual internal computation, offering only a plausible-sounding post-hoc narrative. Similarly, Huang et al. (2025) show that VLMs tend to mimic rather than reason from ICL context, further questioning whether CoT-style outputs constitute genuine understanding. This gap motivates the search for *concept-based explanations* (Poeta et al., 2025) —formal statements that are both human-readable and machine-verifiable.

We approach this problem through the lens of eXplainable AI (XAI) (Ali et al., 2023) and Description Logics (Baader et al., 2009), a decidable fragment of first-order logic widely used for knowledge representation. By designing a hierarchy of five explanation conditions of increasing formal complexity, we systematically investigate whether state-of-the-art MLLMs can extract the discriminative visual features of an image class, formalise them as IF–THEN rules, and ultimately express them as DL. We evaluate explanation quality using an independent LLM-as-a-judge pipeline and a set of contributed XAI metrics.

Studying explanations within few-shot image classification provides a highly controlled setting. It encourages the model to ground its responses in newly introduced classes, helping to mitigate the influence of broad pre-training biases. Our primary goal is to assess how MLLMs explain the knowledge derived from the prompt. We believe that narrowing the scope to visual data under the widely adopted few-shot paradigm provides a controlled and reproducible setting for studying concept-based explainability, as visual features are directly verifiable against the query image.

Contributions. This paper makes the following empirical contributions:

- A systematic study of five explanation conditions of in-

[1]Universidad de Granada, Spain [2]Universidad de Buenos Aires, Argentina. Correspondence to: Carmen Quiles-Ramírez <carmenqr@ugr.es>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. *Copyright 2026 by the author(s).*

creasing formal rigour—from free-text NLE to DL axiom generation—applied to few-shot image classification under ICL.
- A reproducible experimental protocol over four image classification datasets and four MLLMs, using fixed few-shot samples for direct cross-model and cross-condition comparability.
- An LLM-as-a-judge evaluation framework with nine explanation-quality metrics covering textual groundedness, hallucination free, concept counting, comprehensibility, conciseness, specificity, local discriminativeness, instruction following and logical coherence.
- Empirical evidence that concept-based explanation capability varies substantially across current MLLMs, and that classification accuracy alone does not reflect the quality of the generated explanations.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 presents our methodology for assessing the concept-based explanation capabilities of MLLMs in visual classification tasks, formalizing the distinct types of explanations evaluated. Section 4 details the semantic explainability evaluation framework and introduces the judge model used to systematically assess these explanations. Section 5 describes the experimental design, outlining the selected models, datasets, and overall setup. Section 6 reports our experimental results, followed by a comprehensive discussion of these findings in Section 7. Finally, Section 8 concludes our study and highlights promising avenues for future work.

2. Related Work

Few-shot In-Context Learning (Few-shot ICL). It has emerged as a powerful paradigm for adapting Large Language Models to novel tasks using few-shot demonstrations, avoiding the need for weight updates (Brown et al., 2020; Dong et al., 2024). Extensive research has focused on optimizing this technique and understanding its vulnerabilities, such as prompt sensitivity and the impact of demonstration quality (Zhao et al., 2021; Lu et al., 2022). A critical ongoing debate surrounding ICL questions the extent to which models truly learn from the provided context versus merely retrieving prior knowledge from their pre-training phase (Min et al., 2022; Webson & Pavlick, 2022). Recently, this paradigm has been extended to the visual domain, where works such as Zhang et al. (2023), Zhou et al. (2024), and Sun et al. (2025) explore strategies to improve multimodal ICL performance. However, these studies evaluate success almost exclusively by classification accuracy. Crucially, in all ICL settings — including zero-, one-, and few-shot — the model's weights remain frozen; no gradient update occurs (Brown et al., 2020). Our work operates in this paradigm, investigating whether frozen MLLMs can produce concept-based explanations to justify their in-context behaviour.

Explainability of language models. Natural language explanations (NLEs) for model predictions have been studied as a route to interpretability (Wang et al., 2025). Barez et al. (2025) provide a critical analysis showing that CoT traces do not reliably reflect internal computation. Furthermore, previous authors warn of the safety risks associated with relying on CoT explanations, demonstrating that models frequently ignore their own reasoning traces during inference (Lanham et al., 2023) and often produce plausible but misleading rationales (Turpin et al., 2023) Probing studies have further shown that pre-trained language models specialise at different aspects of language understanding, with performance varying substantially across linguistic tasks (Gjinika et al., 2024). The broader landscape of XAI in the era of LLMs is surveyed in Polignano et al. (2024), which highlights the growing need for systematic evaluation frameworks. We take this body of work as a starting point and focus specifically on a gap it does not address: the capacity of MLLMs to *explain* image classification decisions under few-shot ICL—as opposed to merely producing accurate labels.

Structured and neurosymbolic explanations. Description Logics (Baader et al., 2009) provide a decidable, ontology-compatible formalism for expressing class definitions as verifiable axioms. Recent neurosymbolic work has explored symbol grounding (Castellano Ontiveros et al., 2025; Marconato et al., 2025) and logic-explained networks (Ciravegna et al., 2023), but these approaches require supervised training. Our contribution is complementary to these supervised approaches: rather than training models to produce formally verifiable axioms, we probe whether frozen MLLMs can spontaneously generate DL concept-based explanations at test time—without any fine-tuning—and evaluate their quality through an independent judge pipeline.

The cost of explicit reasoning. While explanations are desirable for transparency, recent work suggests that forcing models to articulate intermediate steps can sometimes degrade performance. Liu et al. (2025) demonstrated that Chain-of-Thought reasoning can reduce accuracy on tasks where human performance also drops when forced to think step-by-step, as it biases the model toward over-generalised rules rather than explicit context. Furthermore, the interplay between in-context examples and pre-trained knowledge remains complex, with models often struggling to suppress strong pre-training priors even when provided with explicit demonstrations (Zhang et al., 2024). This highlights a potential conflict between optimizing for predictive accuracy versus explanatory transparency under ICL.

3. Method: Assessing Concept-based Explanation Capabilities of MLLMs for Image Classification

In this section, we outline our experimental methodology. We first define the formal few-shot in-context learning setup for image classification, and subsequently detail the five explanation conditions designed to systematically probe the models' reasoning and explainability capabilities.

3.1. Problem Setup and methodology

We frame the task as N-way, few-shot (K-shot) image classification under ICL. Given a support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$ with N classes and K labelled examples per class, and a query image x_q, the model must predict $\hat{y}_q \in \mathcal{Y}$ using only the information provided in the context window—no fine-tuning or weight updates are allowed (Wang et al., 2020; Vinyals et al., 2016). We extend this standard setup by additionally requiring the model to produce a structured *explanation* of its prediction according to one of five explanation conditions (Section 3.2).

3.2. Evaluating five kinds of concept-based explanations with incremental semantic complexity.

We design five explanation types of increasing formal complexity to evaluate the explanation capabilities of multimodal LLMs. All conditions requested share a common system prompt with three constraints enforced via structured output formatting: (i) predictions must be grounded exclusively in visually observable evidence from the provided examples; (ii) external world knowledge, hidden assumptions, and speculative attributes are explicitly prohibited; (iii) the final class label must appear inside a `<response>` XML tag, whose content must be copied verbatim from the provided options list, to enable deterministic parsing. The common theme across all five conditions is that they are centered around concept-based explainability (Poeta et al., 2025): unlike saliency maps or gradient-based methods, all explanation types rely on human-understandable visual concepts. The exact system prompts and formatting constraints used across all explanation conditions are detailed in Appendix A.

E1 — Classification with no explanation (Baseline). We request the model to output only the final class label inside `<response>`. No explanation nor intermediate output is requested. This establishes the accuracy baseline against which all other explanation conditions are compared.

E2 — Natural Language Explanation (NLE). We ask the model to produce a concise natural language explanation inside an `<explanation>` tag, grounded only in observable visual evidence, briefly justifying why that evidence supports the selected class over the alternatives. This corresponds to standard CoT-style prompting.

E3 — Features-based explanation. We prompt the model to identify the minimal sufficient set of critical, concrete, observable visual features needed to support the classification, expressed as short noun phrases in bullet-point form inside a `<features>` tag.

E4 — Feature-value pairs explanation. We ask the model to list the observable features of the query image in a `<features>` tag, then to construct a knowledge base of IF–THEN rules grounded in visual patterns from the labelled examples in a `<kb>` tag, and finally we ask it to identify which rule is best satisfied by the query image in a `<rule_check>` tag—referring only to the features already listed and introducing no new evidence. This tests whether the model can perform structured logical reasoning from few-shot examples.

E5 — DL Axioms explanation. Description Logics (DL) (Baader et al., 2009) is a decidable fragment of first-order logic designed for knowledge representation. A *knowledge base* $\mathcal{KB} = (\mathcal{T}, \mathcal{A})$ consists of a TBox $\mathcal{T}$ (terminological axioms defining concept hierarchies and roles) and an ABox $\mathcal{A}$ (assertional facts about individuals). DL reasoning supports inference operations including subsumption, consistency checking, and instantiation, making axiom-based explanations formally verifiable.

We ask the model to present the classification evidence using ontological DL statements (Baader et al., 2009): a `<tbox>` tag defines formal class conditions as necessary, sufficient, or necessary-and-sufficient axioms using `hasVisualFeature` roles derived from the labelled examples; an `<abox>` tag records the specific property assertions observed in the query image; and a `<dl_explanation>` tag explains how the ABox assertions satisfy the TBox axioms to logically derive the predicted class. This is the most constrained and formally rigorous explanation condition.

4. Semantic explainability evaluation

We use accuracy to measure the quality of classified instances. Due to the fact that the classification accuracy metric cannot capture the *quality* of the generated explanations, we use an independent *LLM-as-a-judge* pipeline (Figure 1). Our approach uses a single answer that scores 9 different XAI metrics, as shown in Table 1. This enables scalable benchmarking and fast iterations at the scale of the full experiment; the alignment of automated scores with human judgement is validated in Appendix H. This work proposes a standardized way of evaluating the explanations.

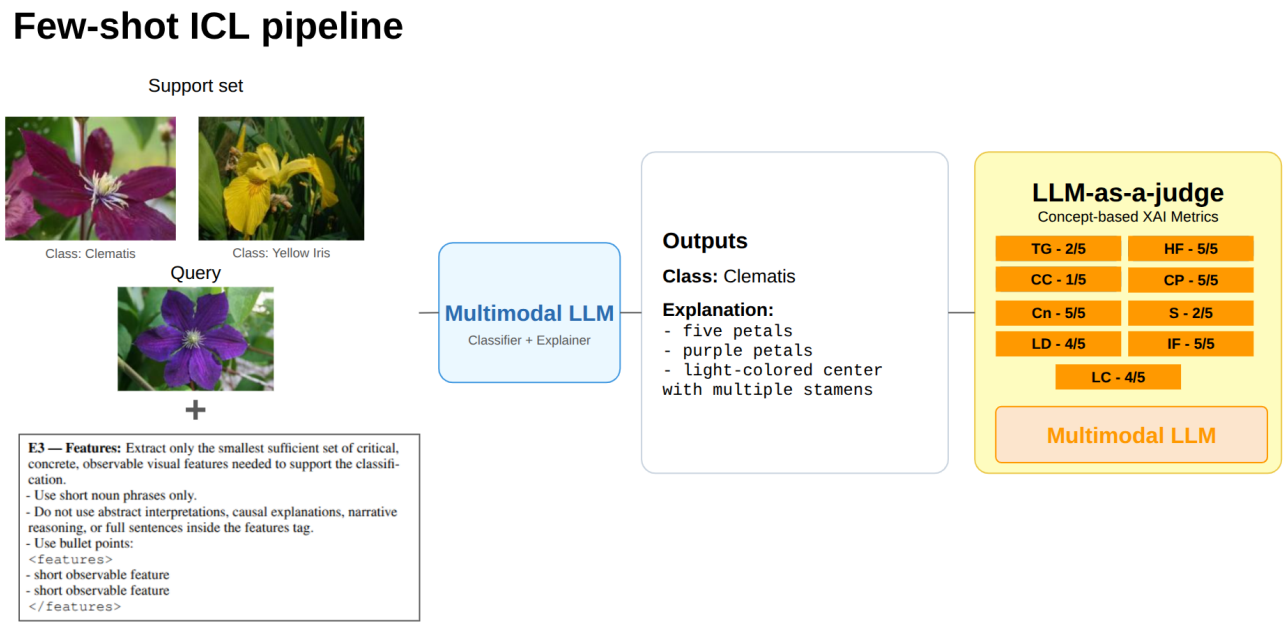

Figure 1. Overview of the In-Context Learning *Explain-Classify and Evaluate* pipeline. The Multimodal LLM acts as both classifier and explainer, while a separate LLM-as-a-judge evaluates nine concept-based XAI metrics (1–5 ↑): see Table 1. The example shown corresponds to E3 (feature-based explanation), where the model is prompted to identify the minimal sufficient set of critical, concrete, and observable visual features needed to support the classification. The judge, using a different MLLM and prompt, scores the generated explanation across the same nine evaluation dimensions.

Table 1. Explanation quality metrics we define to be scored by the LLM-as-a-judge (1 = worst, 5 = best). The complete scoring rubrics for each proposed metric are detailed in Appendix C.

Concept-based XAI Metric	Description
Textual Groundedness (TG)	All salient image concepts are mentioned in the explanation.
Hallucination Free (HF)	Every claim is visually verifiable in the image.
Concept Counting (CC)	Concept counts match the image exactly.
Comprehensibility (CP)	Explanation readable, accessible to non-expert users.
Conciseness (Cn)	Conveys only necessary information without redundancy.
Specificity (S)	Relies on precise, non-generic details of the sample.
Local Discriminativeness (LD)	Highlights features distinguishing predicted class from others.
Instruction Following (IF)	Complies with all explicit format and structure constraints.
Logical Coherence (LC)	Sentences build into a smooth, valid deduction.

For each query trial the judge receives: (i) the query image, (ii) the set of candidate class labels, (iii) the classifier's full output including the predicted class label, (iv) a description of the explanation condition being evaluated (E1–E5), and (v) the scoring rubric.

The judge operates in a separate pipeline only receiving the detailed information and in a zero-shot setting: it does not receive the few-shot support images seen by the classifier model, due to compute constraints. Evaluation is grounded exclusively in the query image as the visual source of truth. For the one metric that relies on the judge understanding attributes about the set of candidate classes (e.g. Local Discriminativeness), we expect the judge model's prior knowledge to be sufficient within the selected datasets and examples.

To optimise this expectation while remaining below compute budget limits, we choose `gpt-5-thinking-mini` (OpenAI, 2025) as the judge model, a medium-size state-of-the-art reasoning model, accessed via the OpenRouter API with `reasoning_effort` set to `medium`. To assess whether judge access to support images affects the key Local Discriminativeness metric, we conduct a controlled ablation study using a second judge model; see Appendix G.

To further validate the judge pipeline, we conducted an independent human evaluation in which three researchers manu-

ally scored a stratified sample of 192 trials using the same 1–5 rubrics (see Appendix H). The judge demonstrates strong alignment with human judgements: Spearman $\rho > 0.60$ against the human mean across 6 of 9 metrics, including Local Discriminativeness ($\rho = 0.656$), the metric most central to our findings, confirming the validity of the automated evaluation pipeline.

5. Experiments

In this section, we present our experimental setup.

Datasets and Models

Following the standard VLM few-shot benchmark established by Zhou et al. (2022) and widely adopted in the VLM literature, we used four datasets that cover a diverse range of visual recognition tasks: CIFAR-10 (Krizhevsky, 2009) (10 generic object categories, low-resolution images), DTD (Cimpoi et al., 2014) (47 texture classes), Oxford Flowers 102 (Nilsback & Zisserman, 2008) (102 flower species), and Oxford-IIIT Pets (Parkhi et al., 2012) (37 fine-grained pet breed categories). Beyond their role as a standard benchmark, these datasets are particularly well-suited for evaluating the quality of textual explanations: Oxford-IIIT Pets and Flowers 102 involve classes that are highly visually describable, providing rich discriminative features for the model to articulate; DTD requires describing abstract perceptual properties such as texture and material, testing whether models can express non-object concepts; and CIFAR-10 serves as a coarser-grained contrast baseline.

We evaluated four state-of-the-art MLLMs selected to represent a diverse range of model families, scales, and access types (Table 2). All models support native multi-image input, a requirement for the N-way K-shot context format used in our experiments, and are instruction-tuned (Wei et al., 2022; Ouyang et al., 2022). All are accessed via the OpenRouter API at temperature T=0. Model selection was further guided by recent systematic benchmarks of multimodal foundation models (Ramachandran et al., 2025) and multimodal ICL tasks (Huang et al., 2025)

Table 2. MLLMs evaluated in this study. † 17B active (MoE). ‡ 4B active (MoE). CTX = context window. See Appendix B for details.

Model	Type	Params	CTX
Gemini 2.5 Flash	Proprietary	—	1,048K
Gemma 4 26B	Open (MoE)	26B‡	262K
Qwen3 VL 8B	Open	8.77B	131K
LLaMA 4 Scout	Open (MoE)	109B†	328K

Experimental Setup

We generate support sets following the standard episodic few-shot evaluation protocol (Vinyals et al., 2016; Snell et al., 2017). Each support set of N classes with K labelled images per class, and a single query image (Q=1).

Setting Q=1 serves two purposes. First, it ensures that each trial constitutes a fully independent sample—with its own support set—satisfying the independence assumption required by the non-parametric statistical tests used in our analysis (McNemar, Wilcoxon signed-rank, Friedman). Increasing Q within a fixed support set would yield pseudo-replicates conditioned on the same context, artificially reducing variance estimates and inflating statistical power. Second, since our primary interest lies in the explanations generated under each explanation condition, Q=1 ensures that each explanation is produced independently, without influence from prior model outputs within the same support set—critical for explanation conditions that elicit structured reasoning (E4, E5), where accumulated context could introduce consistency biases unrelated to the condition.

We evaluated all combinations of $N \in \{2, 3, 4\}$ and $K \in \{1, 5\}$, following the canonical 1-shot and 5-shot evaluation protocol established in the few-shot learning literature (Vinyals et al., 2016; Snell et al., 2017). The 1-shot setting tests whether a single labelled example per class is sufficient for the model to both classify and explain; the 5-shot setting tests whether additional context improves accuracy and, crucially, explanation quality. The range $N \in \{2, 3, 4\}$ spans binary discrimination (N=2) through increasingly fine-grained multi-class settings (N=4), allowing us to assess how explanation quality scales with classification difficulty. The number of independent support sets repetitions per configuration is chosen such that $\text{Reps} \times N = 12$ for all values of N, ensuring that each difficulty level contributes an equal number of class presentations to the aggregate evaluation. Without this constraint, configurations with smaller N would yield more trials in total, inflating performance estimates for easier settings and confounding comparisons across difficulty levels. All support sets are generated once with a fixed random seed (42) and reused across all models and explanation conditions. Table 3 summarises the full configuration grid.

Table 3. Few-shot ICL configurations evaluated. Reps balanced so that Reps $\times$ N = 12. Total: 26 $\times$ 4 models $\times$ 4 datasets $\times$ 5 conditions = 2,080 runs.

Trial	N	K	Q	Runs	Total runs / Trial
1	2	1	1	6	480
2	2	5	1	6	480
3	3	1	1	4	320
4	3	5	1	4	320
5	4	1	1	3	240
6	4	5	1	3	240
Total	—	—	1	26	2,080

6. Results

Classification Accuracy. Tables 4 and 5 report mean accuracy aggregated over all models and datasets respectively. Overall accuracy across all conditions and models reaches 92.6%. Gemini 2.5 Flash is the top-performing model in every condition (Table 4), and Flowers is the easiest dataset, with all models achieving near-perfect accuracy regardless of condition (Table 5). Critically, accuracy decreases monotonically as explanation complexity increases: from 93.8% at E1 (Classification) to 90.1% at E5 (DL Axioms), confirming that demanding richer explanations imposes a cost on predictive performance. The accuracy gap between conditions is most pronounced for Qwen3 VL 8B, which drops from 95.1% at E1 to 83.0% at E5, while Gemini 2.5 Flash remains above 95% across all conditions. Dataset difficulty also varies: DTD shows the steepest decline across conditions (from 85.8% at E1 to 82.6% at E5), consistent with its abstract textural content being harder to verbalise formally. Detailed per-model and per-configuration breakdowns are provided in Appendix D.

Table 4. Mean accuracy (%) by explanation condition and model, aggregated over 4 datasets and 6 few-shot configurations ($N \in \{2, 3, 4\}$, $K \in \{1, 5\}$): $4 \times 26 = 104$ observations per cell. †Classification-only baseline. Values: mean (±SE). Best per column in bold.

XAI Condition	Gem. 2.5F	Gem. 4	Qwen	LLaMA
E1 - Classification†	96.9 (1.0)	**94.4** (1.3)	**95.1** (1.3)	88.5 (1.9)
E2 - NLE	**97.2** (1.0)	94.1 (1.4)	92.7 (1.5)	**90.3** (1.7)
E3 - Features	96.9 (1.0)	93.1 (1.5)	93.8 (1.4)	88.5 (1.9)
E4 - Feature-value pairs	95.8 (1.2)	**94.4** (1.3)	92.4 (1.6)	86.5 (2.0)
E5 - DL Axioms	96.2 (1.1)	92.4 (1.6)	83.0 (2.2)	88.9 (1.9)

Table 5. Mean accuracy (%) by explanation condition and dataset, aggregated over 4 models and 6 few-shot configurations: $4 \times 26 = 104$ observations per cell. †Classification-only baseline. Values: mean (±SE). Best per column in bold.

XAI Condition	Flowers	Pets	CIFAR	DTD
E1 - Classification†	**100.0** (0.0)	**95.1** (1.3)	94.1 (1.4)	**85.8** (2.1)
E2 - NLE	**100.0** (0.0)	93.8 (1.4)	**94.8** (1.3)	**85.8** (2.1)
E3 - Features	**100.0** (0.0)	93.4 (1.5)	**94.8** (1.3)	84.0 (2.2)
E4 - Feature-value pairs	99.7 (0.3)	93.1 (1.5)	92.4 (1.6)	84.0 (2.2)
E5 - DL Axioms	96.5 (1.1)	91.0 (1.7)	90.3 (1.7)	82.6 (2.2)

Effect of K and N on explanation quality. Increasing support images from K=1 to K=5 improves accuracy by 7.0 pp ($p = 2.0 \times 10^{-13}$), with the largest explanation quality gain concentrated in Local Discriminativeness (Δ=+0.26, $p_{\text{Bonf}} = 1.5 \times 10^{-6}$). Conversely, increasing N degrades accuracy monotonically and, among all metrics, only Local Discriminativeness falls significantly (3.86 at N=2 to 3.40 at N=4; $p_{\text{Bonf}} = 1.9 \times 10^{-10}$), confirming that adding classes stresses discriminative reasoning specifically. Results are detailed in Appendix F (Figure 6).

Explanation Quality across explanation conditions. Table 6 reports mean LLM-as-a-judge scores across the nine quality metrics. DL Axioms (E5) scores substantially lower than other conditions on Textual Groundedness (2.31), Specificity (2.85), Local Discriminativeness (3.10), Instruction Following (3.05), and Logical Coherence (2.97), while performing comparably on Hallucination Free (4.40) and Conciseness (4.94). This pattern is clearly visible in Figure 2: DL Axioms produces a markedly smaller polygon than the other conditions, confirming that models produce outputs that are syntactically compliant (when they succeed) but fail to anchor explanations in discriminative visual evidence. Feature-based and Feature-value pairs conditions achieve the highest scores on most metrics, confirming that these intermediate formats are best handled by current MLLMs.

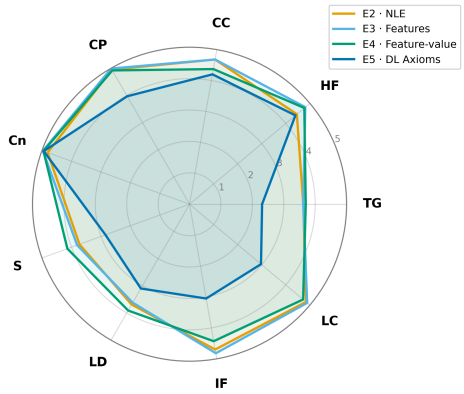

Figure 2. LLM-as-a-judge scores across nine explanation quality metrics by condition (E2–E5), averaged over all models and datasets. DL Axioms (E5) shows a markedly smaller polygon, particularly on Textual Groundedness, Specificity, Instruction Following, and Logical Coherence. Metric abbreviations as in Table 1.

Figure 3 shows the Spearman correlation between each judge metric and classification accuracy, with Bonferroni correction over the 36 simultaneous tests (9 metrics $\times$ 4 conditions). Local Discriminativeness (LD) is the only metric that is both consistently and strongly associated with accuracy across all four conditions (ρ = 0.30–0.35 for E2–E4, ρ = 0.21 for E5; all $p_{\text{Bonf}} < 10^{-10}$). Two further metrics—Textual Groundedness (TG) and Specificity (S)—reach significance in every condition but with substantially smaller magnitudes ($\rho \approx$ 0.13–0.18), confirming LD as the dominant predictor. Logical Coherence (LC) is the strongest predictor for E2 ($\rho = 0.36$, $p_{\text{Bonf}} = 3 \times 10^{-35}$) and remains significant for E3 and E4, but fails to reach significance for E5 DL Axioms ($\rho = 0.04$, $p_{\text{Bonf}} = 1.0$). More broadly, under E5 three dimensions—Logical Coherence (LC), Concept Counting (CC), and Conciseness (Cn)—are

Table 6. Mean LLM-as-a-judge semantic explainability scores (1–5 ↑) by explanation condition, aggregated over 4 models, 4 datasets, and 6 few-shot configurations: $4 \times 26 = 104$ observations per cell. Values: mean (±SE). **Bold** = best per column.

Condition	TG	HF	CC	CP	Cn	S	LD	IF	LC
NLE	3.62 (.02)	4.46 (.03)	4.68 (.03)	4.95 (.01)	4.81 (.01)	3.73 (.02)	3.69 (.04)	4.70 (.02)	4.84 (.01)
Features	3.62 (.03)	**4.81** (.02)	**4.68** (.03)	**4.99** (.00)	**4.97** (.01)	3.81 (.03)	3.62 (.04)	**4.82** (.01)	**4.89** (.01)
Feature-value pairs	**3.70** (.03)	4.77 (.02)	4.37 (.04)	4.92 (.01)	4.95 (.01)	**4.14** (.02)	**3.91** (.04)	4.43 (.03)	4.72 (.02)
DL Axioms	2.31 (.04)	4.40 (.03)	4.20 (.05)	3.97 (.02)	4.94 (.01)	2.85 (.04)	3.10 (.05)	3.05 (.03)	2.97 (.04)

indistinguishable from zero after correction, and the largest significant ρ for E5 is only 0.21 (LD), versus 0.35–0.36 for E2–E3. This pattern is consistent with the degraded explanation quality observed for E5: when the explanation channel is unreliable, most of its dimensions stop being diagnostic of the final classification decision.

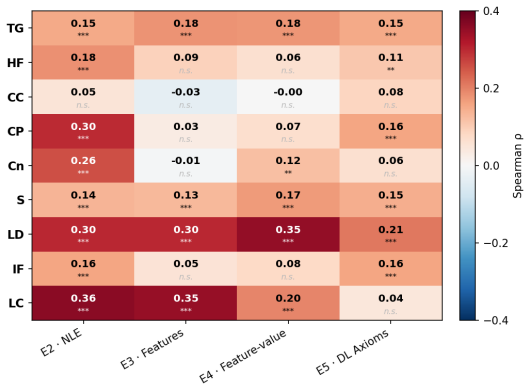

Figure 3. Spearman ρ between LLM-as-a-judge XAI metrics and classification accuracy across explanation conditions E2–E5. Significance markers indicate Bonferroni-corrected p-values over $9 \times 4 = 36$ tests: * $p_{\text{Bonf}} < 0.05$, ** $p_{\text{Bonf}} < 0.01$, *** $p_{\text{Bonf}} < 0.001$; *n.s.* not significant. Metric abbreviations as in Table 1.

7. Discussion

Explaining is Harder Than Predicting Alone. A central finding of our study is that forcing an MLLM to explain its decision does not improve classification accuracy; in fact, we observe the exact opposite. Accuracy drops monotonically as the complexity of the requested explanation increases, reaching its peak when no explanation is requested (E1). This suggests that deciding and explaining simultaneously is fundamentally more complex than deciding alone, contradicting the hypothesis that structured reasoning universally acts as a form of test-time regularisation.

Accuracy does not imply explanation quality. A key observation from our preliminary experiments is that a model can produce the correct class label while generating a factually incorrect explanation—for instance, correctly identifying a flower species while misreporting the number of petals visible in the image. This decoupling motivates our two-track evaluation (accuracy + judge scores) and underscores the importance of assessing explanation quality independently of predictive performance.

Discriminativeness as a proxy for correct decisions. Despite the drop in accuracy for highly concept-based explanations such as DL Axioms, Local Discriminativeness (LD) emerges as the metric most strongly associated with correct predictions. When models succeed, they tend to justify their decisions using visual features that effectively separate the target class from alternatives, suggesting that access to discriminative evidence is closely tied to correctness. Model behaviour varies substantially across architectures and prompting conditions. For instance, Qwen3 VL 8B collapses on Textual Groundedness (TG) under E5 (TG = 1.06 in Table 18), whereas Gemini 2.5 Flash maintains comparatively stable performance across metrics and conditions, indicating that architectural and training differences significantly affect concept-based explanation capability beyond classification accuracy. An exception is LLaMA 4 Scout on DTD, where DL Axioms achieves the highest accuracy (86.1%), surpassing the classification-only baseline (76.4%), suggesting that structured axiom-style prompting may improve grounding for abstract texture classes. Overall, these results admit a mechanistic interpretation. Across settings, improving access to discriminative information—e.g., by increasing the number of support images (K=5)—consistently enhances both accuracy and LD, as models become better at isolating features that distinguish the correct class. In contrast, increasing the number of candidate classes (N) systematically degrades both accuracy and LD, while leaving other explanation dimensions largely stable, indicating that the effect is specific to discriminative reasoning. Together, this suggests that LD is a strong proxy for whether the model has correctly identified class-distinguishing evidence (Figure 4): when such evidence is successfully extracted, the model not only predicts correctly but can also verbalise the underlying rationale.

The Instruction-Tuning Hypothesis. We hypothesise that the capability gap between predicting and formally explaining stems from how these models are instruction-tuned. While current MLLMs excel at concrete tasks like classification, they are not explicitly optimised to generate formal logical representations in tandem with classification. Asking a model to perform both tasks simultaneously within a single

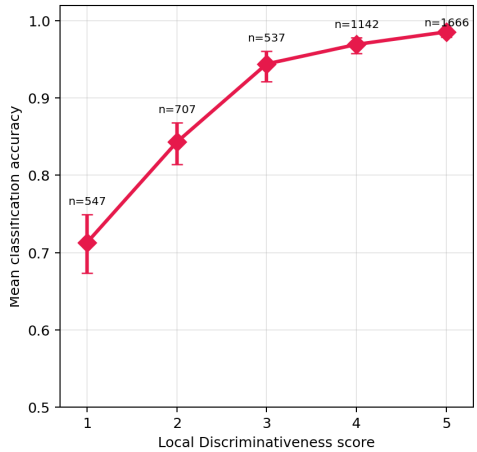

Figure 4. Mean classification accuracy as a function of Local Discriminativeness (LD) score, pooled across all models, datasets, and explanation conditions (E2–E5).

prompt overcomplicates generation, a phenomenon aligned with recent work showing that forcing explicit reasoning steps can sometimes degrade performance by biasing models toward generalised rules rather than contextual evidence (Liu et al., 2025). This is further evidenced by the contrasting pattern observed under E5: while Hallucination Free (HF) and Conciseness (Cn) scores remain high (4.40 and 4.94 respectively), Textual Groundedness (TG), Local Discriminativeness (LD), and Logical Coherence (LC) collapse — suggesting that models learn to reproduce the *syntactic structure* of DL reasoning without filling it with discriminative visual content. This mirrors a broader question central to compositional AI: whether models truly compose symbolic representations or merely imitate their surface form (Huang et al., 2025).

Limitations. Our evaluation framework has several constraints. First, our results are limited to four MLLMs and four datasets. Additionally, due to the independent nature of episodic ICL trials, models do not accumulate knowledge iteratively across queries. This setup makes it fundamentally difficult to assess true "zero knowledge" prior to prompting, or to perform ablation studies that guarantee a complete lack of data leakage from the models' pre-training phase. Consequently, separating genuine in-context reasoning from pre-training memory retrieval remains an inherent limitation of this paradigm — further compounded by the fact that LLMs cannot fully suppress pre-trained knowledge when processing in-context demonstrations, with smaller models tending to rely more heavily on label semantics than on in-context mappings (Zhang et al., 2024), and overcoming strong prior biases remaining an open challenge (Dong et al., 2024). To mitigate this, our prompts explicitly instruct the model not to rely on external world knowledge or speculative attributes (see Appendix A), grounding classification and explanations in the visual evidence of the support images. A further limitation concerns classes whose discriminative features cannot be described in absolute terms — that is, features that only become discriminative *relative* to the other classes present in the support set. For instance, subtle breed-specific coat markings in Oxford-IIIT Pets may only be identifiable as class-defining when compared against the other breeds in the trial, not in isolation. In such cases, Local Discriminativeness (LD) — which is evaluated by the judge without access to the support set — could in principle underestimate explanation quality for classes whose discriminative features are only articulable relative to the specific episode context. However, our ablation study (Appendix G) shows empirically that LD rankings are robust to support-image access ($\rho = 0.627$, mean score shift $= -0.049$ points), suggesting that the judge's prior knowledge of the candidate classes is sufficient for reliable LD assessment within the benchmarks used in this study. Addressing contrastive-dependent features more generally would require *Global Discriminativeness* (GD), which evaluates whether explanations highlight features that are truly pathognomonic at the dataset level (see Appendix C); however, GD requires knowledge beyond a single trial and is therefore left for future work.

8. Conclusion and Future Work

In this paper, we presented a systematic empirical study evaluating whether frozen MLLMs can produce structured, semantically rich, and verifiable explanations for their in-context image classification decisions. Our findings demonstrate that explaining is genuinely harder than predicting alone. Contrary to the hypothesis that forcing explicit reasoning acts as a test-time regularisation that improves accuracy, we observe a monotonic performance drop: classification accuracy degrades from 93.8% to 90.1% as the required explanations become more formally rigorous (from simple labels to Description Logic ontological axioms).

Furthermore, our LLM-as-a-judge evaluation reveals that classification accuracy does not imply explanation quality. While current models struggle significantly to articulate formal logic structures, their ability to identify and verbalise class-discriminative visual features (Local Discriminativeness) is the strongest predictor of a correct classification.

Future work should explore methods to evaluate contrastive-dependent visual features, exploit the machine-verifiable nature of DL axioms — generated under E5 — as structured inputs for a secondary safe classifier, and investigate whether instruction-tuning can be explicitly optimised for formal explainability without compromising predictive performance.

Ultimately, we argue that just as *current models* are opti-

mized via Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) to align with human intent, future frameworks *should integrate concept-based explanations* directly into the optimization objective. This approach creates a crucial bidirectional link between semantics—the 'why' for human understanding—and formality—the mathematical 'proof' for system integrity. By leveraging these explanations to reach a formal state of explainability, we move beyond simple task performance toward a new generation of models that are inherently and provably transparent.

Code, prompts, and evaluation pipeline are publicly available at `https://github.com/letyrodridc/research-llms-in-context-learning-explained`.

Impact Statement This paper advances explainable AI and ICL by contributing tools and evaluation frameworks for auditing AI decision-making. We foresee no immediate negative societal consequences beyond those generally associated with advancing large language model capabilities.

Acknowledgements

This work was supported by the TSI-100927-2023-1 project under the Recovery, Transformation and Resilience Plan funded by the European Union NextGenerationEU through the Ministry for Digital Transformation and the Civil Service, and grants PID2023-149128NBI00 and PID2023-150070NB-I00 funded by MICIU/AEI /10.13039/501100011033 and by ERDF, EU.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. doi: 10.1016/j.inffus.2023.101805.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2nd edition, 2009. ISBN 978-0-521-87023-0.

Babaei Giglou, H., D'Souza, J., Karras, O., and Auer, S. Ontoaligner: A comprehensive modular and robust python toolkit for ontology alignment. In *European Semantic Web Conference*, pp. 174–191. Springer, 2025.

Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., et al. Chain-of-Thought is Not Explainability, 2025.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Castellano Ontiveros, R., Giannini, F., Gori, M., Marra, G., and Diligenti, M. Grounding methods for neural-symbolic AI. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pp. 4806–4814, 2025.

Chiang, C.-H. and Lee, H.-y. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631. Association for Computational Linguistics, 2023.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Liò, P., Maggini, M., and Melacci, S. Logic explained networks. *Artificial Intelligence*, 314:103822, 2023. doi: 10.1016/j.artint.2022.103822.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. A Survey on In-Context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.64.

Gemma Team, Google DeepMind. Gemma 4 model card. technical report. Technical report, Google DeepMind, 2026. URL `https://ai.google.dev/gemma/docs/core/model_card_4`.

Gjinika, E., Arici, N., Putelli, L., Gerevini, A. E., and Serina, I. An analysis on how pre-trained language models learn

different aspects. In *Proceedings of the 5th Italian Workshop on Explainable Artificial Intelligence, co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024)*, Bolzano, Italy, November 2024. CEUR-WS.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Huang, C. et al. Mimicking or reasoning? Rethinking multi-modal in-context learning in vision-language models, 2025.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., et al. Measuring faithfulness in chain-of-thought reasoning, 2023.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., and Griffiths, T. L. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *The Thirteenth International Conference on Learning Representations*, 2025.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4086–4102. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.305.

Marconato, E., Bortolotti, S., van Krieken, E., Morettin, P., Umili, E., Vergari, A., Tsamoura, E., Passerini, A., and Teso, S. Symbol grounding in neuro-symbolic AI: A gentle introduction to reasoning shortcuts, 2025.

Mathur, N., Baldwin, T., and Cohn, T. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4984–4997. Association for Computational Linguistics, 2020.

Meta AI. Llama 4 Model Card and Technical Specifications, 2025. URL `https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md`. Accessed: 2026-04-30.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

OpenAI. OpenAI GPT-5 System Card, 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., and Baralis, E. Concept-based explainable artificial intelligence: A survey. *ACM Computing Surveys*, 2025. doi: 10.1145/3774643.

Polignano, M., Musto, C., Pellungrini, R., Purificato, E., Semeraro, G., Setzu, M., et al. XAI.it 2024: An overview on the future of AI in the era of large language models. In *Proceedings of the Italian Workshop on Explainable Artificial Intelligence (XAI.it 2024)*. CEUR-WS, 2024.

Qwen Team and Alibaba Cloud. Qwen3-vl technical report, 2025.

Ramachandran, R., Garjani, A., Bachmann, R., Atanov, A., Kar, O. F., and Zamir, A. How well does GPT-4o understand vision? Evaluating multimodal foundation models on standard computer vision tasks, 2025.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sun, Y., Chen, Q., Wang, J., Wang, J., and Li, Z. Exploring effective factors for improving visual In-Context Learning. *IEEE Transactions on Image Processing*, 34, 2025. doi: 10.1109/TIP.2025.3554410.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020. doi: 10.1145/3386252.

Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., and Fei, H. Multimodal chain-of-thought reasoning: A comprehensive survey, 2025.

Webson, A. and Pavlick, E. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2300–2344. Association for Computational Linguistics, 2022.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations*, 2022.

Yang, L., Xu, Y., Yuan, C., Liu, W., Li, B., and Hu, W. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9499–9508, 2022.

Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., and Chen, D. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhang, Y., Zhou, K., and Liu, Z. What makes good examples for visual in-context learning?, 2023.

Zhang, Y., Feng, F., and He, X. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14365–14378. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.795.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 12697–12706, 2021.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

Zhou, Y., Li, X., Wang, Q., and Shen, J. Visual in-context learning for large vision-language models, 2024.

A. Prompt Templates

This appendix provides the exact prompt templates used in our experiments. Variables populated at runtime are enclosed in curly braces, such as {CONDITION_INSTRUCTION}.

A.1. Classification Prompts

All MLLM classifiers were evaluated using a shared system prompt, into which specific explanation formatting constraints were injected based on the evaluated condition.

SHARED_SYSTEM_PROMPT

 # TASK

1. Analyze the provided labeled examples to identify the observable visual features that distinguish each class.
2. Examine the Query Image and compare it strictly against the provided examples.
3. Determine which label from the provided options best matches the Query Image.

 # INSTRUCTIONS

- Base your decision only on observable visual evidence in the examples and the Query Image.
- Use the examples to infer discriminative visual patterns for each class.
- Do not use external world knowledge, hidden assumptions, or speculative attributes.
- Use only labels from the provided final options list.
- Choose exactly one final label.
- Output only the requested XML tags.
- Do not output any text outside the XML tags.
- Do not output any additional XML tags other than those explicitly defined in {CONDITION_INSTRUCTION}.

{CONDITION_INSTRUCTION}

<response>final_class</response>

The content of <response> must be exactly one label copied verbatim from the provided options list.

The {CONDITION_INSTRUCTION} variable varies according to the explanation setting as follows:

CONDITION INSTRUCTIONS BY EXPLANATION TYPE

E1 —Classification with no explanation (Baseline): Provide only the final class label inside the response tag.

E2 — NLE: Write a concise natural language explanation grounded only in observable visual evidence.
<explanation>
Brief explanation of why the visual evidence supports the selected class.
</explanation>

E3 — Features: Extract only the smallest sufficient set of critical, concrete, observable visual features needed to support the classification.
- Use short noun phrases only.
- Do not use abstract interpretations, causal explanations, narrative reasoning, or full sentences inside the features tag.
- Use bullet points:
<features>
- short observable feature
- short observable feature
</features>

E4 — Feature-value pairs: First list the critical observable features present only in the Query Image. Then formulate a knowledge base (KB) with a small set of IF-THEN rules grounded in observable patterns from the labeled examples in the support images and the Query Image, for all classes. Try to unify the conditions so that there is only one rule per class. Rules must refer only to visible features. Then check which rule(s) are best satisfied by the Query Image.
<features>
- observable feature: feature value
</features>
<kb>
- IF [observable visual features = feature value AND/OR observable visual features = feature value] THEN [class1]
- IF [observable visual features = feature value] THEN [class2]
</kb>
<rule_check>
- State which rule(s) from the kb fires for the Query Image and briefly explain why, referring only to the features listed above. Do not discuss other rules or introduce new features.
</rule_check>

E5 — DL Axioms: Represent the classification evidence using minimal and concise Description Logic statements.
 # WHAT TO WRITE
- In <tbox>, define the formal class conditions (Necessary, Sufficient, or Necessary & Sufficient) using Description Logic (DL) axioms in Web Ontology Language (OWL 2) derived from the labelled examples to describe which visual features are associated with each class.
- In <abox>, state only the specific property assertions in DL OWL 2 observed in the Query Image to record which visual features are present.
- In <dl_explanation>, list only the exact TBox rules triggered by the ABox assertions to classify the Query Image. Do not use natural language—include only the fired TBox rules.

 # HOW TO INTERPRET THE STATEMENTS
- [Class1] $\sqsubseteq$ hasVisualFeature([F1]) means: images of this class tend to have this visual feature in the provided examples (necessary condition).
- hasVisualFeature([F1]) $\sqsubseteq$ [Class1] means: observing this visual feature is enough to classify the image as from this class in the provided examples (Sufficient condition).
- [Class1] $\equiv$ hasVisualFeature([F1]) $\sqcap$ hasVisualFeature([F2]) means: this combination of features defines this class in the provided examples (Necessary and sufficient condition).

 # RULES

- Use only observable visual features. Do not use hidden properties, external knowledge, or speculation. Treat all axioms as visually grounded in the provided examples. Keep the output concise.
<tbox>
- [Class1] ⊑ hasVisualFeature([F1])
- hasVisualFeature([F2]) ⊑ [Class2]
- [Class3] ≡ hasVisualFeature([F3]) ⊓ hasVisualFeature([F4])
</tbox>
<abox>
- hasVisualFeature(Query, [F3])
- hasVisualFeature(Query, [F4])
</abox>
<dl_explanation>
- Rule(s) fired: [Class3] ≡ hasVisualFeature([F3]) ⊓ hasVisualFeature([F4])
</dl_explanation>

A.2. LLM-as-a-judge Prompts

JUDGE_PROMPT

You are a fair LLM judge. Your task is providing clear, objective feedback based on specific criteria to evaluate the quality of an explanation generated by an independent LLM query regarding a previous image classification decision. Please, ensure that each assessment reflects the absolute dimensions to score.

You will receive: 1. The Query image. 2. The set of candidate class labels. 3. The candidate model output (explanation). 4. The predicted class. 5. A description of the explanation condition that specifies the type of explanation the model was asked to produce and how its output should be interpreted.

To give feedback, rate the explanation on each of all given dimensions writing a score that is an integer between 1 and 5.

 # IMPORTANT EVALUATION PRINCIPLES
- Evaluate the explanation itself, not whether the response is correct or incorrect.
- Evaluate whether the explanation supports the predicted class, regardless of whether the prediction is ultimately correct with respect to external ground truth.
- Judge only what is explicitly present in the model output.
- Use the Query image as the source of truth for whether mentioned features are visible.

 # CONDITION DESCRIPTION
{CONDITION_DESCRIPTION}

 # DIMENSIONS TO SCORE
(1. Textual Groundedness, 2. Hallucination free, 3. Concept counting, 4. Comprehensibility, 5. Conciseness, 6. Specificity, 7. Discriminativeness, 8. Instruction following, 9. Logical coherence. The complete 1–5 scoring rubrics for each metric provided to the judge are detailed in Appendix C.)

 # OUTPUT INSTRUCTIONS
Return your evaluation using exactly the following XML structure:
<evaluation>
<textual_groundedness>1-5</textual_groundedness>
...
<logical_coherence>1-5</logical_coherence>
</evaluation>
Please, do not generate any other opening, closing, and explanations.

JUDGE CONDITION DESCRIPTIONS

The {CONDITION_DESCRIPTION} variable adapts to the evaluated explanation type:

E2 — NLE: The model was asked to classify the Query image and provide a concise natural-language explanation. Expected format: Only <explanation> and <response> tags. The explanation should be grounded only in observable visual evidence and written in free-form natural language.

E3 — Features: The model was asked to classify the Query image and provide a minimal list of critical observable features supporting the classification. Expected format: Only <features> and <response> tags. Features should be short noun phrases formatted as bullet points, without abstract interpretations or full sentences.

E4 — Feature-value pairs: The model was asked to classify the Query image and provide a rule-based explanation. Expected format: <features>, <kb>, <rule_check>, and <response> tags. Rules in <kb> must be IF-THEN grounded in observable patterns covering all classes. <rule_check> should state which rule fires for the Query image referring only to the listed features.

E5 — DL Axioms: The model was asked to classify the Query image and provide a Description Logic (DL)-based explanation using OWL 2 axioms. Expected format: <tbox>, <abox>, <dl_explanation>, and <response> tags. The <tbox> defines formal class conditions (⊑, ≡, ⊓). The <abox> lists specific property assertions in the Query image. The <dl_explanation> lists exactly the TBox rules fired by the ABox assertions without natural language.

B. Models Details

All models are accessed via the OpenRouter API in inference-only mode, with temperature set to 0 to ensure deterministic outputs and no fine-tuning or weight updates applied.

Gemini 2.5 Flash (Google DeepMind, 2025) is a state-of-the-art proprietary model from Google DeepMind with advanced multimodal reasoning capabilities and a 1M-token context window. It serves as the high-performance proprietary baseline.

Gemma 4 26B (Gemma Team, Google DeepMind, 2026) is an open-weight Mixture-of-Experts model from Google DeepMind (26B total parameters, 4B active per forward

pass) with a 262K-token context window, included for its competitiveness on multimodal tasks (Ramachandran et al., 2025).

Qwen3 VL 8B (Qwen Team & Alibaba Cloud, 2025) is an open-weight dense model from Alibaba Cloud (8.77B parameters) with early-fusion visual tokenisation and a 131K-token context window. It represents the most parameter-efficient model in our selection.

LLaMA 4 Scout (Meta AI, 2025) is an open-weight Mixture-of-Experts model from Meta (109B total parameters, 17B active per forward pass) with native multimodal support and a 328K-token context window via OpenRouter (Meta AI, 2025). It is widely used as a baseline in recent multimodal studies.

C. Semantic Explainability Metric Definitions

1. Textual Groundedness

Desideratum: Every relevant concept in the image is mentioned in the text explanation.

Definition (IMG → NLE): A well grounded vision-to-text explanation mentions every relevant concept in the image in the explanation.

What it measures: The degree to which the generated text captures all salient objects, attributes, and spatial relationships present in the source image. Whether the text provides an exhaustive account of the visual scene without omitting relevant concepts ((Yang et al., 2022)).

Scores:

- **Not Grounded:** Severe omissions of key concepts in the explanation.
- **Weakly Grounded:** Major omissions of key concepts.
- **Moderately Grounded:** Few omissions of key concepts.
- **Strongly Grounded:** One omission of key concepts.
- **Perfectly Grounded:** Zero omissions of key concepts.

2. Hallucination-Free

Desideratum: Every relevant concept in the explanation is visible in the image.

Definition (NLE → IMG): A hallucination-free explanation reflects accurate visual evidence.

What it measures: Evaluates whether the explanation reflects accurate visual evidence (any claims made in the explanation must be visually verifiable in the query image without hallucinating non-existent features).

Scores:

- **Severe hallucinations**
- **Major hallucinations**
- **Few hallucinations**
- **One hallucination**
- **Hallucination-free:** Zero hallucinations.

3. Concept Counting

Desideratum: The explanation accurately quantifies the exact number of relevant concepts present in the image.

Definition: A precisely counted explanation ensures that the elements counted in the NLE faithfully correspond with the number of concepts in the image.

What it measures: Evaluates the quantitative accuracy of the explanation (penalizing any discrepancies between the stated count of elements in the text and the actual number of countable elements visible in the image).

***Note*:** This metric is different from hallucination-free because it does reflect, for instance, a model explaining a flower of class f has 5 petals while the query image shows 6 petals.

Scores:

- **Wrong concept counting:** 0% of the countable features are well counted.
- **Weak concept counting:** 1–33% of the countable features are well counted.
- **Moderate concept counting:** 34–66% of the countable features are well counted.
- **Almost perfect concept counting:** 67–99% of the countable features are well counted.
- **Perfect concept counting:** 100% of the countable features are well counted.

4. Comprehensibility

Desideratum: The explanation is simple, readable, and perfectly accessible to an end-user without unnecessary complexity.

Definition: A highly comprehensible explanation is easy to read, simple, and accessible to end-users.

What it measures: Evaluates whether: 1) explanations are simple and accessible to end-users by measuring readability and penalizing overly complex or unnecessarily sophisticated explanations; 2) descriptions in terms of symbols are

semantically and structurally similar to those of human experts (Arrieta et al., 2020).

Scores:

- **Not Comprehensible:** Highly complex.
- **Weak Comprehensibility:** Mostly complex.
- **Medium Comprehensibility:** Somewhat complex.
- **Strong Comprehensibility:** A light degree of complexity.
- **Perfectly Comprehensible:** No complexity.

5. Conciseness

Desideratum: The explanation conveys only the strictly necessary information efficiently.

Definition: A concise explanation conveys the necessary information efficiently without unnecessary length nor fillers.

What it measures: Evaluates the length of the explanation (penalizing unnecessarily verbose answers, filler words, or repetition).

Scores:

- **Not Concise:** Highly verbose, most content is redundant or removable.
- **Weakly Concise:** Mostly verbose, several sentences can be removed without loss of meaning.
- **Medium Conciseness:** Somewhat verbose, few sentences or parts can be removed while preserving full meaning.
- **Strongly Concise:** Light verbosity, one sentence can be removed without losing meaning.
- **Perfectly Concise:** No unnecessary verbosity, nothing can be removed without losing meaning.

6. Specificity

Desideratum: The explanation relies exclusively on precise, exact, and non-generic details about the sample being classified.

Definition: A highly specific explanation provides precise, concrete, and non-generic information about the classified subject. This metric is evaluated locally, meaning it only considers the individual sample at hand, without requiring knowledge of other classes.

What it measures: Evaluates whether the explanation relies on precise details on a local sample-basis (penalizing vague, broad, or generic statements).

Scores:

- **Not Specific:** Highly vague, 0% of all features are precise, concrete, and non-generic.
- **Weakly Specific:** Mostly vague, 1–33% of all features are precise, concrete, and non-generic.
- **Moderately Specific:** Somewhat vague, 34–66% of all features are precise, concrete, and non-generic features.
- **Strongly Specific:** Minimal vagueness, 67–99% of all features are precise, concrete, and non-generic.
- **Perfectly Specific:** Zero vagueness, 100% of all features are precise, concrete, and non-generic.

7. Global Discriminativeness

Desideratum: The explanation successfully isolates the unique features that separate the predicted class from all others.

Definition: A discriminative explanation highlights the features and values that truly support the predicted class over all others — focusing on pathognomonic features rather than correlated or commonly shared ones. This metric is evaluated globally: the evaluator must be aware of all classes present in the dataset.

What it measures: Evaluates the capability of providing global key reasons causing a prediction.

Scores:

- **Not Discriminative:** 0% pathognomonic features.
- **Weakly Discriminative:** 1–33% pathognomonic features.
- **Moderately Discriminative:** 34–66% pathognomonic features.
- **Strongly Discriminative:** 67–99% pathognomonic features.
- **Perfectly Discriminative:** 100% pathognomonic features.

8. Local Discriminativeness

Desideratum: The explanation successfully isolates the unique features that separate the predicted class from all others.

Definition: A discriminative explanation highlights the features and values that truly support the predicted class over all others, evaluated locally within the trial.

What it measures: Evaluates the capability of providing local key reasons causing a prediction.

Scores:

- **Not Discriminative:** 0% pathognomonic features.
- **Weakly Discriminative:** 1–33% pathognomonic features.
- **Moderately Discriminative:** 34–66% pathognomonic features.
- **Strongly Discriminative:** 67–99% pathognomonic features.
- **Perfectly Discriminative:** 100% pathognomonic features.

9. Instruction Following

Desideratum: The output flawlessly satisfies all explicit structure and format constraints.

Definition: An explanation that follows instructions is compliant with open-ended and format instructions (Zeng et al., 2024).

What it measures: Evaluates how closely the model's output follows all explicit instructions and constraints in the prompt.

Scores:

- **No Adherence:** 0% followed.
- **Weak Adherence:** 1–33% followed.
- **Partial Adherence:** 34–66% followed.
- **Strong Adherence:** 67–99% followed.
- **Perfect Adherence:** 100% followed.

10. Logical Coherence

Desideratum: The sentences connect logically to form a valid, smooth, and complete deduction.

Definition: A logically coherent explanation builds from sentence to sentence into a valid body of information (Liu et al., 2023).

What it measures: Evaluates whether the explanation forms a valid logical progression.

Scores:

- **Not Coherent:** Completely disconnected clauses.
- **Weakly Coherent:** Severe logical errors.
- **Moderately Coherent:** Some logical issues.
- **Strongly Coherent:** Mostly smooth, minor error.
- **Perfectly Coherent:** Flawless deduction.

11. Explanation Reproducibility

Desideratum: The core reasoning remains consistent across identical runs.

Definition: A reproducible explanation maintains consistency in extracted features across repeated runs.

What it measures: Evaluates stability across runs via feature overlap.

Scores:

- **Not Reproducible:** 0% overlap.
- **Weakly Reproducible:** 1–33% overlap.
- **Moderately Reproducible:** 34–66% overlap.
- **Strongly Reproducible:** 67–99% overlap.
- **Perfectly Reproducible:** 100% overlap.

Note on evaluation scope. Global discriminativeness and explanation reproducibility are included for completeness but are not evaluated in the present study. Both require information beyond a single independent trial in the in-context learning setting (e.g., access to the full class set or repeated runs), and are therefore left for future work.

D. Detailed Results per Model

Tables 7–10 report the full accuracy breakdown per model across all datasets, N-way settings, and K-shot settings. Observations per cell vary by N: N=2 $\rightarrow$ 6 observations, N=3 $\rightarrow$ 4 observations, N=4 $\rightarrow$ 3 observations, Q=1. †Classification-only baseline.

E. Detailed LLM-as-a-judge Results

Tables 11–15 report additional accuracy and judge score breakdowns.

F. Effect of N and K on Accuracy and Explanation Quality

This section provides a brief analysis of how the number of classes (N) and support examples (K) influence both accuracy and explanation quality. The main trends are summarised in Figure 6.

Table 7. Full FS-ICL classification accuracy breakdown for model Gemini 2.5 Flash (%). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. †Classification-only baseline.

	Flowers						**Oxford Pets**						**CIFAR-10**						**DTD**					
	2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way	
Condition	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s
Classif.†	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	100.0	100.0	100.0	91.7	100.0	91.7	91.7	91.7	100.0	100.0	100.0	75.0	91.7
NLE	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	100.0	91.7	91.7	83.3	100.0	100.0	100.0	75.0	100.0
Features	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	100.0	91.7	91.7	75.0	100.0	100.0	100.0	75.0	100.0
Feature-value pairs	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	83.3	100.0	100.0	100.0	91.7	100.0	91.7	91.7	83.3	100.0	91.7	100.0	75.0	91.7
DL Axioms	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	100.0	100.0	100.0	91.7	100.0	91.7	91.7	83.3	100.0	91.7	100.0	83.3	83.3

Table 8. Full FS-ICL classification accuracy breakdown for model Gemma 4 26B (%). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. †Classification-only baseline.

	Flowers						**Oxford Pets**						**CIFAR-10**						**DTD**					
	2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way	
Condition	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s
Classif.†	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	100.0	83.3	100.0	100.0	100.0	100.0	100.0	83.3	100.0	75.0	100.0	100.0	83.3	66.7	83.3
NLE	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	100.0	83.3	91.7	100.0	100.0	100.0	100.0	91.7	100.0	83.3	100.0	100.0	83.3	66.7	91.7
Features	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	100.0	83.3	100.0	100.0	100.0	100.0	100.0	91.7	100.0	58.3	100.0	91.7	83.3	58.3	91.7
Feature-value pairs	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	83.3	100.0	91.7	100.0	100.0	100.0	91.7	100.0	83.3	100.0	66.7	100.0	100.0	83.3	91.7	75.0
DL Axioms	100.0	100.0	91.7	100.0	100.0	100.0	100.0	100.0	75.0	100.0	83.3	83.3	100.0	100.0	100.0	100.0	83.3	100.0	75.0	100.0	83.3	83.3	75.0	83.3

Table 9. Full FS-ICL classification accuracy breakdown for model Qwen3 VL 8B (%). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. †Classification-only baseline.

	Flowers						**Oxford Pets**						**CIFAR-10**						**DTD**					
	2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way	
Condition	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s
Classif.†	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	83.3	100.0	83.3	100.0	100.0	100.0	91.7	100.0	83.3	100.0	91.7	100.0	91.7	75.0	100.0	83.3
NLE	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	100.0	83.3	100.0	91.7	100.0	91.7	100.0	83.3	91.7	83.3	100.0	83.3	75.0	100.0	75.0
Features	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	100.0	83.3	91.7	100.0	100.0	91.7	100.0	83.3	91.7	83.3	100.0	100.0	83.3	91.7	75.0
Feature-value pairs	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	58.3	100.0	83.3	91.7	91.7	100.0	91.7	100.0	75.0	100.0	91.7	91.7	91.7	83.3	91.7	75.0
DL Axioms	100.0	100.0	66.7	100.0	83.3	100.0	100.0	100.0	75.0	100.0	66.7	66.7	91.7	100.0	50.0	100.0	58.3	91.7	75.0	75.0	75.0	83.3	66.7	66.7

Table 10. Full FS-ICL classification accuracy breakdown for model LLaMA 4 Scout (%). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. †Classification-only baseline.

	Flowers						**Oxford Pets**						**CIFAR-10**						**DTD**					
	2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way		2-way		3-way		4-way	
Condition	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s	1s	5s
Classif.†	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	100.0	83.3	100.0	91.7	100.0	75.0	100.0	58.3	100.0	66.7	100.0	75.0	75.0	66.7	75.0
NLE	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	100.0	91.7	100.0	91.7	100.0	75.0	100.0	83.3	100.0	75.0	100.0	83.3	75.0	66.7	58.3
Features	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	58.3	100.0	75.0	100.0	91.7	100.0	83.3	100.0	66.7	100.0	66.7	100.0	91.7	75.0	58.3	58.3
Feature-value pairs	100.0	100.0	91.7	100.0	100.0	100.0	100.0	100.0	83.3	100.0	75.0	83.3	83.3	100.0	66.7	91.7	83.3	83.3	75.0	100.0	83.3	66.7	41.7	66.7
DL Axioms	100.0	100.0	100.0	100.0	75.0	100.0	91.7	100.0	50.0	100.0	100.0	100.0	100.0	100.0	66.7	91.7	66.7	91.7	75.0	100.0	91.7	83.3	75.0	75.0

Table 11. Mean accuracy (%) by few-shot ICL configuration and model, aggregated over 4 datasets and 5 explanation types. $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Values: mean (±SE). Best per row in bold.

Config (N, K)	Gem. 2.5F	Gem. 4	Qwen	LLaMA
N=2, K=1	**95.8** (1.3)	92.9 (1.7)	95.0 (1.4)	90.4 (1.9)
N=2, K=5	**100.0** (0.0)	**100.0** (0.0)	98.3 (0.8)	**100.0** (0.0)
N=3, K=1	**97.1** (1.1)	92.5 (1.7)	84.2 (2.4)	80.4 (2.6)
N=3, K=5	**100.0** (0.0)	95.8 (1.3)	95.0 (1.4)	92.9 (1.7)
N=4, K=1	**90.4** (1.9)	85.8 (2.3)	85.8 (2.3)	78.3 (2.7)
N=4, K=5	**96.2** (1.2)	95.0 (1.4)	90.0 (1.9)	89.2 (2.0)

Table 12. Mean accuracy (%) by model and dataset, aggregated over 5 explanation types and 6 few-shot configurations: $5 \times 26 = 130$ observations per cell. $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Values: mean (±SE). Best per column in bold.

Model	Flowers	Pets	CIFAR	DTD
Gemini 2.5 Flash	**100.0** (0.0)	**98.9** (0.6)	95.8 (1.1)	**91.7** (1.5)
Gemma 4 26B	99.7 (0.3)	93.1 (1.3)	**97.5** (0.8)	84.4 (1.9)
Qwen3 VL 8B	98.3 (0.7)	90.3 (1.6)	91.7 (1.5)	85.3 (1.9)
LLaMA 4 Scout	98.9 (0.6)	90.8 (1.5)	88.1 (1.7)	76.4 (2.2)

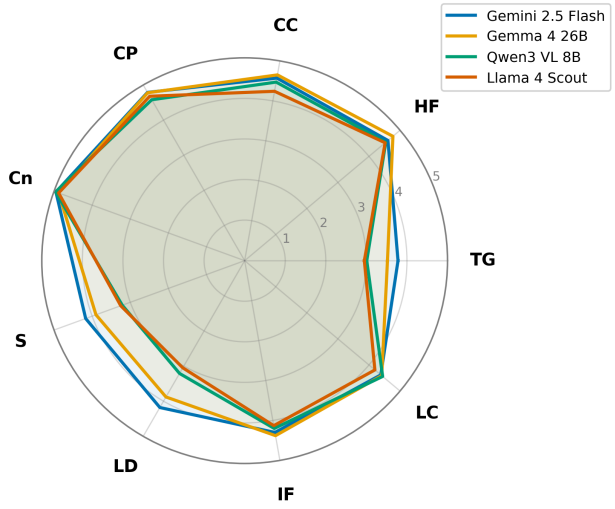

Figure 5. LLM-as-a-judge scores by model, averaged over all conditions and datasets. Gemini 2.5 Flash maintains the largest polygon on TG, S, and LD; Qwen3 VL 8B and LLaMA 4 Scout show the most pronounced collapses. The metrics are: Textual Groundedness (TG), Hallucination Free (HF), Concept Counting (CC), Comprehensibility (CP), Conciseness (Cn), Specificity (S), Local Discriminativeness (LD), Instruction Following (IF), and Logical Coherence (LC)

G. Ablation Study: Effect of Support-Image Access on Judge Scores

The judge pipeline described in Section 4 evaluates explanations using only the query image, without access to the few-shot support images seen by the classifier. Beyond validating this specific design choice, this experiment reveals which evaluation metrics are inherently sensitive to visual access and which are not. All results reported in Section 6 are based exclusively on the original GPT-QO judge and are unaffected by the limitations of the Qwen judge described below; this appendix is a validation of the original judge design, not a replacement or extension of those results.

To isolate the effect of support-image access, we ran a second judge (`Qwen3-VL-32B-Thinking`) on the same explanations under two conditions. **Qwen-QueryOnly** (Qwen-QO) replicates the original setup: query image only, no support images. **Qwen-QueryAndSupport** (Qwen-QAS) additionally provides the full set of labelled support images. The original judge (`gpt-5-thinking-mini`) is labelled **GPT-QueryOnly** (GPT-QO).

This design yields two distinct comparisons: (i) *Qwen-QO vs. Qwen-QAS* isolates the effect of support-image access within a fixed model; (ii) *GPT-QO vs. Qwen-QO* probes between-model calibration under identical conditions. All three judges scored the same 3,504 trials. The DL Axioms prompt type contributes fewer trials (497 vs. ≈1,000; Table 20) due to parse errors by Qwen-QO on structured outputs (the judge model exhausted its token budget while reasoning over the formal DL syntax before producing the scoring XML, leaving those trials without a valid score). We report results both including and excluding this subset to verify that conclusions do not depend on this incomplete subset.

It should be noted that the 497 scored axioms trials are not a random sample of the full axioms set: they are most likely the trials whose classifier outputs were shorter or structurally simpler, since more elaborate DL Axiom outputs tend to exhaust the judge's token budget before the scoring XML is produced. This introduces a selection bias — the correlations reported for the axioms subset may be overestimated relative to what would be observed on the full set. For this reason, the main conclusions of this appendix rest on the three prompt types where coverage is complete (Features, NLE, Rule-based; $n = 3{,}007$ after removing axioms trials), and axioms results are reported as additional evidence with this limitation explicit. Addressing this limitation would require either a non-reasoning judge model (which foregoes the chain-of-thought transparency) or a dedicated formal verifier for OWL 2 outputs; both are left for future work given the computational constraints of this study.

Table 13. Best-performing explanation condition per model and dataset (highest mean accuracy averaged over $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1).

	Flowers		Oxford Pets		CIFAR-10		DTD	
Model	**Best cond.**	**Acc.**	**Best cond.**	**Acc.**	**Best cond.**	**Acc.**	**Best cond.**	**Acc.**
Gemini 2.5 Flash	Classification	100.0	Classification	100.0	Classification	95.8	Classification	93.1
Gemma 4 26B	Classification	100.0	Classification	95.8	NLE	98.6	NLE	87.5
Qwen3 VL 8B	Classification	100.0	Classification	94.4	Classification	95.8	Classification	90.3
LLaMA 4 Scout	Classification	100.0	NLE	93.1	NLE	91.7	DL Axioms	86.1

Table 14. Mean LLM-as-a-judge score (1–5 ↑) by model and metric, aggregated over 4 explanation conditions, 4 datasets, and 6 few-shot configurations: $4 \times 26 = 104$ observations per cell. $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Values: mean (±SE). Best per XAI metric in **bold**.

Model	TG	HF	CC	CP	Cn	S	LD	IF	LC
Gemini 2.5 Flash	**3.78** (.03)	4.61 (.03)	4.57 (.04)	4.79 (.01)	**4.95** (.01)	**4.17** (.02)	**4.18** (.03)	4.30 (.03)	4.38 (.03)
Gemma 4 26B	3.52 (.03)	**4.77** (.02)	**4.65** (.03)	**4.78** (.01)	4.88 (.01)	3.90 (.03)	3.88 (.04)	**4.38** (.03)	**4.41** (.03)
Qwen3 VL 8B	3.01 (.04)	4.54 (.03)	4.47 (.04)	4.58 (.02)	4.96 (.01)	3.21 (.04)	3.21 (.04)	4.20 (.03)	4.44 (.03)
LLaMA 4 Scout	2.95 (.03)	4.52 (.03)	4.24 (.04)	4.68 (.02)	4.88 (.01)	3.25 (.03)	3.05 (.04)	4.13 (.03)	4.19 (.04)

Table 15. Mean LLM-as-a-judge score (1–5 ↑) by dataset and metric, aggregated over 4 models, 4 explanation conditions, and 6 few-shot configurations: $4 \times 26 = 104$ observations per cell. $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Values: mean (±SE). **Bold** = best per XAI metric.

Dataset	TG	HF	CC	CP	Cn	S	LD	IF	LC
Flowers	**3.38** (.03)	**4.64** (.02)	4.36 (.04)	4.72 (.02)	**4.96** (.01)	**3.79** (.03)	**3.64** (.04)	**4.36** (.03)	**4.50** (.03)
Oxford Pets	3.20 (.03)	4.71 (.02)	**4.76** (.03)	**4.74** (.02)	4.96 (.01)	3.66 (.03)	3.55 (.04)	4.35 (.03)	4.41 (.03)
CIFAR-10	3.21 (.04)	4.56 (.03)	4.45 (.04)	4.70 (.02)	4.92 (.01)	3.37 (.03)	3.60 (.04)	4.16 (.03)	4.27 (.04)
DTD	3.46 (.04)	4.53 (.03)	4.36 (.04)	4.68 (.02)	4.83 (.02)	3.71 (.03)	3.53 (.04)	4.14 (.03)	4.23 (.04)

Table 16. Full LLM-as-a-judge score breakdown (1–5 ↑) for model (Gemini 2.5 Flash). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Observations per cell vary by N: N=2 → 6 observations, N=3 → 4 observations, N=4 → 3 observations. Values: mean (±SE). **Bold** indicates best score per XAI metric.

Condition	TG	HF	CC	CP	Cn	S	LD	IF	LC
NLE	3.76 (.04)	4.49 (.05)	**4.77** (.05)	**4.99** (.01)	4.90 (.02)	4.02 (.04)	4.17 (.06)	4.67 (.03)	**4.96** (.02)
Features	3.94 (.05)	4.77 (.04)	4.63 (.06)	4.97 (.01)	4.97 (.01)	4.08 (.05)	4.15 (.07)	**4.72** (.04)	4.93 (.02)
Feature-value pairs	**4.24** (.03)	**4.82** (.03)	4.51 (.07)	4.97 (.01)	**4.98** (.01)	**4.61** (.03)	**4.39** (.06)	4.70 (.04)	4.76 (.04)
DL Axioms	3.16 (.06)	4.38 (.07)	4.39 (.09)	4.20 (.03)	4.94 (.02)	3.98 (.06)	4.01 (.07)	3.12 (.05)	2.86 (.07)

Table 17. Full LLM-as-a-judge score breakdown (1–5 ↑) for model (Gemma 4 26B). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Observations per cell vary by N: N=2 → 6 observations, N=3 → 4 observations, N=4 → 3 observations. Values: mean (±SE). **Bold** indicates best score per XAI metric.

Condition	TG	HF	CC	CP	Cn	S	LD	IF	LC
NLE	3.48 (.05)	4.61 (.05)	4.63 (.07)	4.93 (.02)	4.83 (.04)	3.61 (.05)	3.69 (.08)	4.82 (.03)	4.81 (.03)
Features	3.71 (.05)	**4.91** (.02)	**4.88** (.04)	**5.00** (.00)	**4.99** (.01)	3.92 (.05)	3.74 (.08)	**4.95** (.01)	**4.93** (.02)
Feature-value pairs	**3.92** (.04)	4.88 (.03)	4.61 (.07)	4.90 (.02)	4.86 (.03)	**4.29** (.04)	**4.11** (.07)	4.56 (.05)	4.69 (.04)
DL Axioms	2.96 (.06)	4.66 (.05)	4.50 (.08)	4.30 (.04)	4.86 (.03)	3.79 (.05)	3.99 (.07)	3.17 (.05)	3.20 (.08)

Table 18. Full LLM-as-a-judge score breakdown (1–5 ↑) for model (Qwen3 VL 8B). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Observations per cell vary by N: N=2 → 6 observations, N=3 → 4 observations, N=4 → 3 observations. Values: mean (±SE). **Bold** indicates best score per XAI metric.

Condition	TG	HF	CC	CP	Cn	S	LD	IF	LC
NLE	**3.76** (.04)	4.35 (.06)	**4.77** (.05)	4.96 (.01)	4.84 (.02)	3.85 (.04)	3.75 (.07)	4.64 (.04)	4.85 (.03)
Features	3.55 (.06)	**4.83** (.03)	4.76 (.05)	**4.99** (.01)	4.99 (.01)	3.79 (.05)	3.42 (.08)	**4.90** (.02)	4.89 (.03)
Feature-value pairs	3.66 (.05)	4.81 (.03)	4.38 (.08)	**4.99** (.01)	**5.00** (.00)	**4.08** (.04)	**3.97** (.07)	4.24 (.06)	**4.92** (.02)
DL Axioms	1.06 (.02)	4.16 (.07)	3.97 (.10)	3.39 (.04)	**5.00** (.00)	1.09 (.03)	1.69 (.07)	3.00 (.06)	3.10 (.09)

Table 19. Full LLM-as-a-judge score breakdown (1–5 ↑) for model (LLaMA 4 Scout). $N \in \{2,3,4\}$, $K \in \{1,5\}$, Q=1. Observations per cell vary by N: N=2 → 6 observations, N=3 → 4 observations, N=4 → 3 observations. Values: mean (±SE). **Bold** indicates best score per XAI metric.

Condition	TG	HF	CC	CP	Cn	S	LD	IF	LC
NLE	**3.48** (.05)	4.40 (.06)	**4.56** (.07)	4.92 (.02)	4.67 (.04)	3.43 (.05)	3.15 (.08)	4.68 (.04)	4.74 (.04)
Features	3.30 (.06)	**4.72** (.04)	4.47 (.08)	**4.99** (.01)	4.93 (.02)	3.44 (.05)	3.17 (.08)	**4.73** (.04)	**4.79** (.04)
Feature-value pairs	2.97 (.06)	4.57 (.05)	4.00 (.10)	4.83 (.03)	**4.97** (.01)	**3.57** (.05)	**3.19** (.08)	4.22 (.06)	4.49 (.05)
DL Axioms	2.05 (.07)	4.40 (.07)	3.94 (.10)	4.00 (.04)	4.94 (.03)	2.57 (.08)	2.69 (.09)	2.91 (.05)	2.72 (.07)

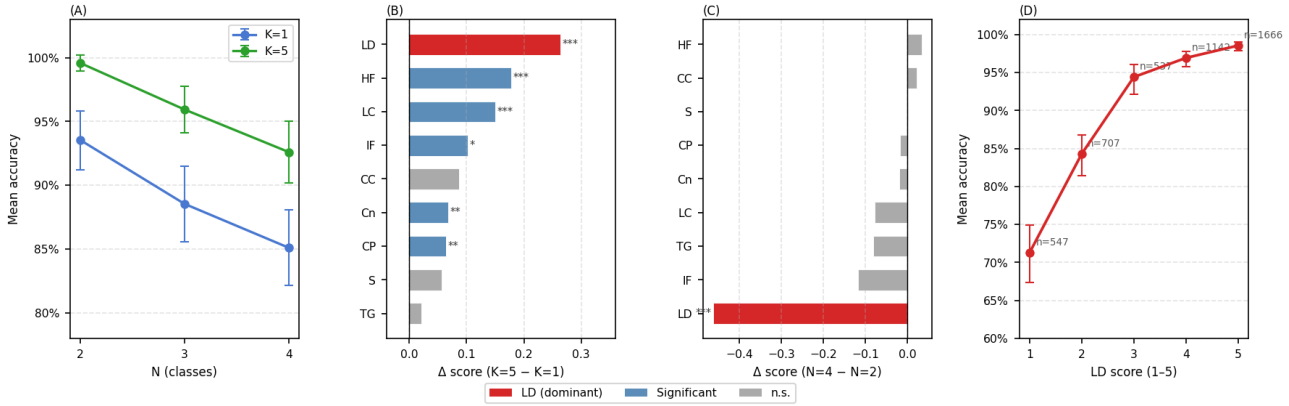

Figure 6. (A) Mean classification accuracy vs. number of classes N, by support shots K. K=5 raises accuracy by +7.0 pp on average (p=2.0×10^{-13}, pooled Wilcoxon); accuracy falls monotonically with N at both K levels (Friedman χ^2=75.2, p=4.7×10^{-17}). Error bars: 95% CI. (B) Effect of K on each XAI metric (Δ = K=5 − K=1, pooled across N; Bonferroni ×40). Local Discriminativeness shows the largest gain (Δ=+0.26, p_{Bonf}=1.5×10^{-6}); TG, CC, and S are unaffected. (C) Effect of N on each XAI metric (Δ = N=4 − N=2, pooled across K; Bonferroni ×30). Only Local Discriminativeness shows a significant N-trend (χ^2=51.5, p_{Bonf}=1.9×10^{-10}). (D) Mean classification accuracy as a function of LD score (pooled across all models, datasets, and conditions E2–E5; n = trials per bin; error bars: 95% CI). Significance markers: * p_{Bonf}<0.05, ** p_{Bonf}<0.01, *** p_{Bonf}<0.001. Metric abbreviations as in Table 1.

Table 20. Trial counts at the intersection of all three judge conditions by prompt type. The reduced count for DL Axioms is due to Qwen-QO parse errors. Individual-judge counts reflect all parseable outputs before intersection.

Prompt type	Intersect.	GPT-QO	Qwen-QO	Qwen-QAS
DL Axioms	497	1,146	588	985
Features	999	1,152	1,062	1,084
NLE	1,030	1,151	1,075	1,101
Rule-based	978	1,150	1,050	1,069
Total	**3,504**			

Results

Table 21 reports Spearman ρ and mean score differences for all nine metrics, Bonferroni-corrected over 27 simultaneous tests (9 metrics $\times$ 3 pairs).

Interpretation

The nine metrics fall into three groups according to how the Qwen-QO vs. Qwen-QAS comparison (the clean within-model test) behaves.

Metrics robust to visual access (Comprehensibility, Specificity, Local Discriminativeness). These three metrics show the highest within-model agreement across conditions ($\rho = 0.62$–0.83) and negligible mean score shifts ($|\Delta| \leq 0.06$ for the Qwen pair). This pattern makes conceptual sense: Comprehensibility and Specificity evaluate properties of the text itself that a judge can assess from the explanation alone, regardless of what the support images look like. Local Discriminativeness is notable in this group: despite requiring class-contrastive reasoning, the ranking of explanations on LD is essentially unaffected by whether the judge sees the support set ($\rho = 0.627$, mean diff $= -0.049$). This supports the design rationale that LD can be reliably assessed through the judge's prior class knowledge combined with the query image. The GPT-QO vs. Qwen-QAS correlation for LD ($\rho = 0.531$) confirms that even across different model families the ranking order is preserved, with only a constant calibration offset of -0.421 points that is uniform across prompt types.

Metrics sensitive to visual access (Textual Groundedness, Hallucination Free). Both metrics show notably lower ranking agreement between the two Qwen conditions — that is, lower Qwen-QO vs. Qwen-QAS correlation — ($\rho = 0.43$–0.56) than the three metrics in the robust group (Comprehensibility, Specificity, and Local Discriminativeness, $\rho = 0.62$–0.83). This is precisely what their definitions predict: Textual Groundedness measures whether the explanation covers all salient image concepts, and Hallucination Free measures whether every claim is visually verifiable — both are judgments that require seeing the images. When the classifier generates an explanation that references visual patterns from the support set (e.g. "unlike the other examples, this flower has broader petals"), a query-only judge cannot verify whether that comparison is accurate, because it never saw those support images. A judge with access to the full support set can, and will sometimes score the same explanation differently as a result. The lower Qwen-QO vs. Qwen-QAS agreement for these two metrics therefore reflects a genuine limitation of query-only evaluation for TG and HF specifically: the judge is missing part of the visual evidence these metrics are defined against. This is a known cost of our compute-constrained design and points towards future evaluation pipelines that provide the full visual context to the judge for these two metrics in particular.

Intermediate metrics: form-dependent but context-sensitive (Conciseness, Instruction Following, Logical Coherence). These three metrics might be expected to be entirely image-independent because they evaluate structural or linguistic properties of the explanation text. Indeed, their mean score shifts between Qwen-QO and Qwen-QAS are negligible ($|\Delta| < 0.06$). However, their ρ values are substantially lower than those of the robust group (Comprehensibility, Specificity, and Local Discriminativeness; $\rho = 0.62$–0.83), reaching only 0.32–0.50, and this gap is largely driven by the DL Axioms subset. Excluding axioms, the within-model correlations for Conciseness, Instruction Following, and Logical Coherence recover considerably ($\rho = 0.37, 0.30, 0.48$ respectively), and cross-model agreement for Logical Coherence approaches zero mean difference. This suggests the lower correlations in the 3,504-trial set (which includes the 497 DL Axiom trials) reflect the difficulty of scoring those outputs consistently, rather than genuine sensitivity to visual access per se. When the classifier produces a malformed or partial OWL 2 axiom, metrics like Instruction Following and Logical Coherence become ambiguous to score: the output is neither well-formed natural language nor valid formal syntax, so any LLM judge that has not been specifically calibrated on that format will make inconsistent judgments, and the two Qwen conditions, despite sharing the same model, will sometimes disagree for that reason alone. Rigorously evaluating DL Axiom outputs would require formal verification with a dedicated OWL 2 reasoner (e.g. OntoAligner (Babaei Giglou et al., 2025) or similar tools), which could check structural validity and logical consistency directly rather than delegating that judgment to an LLM; we leave this for future work.

Concept Counting. Concept Counting stands apart: its within-model agreement is moderate ($\rho = 0.52$) but its cross-model agreement is very weak ($\rho \approx 0.17$), the lowest of all nine metrics. This pattern indicates that GPT and Qwen apply fundamentally different interpretations of what constitutes a countable concept in an explanation, a model-level disagreement that is independent of image access. We

Table 21. Cross-judge validation: Spearman ρ [95 % CI] between judge conditions and mean score difference (GPT-QO − Qwen-QAS) for all nine XAI metrics ($n = 3{,}504$ trials; all $p_{\text{Bonf}} < 0.001$). **Bold** highlights the Local Discriminativeness row. The Qwen-QO vs. Qwen-QAS column isolates the effect of support-image access within a fixed model.

Metric	GPT-QO vs. Qwen-QO ρ [CI]	GPT-QO vs. Qwen-QAS ρ [CI]	Qwen-QO vs. Qwen-QAS ρ [CI]	Mean diff. GPT-QO−QAS
Comprehensibility	0.691 [0.655, 0.725]	0.733 [0.699, 0.763]	0.827 [0.801, 0.853]	+0.193
Specificity	0.549 [0.523, 0.573]	0.562 [0.536, 0.586]	0.620 [0.591, 0.648]	−0.478
Local Discriminativeness	**0.536 [0.509, 0.562]**	**0.531 [0.502, 0.558]**	**0.627 [0.598, 0.653]**	**−0.421**
Textual Groundedness	0.456 [0.426, 0.484]	0.494 [0.468, 0.521]	0.562 [0.536, 0.590]	−0.339
Hallucination Free	0.339 [0.298, 0.377]	0.278 [0.237, 0.319]	0.427 [0.388, 0.466]	−0.031
Conciseness	0.445 [0.391, 0.497]	0.361 [0.304, 0.419]	0.316 [0.266, 0.364]	+0.090
Instruction Following	0.246 [0.207, 0.283]	0.315 [0.278, 0.351]	0.339 [0.286, 0.390]	−0.423
Logical Coherence	0.411 [0.370, 0.450]	0.381 [0.340, 0.421]	0.496 [0.451, 0.538]	−0.235
Concept Counting	0.169 [0.127, 0.212]	0.170 [0.127, 0.214]	0.523 [0.477, 0.566]	−0.153

advise treating Concept Counting results with additional caution. Unlike Local Discriminativeness — where GPT and Qwen produce different absolute scores but preserve a similar ranking order ($\rho = 0.53$) — for Concept Counting not even the ranking is preserved across models ($\rho \approx 0.17$). This means the discrepancy is not a matter of calibration that cancels in relative comparisons: the two models are to some extent measuring different things. Consequently, if this experiment were reproduced with a different judge model, the Concept Counting scores would likely not be comparable to ours — not because the explanations changed, but because a different judge may count concepts differently.

Summary and Implications

Local Discriminativeness, the metric most central to the paper's findings, falls in the robust group. Two independent pieces of evidence support this. First, within the same model, providing or withholding the support images produces no meaningful reordering of explanations on LD ($\rho = 0.627$, mean score shift $= -0.049$ points on a 1–5 scale): a judge that has never seen the support set reaches essentially the same LD ranking as one that has. Second, even across two different model families with different visual access — GPT-QO and Qwen-QAS — the LD ranking is preserved ($\rho = 0.531$), with only a constant calibration offset of -0.421 points that is uniform across all four prompt types (NLE, Features, Rule-based, DL Axioms) and therefore cancels entirely in the between-condition comparisons on which the paper's conclusions rest. In other words, the finding that LD is the strongest predictor of classification accuracy, and that it responds specifically to K and N manipulations, holds regardless of which judge model is used and regardless of whether that judge sees the support images.

A natural concern is whether a judge without support images can reliably assess a metric that requires class-contrastive reasoning. The answer, empirically, is yes for LD: the judge's prior knowledge of the candidate classes — which are drawn from standard benchmarks (CIFAR-10, DTD, Oxford Flowers, Oxford-IIIT Pets) — is sufficient to assess whether an explanation highlights features that separate the predicted class from the alternatives. The metrics that genuinely require full visual access are Textual Groundedness and Hallucination Free, and this is theoretically expected: both are defined in terms of verifying claims against the images, so a judge that sees more images will score them differently. This is a scoping limitation we acknowledge and address in future work, not a threat to the validity of the LD-based conclusions.

H. Human Validation of the LLM-as-a-Judge Pipeline

To complement the automated ablation study in Appendix G, we conducted an independent human evaluation in which three researchers manually scored a stratified sample of explanations using the same rubrics provided to the judge model. This appendix describes the annotation interface, the sampling protocol, the statistical methods used, and the results of the human–judge alignment analysis.

Annotation Interface

To collect human judgements, we developed a local web application that presents each trial to the annotator in a three-column layout: the query image with the predicted class label (highlighted in green if correct, red if incorrect) on the left; the full explanation generated by the classifier model, together with the labelled support images in a collapsible panel, in the centre; and a scoring form on the right, where each of the nine XAI metrics is rated via radio buttons (1–5), with a per-metric expandable rubric available at all

times via an info button (**i**). Annotations are written to disk immediately upon clicking *Save & Next*, so that progress is preserved across sessions. The application is included in the public repository.

Sampling Protocol

From the 4,608 trials with generated explanations (conditions E2–E5), we selected 192 via stratified random sampling across 64 strata—unique combinations of explanation condition, model, and dataset ($4 \times 4 \times 4$)—with 3 items per stratum. This corresponds to 4.2% of the total, a proportion consistent with established practice for human validation of LLM-as-a-judge pipelines (Zheng et al., 2023; Chiang & Lee, 2023). Stratified sampling ensures uniform coverage of all experimental combinations, which is methodologically superior to a larger simple random sample that could under-represent rare or difficult conditions (Chiang & Lee, 2023).

Within each stratum, items were selected at random (seed = 42) with the constraint that at least one incorrectly classified trial be included, to avoid positive-outcome bias — models tend to produce higher-quality explanations when the classification is correct. The 14 strata for which no incorrect predictions were available (concentrated in Flowers and Pets, where model accuracy exceeds 99%) were filled entirely with correct predictions; these exceptions are documented in the sampling CSV provided in the repository.

The 192 items were distributed among three researchers such that each item was scored by exactly two independent annotators, providing complete double-coverage of the sample. Each researcher evaluated 128 items. Due to sequential ordering of items by explanation condition prior to block assignment, the distribution of conditions is not uniform across annotator pairs; we therefore focus our analysis on judge–human correlation rather than inter-annotator agreement, which would conflate annotator calibration differences with material difficulty.

Evaluation Metrics

We report two complementary statistics (Mathur et al., 2020) and an additional proposed metric to capture the singularities of the evaluation.

Spearman rank correlation (ρ) measures whether the judge LLM ranks items from highest to lowest explanation quality in the same order as the human annotators. It is computed between the judge score and the mean of the two human scores for each of the 192 items. ρ ranges from -1 to 1: a value of 1 indicates identical ranking, 0 indicates no relationship, and negative values indicate opposite orderings. Unlike Pearson correlation, Spearman makes no assumption of linearity or normality, making it appropriate for ordinal 1–5 scales. We adopt $\rho \geq 0.60$ as the threshold for substantial agreement, following the scale of Landis & Koch (1977) as applied to rank-based measures in natural language generation (NLG) evaluation (Zheng et al., 2023; Liu et al., 2023; Chiang & Lee, 2023).

Mean Absolute Error (MAE) measures the average absolute difference in points between the judge score and the human mean. It ranges from 0 (perfect agreement) to 4 (maximum possible error on a 1–5 scale). MAE complements ρ by detecting systematic magnitude biases: a judge that consistently scores two points above or below humans would yield a high MAE even if ρ is high, because ρ is sensitive only to ranking order, not to absolute values.

Tolerance-Weighted Agreement (TWA) To complement error magnitude (MAE) and rank correlation (Spearman ρ) evaluations, we introduce the Tolerance-Weighted Agreement (TWA) metric. Since qualitative 1-5 ordinal scales are inherently susceptible to subjective human calibration biases, a one-point divergence (e.g., a 4 versus a 5) often reflects natural annotator noise rather than structural semantic disagreement. To mitigate this effect and capture the operational utility of the LLM-as-a-judge, TWA acts as a relaxed categorical accuracy that grants partial credit to adjacent predictions. Let $y_{human}, y_{judge} \in \{1, 2, 3, 4, 5\}$ represent the scores assigned by the human and the algorithmic judge, respectively. Defining the absolute distance between the judge and human scores as $d = |y_{judge} - y_{human}|$, the metric is formulated as:

$$S(d) = \begin{cases} 1 & \text{if } d = 0 \\ 0.75 & \text{if } d = 1 \\ 0 & \text{if } d > 1 \end{cases}$$

TWA provides a pragmatic assessment of functional alignment. By highly rewarding $d = 1$ distances (buffering human subjectivity) and applying a strict penalty to larger divergences ($d > 1$), this metric interpretably isolates the proportion of algorithmic verdicts that are practically equivalent to human judgments. Its inclusion alongside MAE and Spearman enables robust triangulation: TWA confirms practical, item-by-item consensus.

TWA score equal or greater than 0.75 indicates that the judge and human ratings differ by at most one point on average, with values approaching 1.0 reflecting exact agreement. Since most dimensions evaluated yielded scores at or above this threshold, the pragmatic discrepancies between the LLM and human annotators are largely confined to adjacent ratings rather than severe misalignments. To ensure the metric remains rigorously intolerant of severe evaluative misalignments, any absolute distance strictly greater than one ($d > 1$) between the judge and human scores is penalized with a score of 0. Eliminating partial credit for these severe cases prevents artificial inflation of the agreement

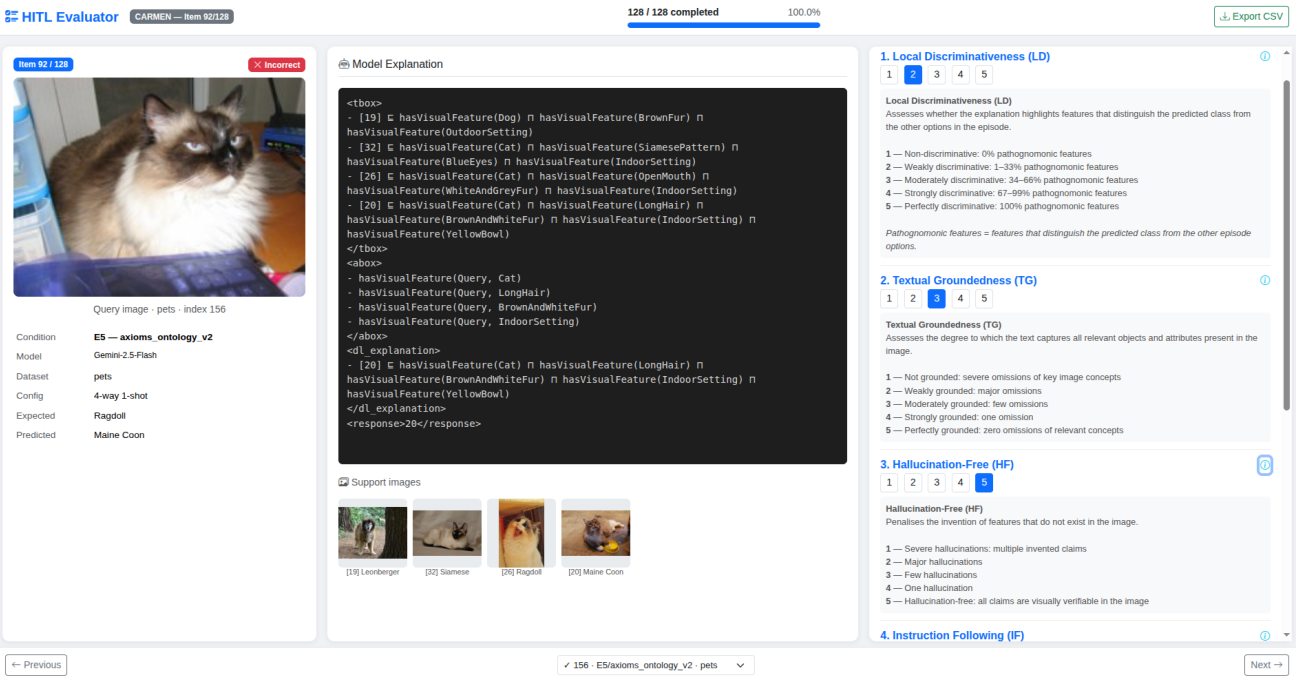

Figure 7. Screenshot of the human annotation interface used for judge validation. **Left column:** query image with ground-truth class and predicted label; a green border indicates a correct prediction, red an incorrect one. **Centre column:** explanation generated by the classifier model (preserving XML tag formatting); below it, a collapsible panel displays the labelled support images with their class names. **Right column:** scoring form for the nine XAI metrics; each metric shows radio buttons for scores 1–5, with an expandable rubric (button **i**) providing the full scoring criteria; a progress bar at the top indicates how many of the 128 assigned items have been completed.

score. This adjustment ensures that the Tolerance-Weighted Agreement (TWA) is both operationally pragmatic regarding minor human subjectivity and rigorously intolerant of severe evaluative misalignments.

Results

Table 22 reports Spearman ρ and MAE for all nine metrics, and Figure 8 visualises the ρ values against the $\rho = 0.60$ threshold. TWA scores are reported by XAI metric in Figure 9, by explanation condition in Figure 10, and by dataset in Figure 11.

Table 22. Human validation results for the LLM-as-a-judge pipeline ($n = 192$ stratified trials, 3 annotators, each item scored by 2 annotators independently). **Spearman** ρ: rank correlation between judge score and mean human score (range -1 to 1; higher is better; $\rho \geq 0.60$ indicates substantial agreement). **MAE**: mean absolute error in points between judge score and mean human score (range 0–4; lower is better). A horizontal rule separates metrics above and below the $\rho = 0.60$ threshold.

Metric	Spearman ρ	MAE
Textual Groundedness (TG)	0.741	0.672
Comprehensibility (CP)	0.732	0.276
Instruction Following (IF)	0.729	0.555
Local Discriminativeness (LD)	0.656	0.865
Logical Coherence (LC)	0.624	0.560
Specificity (S)	0.607	0.646
Conciseness (Cn)	0.498	0.198
Hallucination Free (HF)	0.376	0.474
Concept Counting (CC)	0.234	0.867

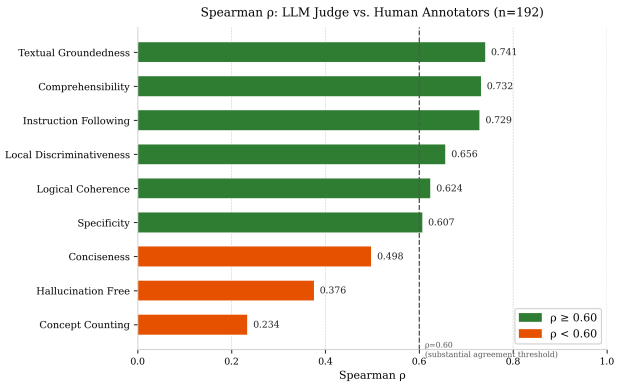

Figure 8. Spearman ρ between the LLM-as-a-judge scores and the mean of two independent human annotators, for each of the nine XAI metrics ($n = 192$ stratified trials). Each bar shows the ρ value for one metric; bars are ordered from highest to lowest. **Green bars** ($\rho \geq 0.60$) indicate metrics for which the judge's ranking closely matches the human ranking, consistent with substantial agreement (Landis & Koch, 1977). **Orange bars** ($\rho < 0.60$) indicate metrics where alignment is weaker. The dashed vertical line marks the $\rho = 0.60$ threshold. Six of nine metrics exceed this threshold, including Local Discriminativeness ($\rho = 0.656$), the metric most central to the paper's conclusions.

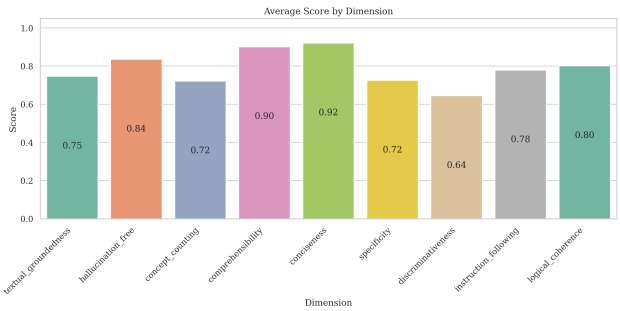

Figure 9. Average TWA scores by XAI metric. The LLM-as-a-judge demonstrates strong practical alignment on dimensions like Conciseness (0.92) and Comprehensibility (0.90). Evaluative dimensions requiring class-contrastive logic, such as Local Discriminativeness (0.64), maintain robust but comparatively lower agreement.

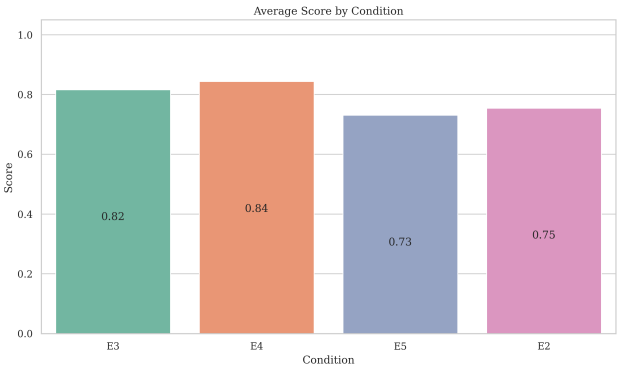

Figure 10. Average TWA scores categorized by explanation condition. Structured intermediate formats like Feature-value pairs (E4, 0.84) and Features (E3, 0.82) yield the highest human-judge alignment, outperforming free-text NLE (E2, 0.75) and formal DL Axioms (E5, 0.73).

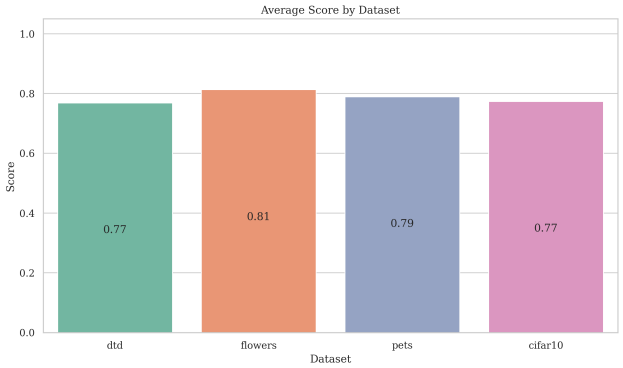

Figure 11. Average TWA scores across the four evaluated image datasets. Human-judge agreement remains highly consistent across different visual domains, ranging from 0.77 on abstract textures (DTD) and generic objects (CIFAR-10) to 0.81 on fine-grained categories (Flowers), indicating stability regardless of dataset complexity.

Discussion

The judge achieves $\rho > 0.60$ on the six metrics most relevant to the between-condition comparisons reported in the paper: Textual Groundedness, Comprehensibility, Instruction Following, Local Discriminativeness, Logical Coherence, and Specificity. The lower ρ values for Conciseness, Hallucination Free, and Concept Counting do not reflect judge unreliability. The first two suffer from a ceiling effect: most explanations score near 5 on these dimensions across all conditions, so there is little variation to rank. Concept Counting is structurally different: it requires exact enumeration of visual elements in the image, a task that proved difficult to apply consistently even among human annotators. The higher MAE values observed for Local Discriminativeness and Textual Groundedness reflect annotator calibration differences rather than systematic judge error: the judge's absolute scores align closely with two of the three annotators, while the third annotator applied consistently higher ratings across all metrics, inflating the human mean for those dimensions.

To reconcile these statistical nuances, the Tolerance-Weighted Agreement (TWA) provides a clearer picture of the judge's operational reliability. By mitigating the impact of the ceiling effects mentioned above and the calibration discrepancies—granting partial credit for adjacent scores that reflect subjective human noise while strictly penalizing severe semantic disagreements ($d > 1$) with a score of zero—the TWA confirms that the judge's evaluations remain functionally aligned with human consensus. Ultimately, the consistently high TWA scores confirm that the LLM-as-a-judge succeeds as a robust and practical evaluator, even in dimensions where ρ or MAE penalize strict numerical divergence.