

# The Imperfective Paradox in Large Language Models

Bolei Ma\*  
LMU Munich & MCML  
bolei.ma@lmu.de

Yusuke Miyao  
The University of Tokyo  
yusuke@is.s.u-tokyo.ac.jp

## Abstract

Do Large Language Models (LLMs) genuinely grasp the compositional semantics of events, or do they rely on surface-level probabilistic heuristics? We investigate the **Imperfective Paradox**, a logical phenomenon where the past progressive aspect entails event realization for activities (e.g., *running* → *ran*) but not for accomplishments (e.g., *building* → *built*). We introduce IMPERFECTIVENLI, a diagnostic dataset designed to probe this distinction across diverse semantic classes. Evaluating state-of-the-art open-weight models, we uncover a pervasive **Teleological Bias**: models systematically hallucinate completion for goal-oriented events, even overriding explicit textual cancellation. Prompting interventions partially reduce this bias but trigger a calibration crisis, causing models to incorrectly reject valid entailments for atelic verbs. Representational analyses further show that while internal embeddings often distinguish progressive from simple past forms, inference decisions are dominated by strong priors about goal attainment. Taken together, our findings indicate that these current open-weight LLMs operate as predictive narrative engines rather than faithful logical reasoners, and that resolving aspectual inference requires moving beyond prompting toward structurally grounded alignment.<sup>1</sup>

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable proficiency in natural language inference (NLI) and complex reasoning tasks (Brown et al., 2020; Touvron et al., 2023). Despite these advancements, questions remain regarding the robustness of their reasoning: do models genuinely grasp the compositional semantics of events, or do they rely on surface-level probabilistic heuristics (Bender et al., 2021; Dziri et al., 2023)? This

\*Work done while visiting The University of Tokyo.

<sup>1</sup>The data and code are available at: <https://github.com/boleima/ImperfectiveParadox>.

Class	Examples (Premise → Hypothesis)	Entail?
Telic	<b>P:</b> The carpenter <i>was building</i> a gazebo. <b>H:</b> The carpenter <i>built</i> a gazebo.	✗ No
Atelic	<b>P:</b> The boy <i>was running</i> in the park. <b>H:</b> The boy <i>ran</i> in the park.	✓ Yes

Table 1: Examples for the Imperfective Paradox. Telic verbs do not entail completion in the progressive form.

distinction is critical when processing temporal events. A faithful reasoning agent must distinguish between an action intended (or in progress) and an action realized.

Consider the following sentence: “*The carpenter was building a gazebo*”, as shown in Table 1. Logically, this description of an ongoing process does not entail that the gazebo was ever finished. The construction could have been abandoned or interrupted. This lack of entailment for telic (goal-oriented) events is known in formal semantics as the **Imperfective Paradox** (Dowty, 1977, 1979). In contrast, for atelic verbs like “*run*”, the process itself constitutes the event (Vendler, 1957; Bennett and Partee, 1978; Smith, 1997): “*The boy was running*” entails “*The boy ran*”.

This phenomenon is grounded in the theory of (**Lexical**) **Aspect**, which categorizes events by their telicity (boundedness) (Vendler, 1957; Comrie, 1976; Leiss, 1992). Unlike atelic activities where the process implies the event (e.g., running), telic accomplishments possess a distinct culmination point that separates the process from the result (Moens and Steedman, 1988). Therefore, the progressive aspect acts as a logical filter: it entails realization for activities but suspends it for accomplishments. A robust reasoner must respect this structural boundary, distinguishing between the occurrence of a process and the attainment of a goal.

While humans intuitively navigate these aspectual distinctions, we hypothesize that LLMs exhibit a **Teleological Bias**: a tendency to assume that

goal-directed actions inevitably lead to their successful completion. This bias likely stems from the “reporting bias” in pre-training corpora (Gordon and Van Durme, 2013), where narratives typically mention goals only when they are relevant to the outcome, often implying success (Grice, 1975). Consequently, models may hallucinate event completion based on statistical likelihood rather than logical entailment, effectively conflating the intent with the result (Farquhar et al., 2024).

Prior work in NLP has largely focused on aspect classification or modeling, training models to label verbs as stative, dynamic, telic, or atelic (see Friedrich et al., 2023). However, high classification accuracy does not guarantee robust inference. It remains unexplored whether generative models can apply these aspectual properties to resist fallacious entailments in zero-shot reasoning scenarios.

To bridge this gap, we introduce IMPERFECTIVENLI, a diagnostic dataset to probe the Imperfective Paradox. We construct minimal pairs across four logical conditions, contrasting interrupted and ambiguous contexts for both telic and atelic verbs. We evaluate state-of-the-art open-weight models and uncover following critical insights:

- 1) **Teleological Bias Dominates Model Inference:** Models exhibit a pervasive tendency to hallucinate completion for goal-oriented events. Crucially, this bias is robust enough to override explicit textual negation (e.g., predicting *completion* even when an action is interrupted), revealing a fundamental failure in *contextual fidelity*.
- 2) **Prompting Interventions Trigger a Calibration Crisis:** Strategies like Chain-of-Thought or forced counterfactual non-endpoints can reduce hallucinated completions, but often over-correct, causing models to incorrectly reject valid entailments for atelic activities. Models struggle to adapt reasoning dynamically based on event type.
- 3) **Representation and Reasoning Are Dissociated:** We find that the bias is not representational; embeddings successfully distinguish *process* (progressive form) from *result* (simple past form). Instead, inference is dominated by prior expectations about goal completion. The failure is likely driven by predictive priors during decoding, rather than an inherent inability of Transformer architectures to encode the aspectual distinction.

Ultimately, our findings underscore that current LLMs struggle to resolve the Imperfective Paradox, operating as probabilistic predictors of narrative outcomes rather than faithful logical reasoners. Bridging this gap requires not just better prompting, but a deeper alignment between model behavior and the formal logic of event semantics.

## 2 Background

### 2.1 Vendler’s Aspectual Classes: Telicity

Vendler (1957) categorized verbs into four aspectual classes (activity, accomplishment, achievement, state) based on their temporal properties. In this study, we focus on the distinction between Activities and Accomplishments, defined by the property of *telicity*, whether an event involves an inherent endpoint. Telicity often correlates with perfective interpretations, which present an eventuality from a holistic perspective (Filip, 2008).

- **Activities (Atelic):** Events with no inherent endpoint (e.g., “run”, “swim”, “push a cart”). They are homogeneous; any part of the process counts as an instance of the same event.
- **Accomplishments (Telic):** Events that proceed toward a specific terminus or goal (e.g., “build a house”, “write a letter”). Such events are only completed and receive canonical perfective interpretations, when the endpoint is reached.

### 2.2 The Imperfective Paradox

Semantically, in English, the progressive form has imperfective reading, while the simple past form has perfective reading (Comrie, 1976; Smith, 1997). The Imperfective Paradox arises when we analyze the logical entailment between the progressive form (Imperfective) and the simple past form (Perfective) of these verb classes. Let  $\phi$  be an event predicate and  $PROG(\phi)$  denote its progressive form.

For Atelic predicates (Activities), the entailment holds:

$$PROG(\text{run}) \models \text{run}$$

If  $x$  was running, then it is true that  $x$  ran. This is often referred to as the Sub-interval Property: if an activity holds true for an interval, it holds true for any sub-interval (Bennett and Partee, 1978).

For Telic predicates (Accomplishments), the entailment fails:

$$PROG(\text{build a house}) \not\models \text{build a house}$$

If  $x$  was building a house, it is NOT necessarily true that  $x$  built a house.

The paradox highlights that the progressive form of an accomplishment describes a process leading to a result, but does not assert the result itself (Parsons, 1990). Dowty (1977) formalized this using possible worlds semantics, suggesting that  $PROG(\phi)$  is true if  $\phi$  completes in all “inertia worlds” (worlds where the course of events proceeds without interruption). However, the real world is not necessarily an inertia world and interruptions occur.

### 2.3 Related Work: Interpreting and Modeling “Aspect” in NLP

Computational modeling of aspect has evolved from feature-engineering approaches to representation learning (Friedrich et al., 2023). Siegel and McKeown (2000) pioneered the use of supervised machine learning for aspectual classification. Subsequent work extensively explored semi-supervised approaches, combining linguistic and distributional features to predict stativity, duration, and situation entity types, while also establishing essential annotated datasets (Friedrich and Palmer, 2014; Friedrich and Pinkal, 2015; Friedrich et al., 2016; Friedrich and Gateva, 2017). Similar methodologies were applied to German verbs (Hermes et al., 2015) and multilingual morphosyntactic annotation (Ramm et al., 2017).

Moving toward compositional semantics, Kober et al. (2020) introduced an NLI-based dataset to model aspect in context. Later work showed that transformer models capture some aspectual cues but rely heavily on surface features such as tense and word order (Metheniti et al., 2022). More recently, Ma (2024) found that large-scale LLMs still struggle with lexical aspect identification. Closely related are two studies on the Imperfective Paradox (Pruš et al., 2024, 2025), which focus primarily on modeling human judgments of progressive versus simple past forms, treating LLM behavior as a secondary comparison. As such, neither study provides a systematic diagnostic evaluation of LLM aspectual inference. Our work directly addresses this gap, offering the first controlled investigation of whether LLMs respect the Imperfective Paradox or instead hallucinate event completion by conflating goal-directed processes with their result states.

## 3 Dataset - IMPERFECTIVENLI

To evaluate LLMs’ ability to reason about the imperfective paradox and teleological aspectual semantics, we introduce IMPERFECTIVENLI, a diagnostic NLI dataset in English.

### Dataset Construction with Different Verbs.

The construction of IMPERFECTIVENLI employs a template-based generation approach to ensure syntactic uniformity while isolating semantic reasoning. We curated a balanced lexicon of verbs based on Vendler’s aspectual classification:

- **100 Telic Verbs (Accomplishments):** Predicates involving a natural endpoint or result state, where the process is distinct from the culmination (e.g., *build, write, fix, paint, bake*).
- **100 Atelic Verbs (Activities):** Predicates describing homogeneous processes without a necessary endpoint, where the process itself constitutes the event (e.g., *run, swim, wander, play, vibrate*).

For each verb, we generated context-hypothesis pairs using Gemini-assisted rewriting (Team et al., 2025), followed by strict human verification. The dataset is structured as a controlled minimal-pair experiment across four logical conditions, resulting in a total of  $N = 400$  examples.

To investigate whether teleological bias varies across different types of goal-directedness, we further annotated the 100 telic verbs into four fine-grained semantic classes (drawn on the theory of Levin, 1993): *Change of State* (44), *Creation* (39), *Consumption* (9), and *Motion to Goal* (8).

We restricted this sub-classification to the telic group because, unlike homogeneous activities, telic predicates inherently encode a distinct *resultant state* (Beavers and Koontz-Garboden, 2012). This classification categorizes the nature of that result: *Change of State* verbs denote a transformation in the physical or abstract condition of an existing entity (e.g., *melt, break*); *Creation* verbs involve bringing an object into existence (e.g., *build, knit*); *Consumption* verbs entail the destruction or complete ingestion of an object (e.g., *eat, burn*) (Beavers, 2010); and *Motion to Goal* verbs describe movement towards a specific spatial endpoint (e.g., *arrive, enter*) (Talmy, 2000). This stratification enables us to probe whether LLMs are more prone to hallucinating completion for artifact-creating events, where the “teleological” purpose is most salient, compared to simple state changes.

Group	Condition	Premise Example	Hypothesis	Gold Label
A	<b>Interrupted Accompl.</b> (Telic + Cancel)	The carpenter was building a gazebo, <i>but a storm destroyed the frame before the roof was on.</i>	The carpenter built a gazebo.	Contradiction ( <b>X</b> False)
B	<b>Interrupted Activity</b> (Atelic + Stop)	The athletes were running on the track, <i>but it started to hail.</i>	The athletes ran on the track.	Entailment ( <b>✓</b> True)
C	<b>Ambiguous Accompl.</b> (Telic + Process)	The carpenter was building a gazebo.	The carpenter built a gazebo.	Neutral ( <b>?</b> Unknown)
D	<b>Ambiguous Activity</b> (Atelic + Process)	The athletes were running on the track.	The athletes ran on the track.	Entailment ( <b>✓</b> True)

Table 2: Examples from the IMPERFECTIVENLI dataset. Group C is the critical probe, where the imperfective paradox implies a *Neutral* relation, but teleological bias may lead models to predict *Entailment*.

**Imperfective Logical Conditions.** The core contribution of IMPERFECTIVENLI is its minimal-pair design, which can be viewed as a  $2 \times 2$  matrix crossing **Verb Telicity** (Telic vs. Atelic) with **Contextual Information** (Interrupted vs. Ambiguous), described in the following groups A-D. Table 2 illustrates these conditions with examples.

#### Group A: Interrupted Accomplishment.

This group serves as a counterfactual check. The premise describes a telic action in progress but includes an explicit cancellation clause (e.g., “...*but a storm destroyed the frame before the roof was on.*”). This tests whether models can attend to explicit interruption of action and the negation of the result. The logical relation is **Contradiction**.

#### Group B: Interrupted Activity.

This group controls for aspectual class properties. Due to the *sub-interval property* of atelic verbs, stopping an activity does not negate its occurrence (i.e., stopping running implies one has run). Even with an explicit interruption, the premise entails the hypothesis. The logical relation is **Entailment**.

#### Group C: Ambiguous Accomplishment.

This condition is the primary testbed for the Imperfective Paradox. Structurally, Group C is an ablation of Group A, where the interruption clause is removed, leaving only the progressive form of a telic verb. Semantically, the outcome is under-specified: the action was in progress, but completion is not known. The logical relation is **Neutral**. This group probes whether models hallucinate completion in the absence of explicit evidence.

#### Group D: Ambiguous Activity.

Analogous to Group C, this condition is an ablation of Group B, retaining only the progressive form of an activity. Unlike Group C, the entailment holds here because the atelic process itself validates the event. This baseline tests whether models possess basic aspectual competence for atelic verbs. The logical relation is **Entailment**.

While standard NLI tasks utilize the labels *Entailment*, *Contradiction*, and *Neutral* (e.g., Bowman et al., 2015), we map these to the natural

language truth values “**True**”, “**False**”, and “**Unknown**” respectively for our experiments.

**Quality Assurance.** To ensure the validity and naturalness of the generated data, we recruited three native English speakers as evaluators via *Prolific*<sup>2</sup>. They assessed the sentences across three linguistic dimensions, Grammar, Fluency, and Adequacy (details in Appendix C), and were compensated at a rate of £9/hour. Beyond quantitative ratings, evaluators provided corrections for any unnatural or unclear instances. Discrepancies and conflicting revisions were resolved through manual verification, ensuring quality in the final dataset.

## 4 Experimental Settings

### 4.1 Models

We experimented with 7 instruction-tuned open-weight models with their parameter around 7-9B: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Qwen2.5-7B-Instruct (Qwen et al., 2025), deepseek-llm-7b-chat (DeepSeek-AI, 2024), gemma-2-9b-it (Team et al., 2024), glm-4-9b-chat-hf (GLM et al., 2024), Yi-1.5-9B-Chat (AI et al., 2025).

### 4.2 Prompts

To disentangle the model’s intrinsic biases from its instruction-following capabilities, we evaluate four distinct prompting strategies. The detailed prompts given to the models are presented in Appendix B.

**(1) Strict Logic Prompt (Zero-Shot).** Our primary setting employs a simple zero-shot instruction, framing the model as a “strict logician”. This prompt explicitly prohibits common sense assumptions (e.g., the pragmatic inference that “building

<sup>2</sup><https://www.prolific.com/>

usually implies a house”) and forces the model to rely solely on the provided text.

**(2) Definition-Aware Prompt (DAP).** To test whether the failure stems from a lack of linguistic *knowledge* or a failure of *application*, we explicitly inject the linguistic definition of the imperfective aspect for both Activity and Accomplishment into the system prompt. If the bias persists even when the rule is provided, it suggests that the teleological prior is robust enough to override in-context instructions.

**(3) Chain-of-Thought Prompt (CoT).** We adapt the standard “Let’s think step by step” paradigm (Wei et al., 2022b). This setting investigates whether the Teleological Bias is a result of fast, superficial processing. We hypothesize that by forcing the model to articulate the temporal status of the event before predicting the label, it may self-correct the hallucination for Ambiguous Accomplishment.

**(4) Counterfactual Prompt (Counterfactual).** Drawing on Dowty (1977)’s theory that the progressive aspect presupposes an “inertia world” (where events proceed normally), this prompt attempts to explicitly break this assumption. We explicitly instruct the model to first generate three plausible real-world scenarios in which the event is interrupted and its endpoint is not reached, before making a judgment.

### 4.3 Metrics

We report following metrics:

**(1) Group-wise Accuracy (ACC):** We calculate classification accuracy separately for each logical condition (Groups A–D).

**(2) Teleological Bias Rate (TBR):** This metric specifically quantifies the “Hallucination of Completion” in Group C (Ambiguous Accomplishment). Since the Gold Label for Group C is *Unknown*, any prediction of *True* indicates that the model has hallucinated a result. We define TBR as:

$$\text{TBR}_C = \frac{\sum_{i \in C} \mathbb{I}(\hat{y}_i = \text{True})}{|C|} \quad (1)$$

where  $C$  is the set of examples in Group C, and  $\hat{y}_i$  is the model’s prediction. A higher TBR indicates a stronger reliance on teleological priors.

**(3) Aspectual Awareness Gap ( $\Delta_{AA}$ ):** This metric measures the model’s capability to distinguish

the logical implications of *atelic* processes (which imply realization) from *telic* processes (which do not). It is calculated as the difference between the accuracy on the Ambiguous Activity baseline (Group D) and the error rate on the Ambiguous Accomplishment probe (Group C):

$$\Delta_{AA} = \text{ACC}_D - \text{TBR}_C \quad (2)$$

where  $\text{ACC}_D$  represents the correct entailment rate for Activity verbs. A high gap indicates ideal aspectual reasoning: the model correctly infers entailment for activities while correctly suspending judgment (low bias) for accomplishments. A low gap suggests that the model treats all progressive verbs homogeneously, likely applying a shallow heuristic (e.g., “mentioning an action implies its completion”) regardless of lexical aspect.

Note that  $\Delta_{AA}$  is most interpretable when  $\text{ACC}_D$  is high; a low  $\Delta_{AA}$  driven by a low  $\text{ACC}_D$  rather than a high  $\text{TBR}_C$  reflects a distinct failure mode, namely, an inability to recognize valid entailments for atelic verbs, rather than teleological overgeneralization.

## 5 Results and Analysis

In this section, we present the evaluation results on IMPERFECTIVENLI. We aim to answer four key questions: whether LLMs exhibit teleological bias, how prompting strategies mitigate this bias, how the model scale affects the bias, and how these biases manifest across different semantic categories.

### 5.1 The Dominance of Teleological Bias for “If A Subject was V-ing, then the Subject V-ed” in Zero-shot

The first few rows of Table 3 present the performance of seven open-weight models in the zero-shot setting. The results reveal a fundamental inability of current LLMs to decouple the description of a process from the assertion of its result.

**$P(\text{Complete} \mid \text{Past Progressive}) \approx 1?$  – – The Illusion of Competence.** At first glance, models appear competent in aspectual reasoning, achieving near-perfect accuracy on Group D for ambiguous Activity (atelic verbs) (e.g., Llama-3.1: 0.98, Mistral: 1.00). However, this high performance is deceptive. When contrasted with the catastrophic failure on Group C for ambiguous Accomplishment (telic verbs) (e.g., Llama-3.1: 0.02), a different picture emerges. The near-zero  $\Delta_{AA}$  for most

Prompt	Model	Acc A (↑)	Acc B (↑)	Acc C (↑)	Acc D (↑)	TBR <sub>C</sub> (↓)	$\Delta_{AA}$ (↑)
Zero-shot	Llama-3.1-8B-Instruct	0.47	0.85	0.02	0.98	0.98	0.00
	Mistral-7B-Instruct-v0.3	0.37	0.92	0.02	1.00	0.97	0.03
	Qwen2.5-7B-Instruct	0.20	0.98	0.47	0.97	0.53	0.44
	Yi-1.5-9B-Chat	0.35	0.94	0.02	1.00	0.98	0.02
	deepseek-llm-7b-chat	0.04	0.88	0.00	1.00	1.00	0.00
	gemma-2-9b-it	0.03	0.96	0.06	1.00	0.94	0.06
	glm-4-9b-chat-hf	0.14	0.98	0.03	1.00	0.97	0.03
DAP	Llama-3.1-8B-Instruct	0.60	0.95	0.36	0.99	0.45	0.54
	Mistral-7B-Instruct-v0.3	0.55	0.96	0.06	1.00	0.94	0.06
	Qwen2.5-7B-Instruct	0.35	0.97	0.89	0.72	0.09	0.63
	Yi-1.5-9B-Chat	0.57	0.97	0.20	1.00	0.80	0.20
	deepseek-llm-7b-chat	0.12	0.92	0.00	1.00	1.00	0.00
	gemma-2-9b-it	0.08	0.90	0.60	0.88	0.40	0.48
	glm-4-9b-chat-hf	0.08	1.00	0.33	0.98	0.67	0.31
CoT	Llama-3.1-8B-Instruct	0.15	0.39	0.67	0.65	0.33	0.32
	Mistral-7B-Instruct-v0.3	0.23	0.65	0.41	0.66	0.56	0.10
	Qwen2.5-7B-Instruct	0.30	0.80	0.97	0.37	0.03	0.34
	Yi-1.5-9B-Chat	0.56	0.81	0.85	0.44	0.13	0.31
	deepseek-llm-7b-chat	0.05	0.88	0.40	0.94	0.60	0.34
	gemma-2-9b-it	0.15	0.88	0.98	0.15	0.02	0.13
	glm-4-9b-chat-hf	0.24	0.95	0.38	0.97	0.62	0.35
Counterfactual	Llama-3.1-8B-Instruct	0.35	0.00	0.97	0.00	0.00	0.00
	Mistral-7B-Instruct-v0.3	0.28	0.39	0.87	0.19	0.11	0.08
	Qwen2.5-7B-Instruct	0.07	0.57	1.00	0.02	0.00	0.02
	Yi-1.5-9B-Chat	0.07	0.41	1.00	0.10	0.00	0.10
	deepseek-llm-7b-chat	0.25	0.97	0.00	0.98	1.00	-0.02
	gemma-2-9b-it	0.03	0.00	1.00	0.00	0.00	0.00
	glm-4-9b-chat-hf	0.23	0.66	0.73	0.24	0.20	0.04

Table 3: Model performance under Zero-shot, CoT, DAP, and Counterfactual prompting strategies with Acc for A-D Groups, Teleological Bias Rate (TBR), and Aspectual Awareness Gap ( $\Delta_{AA}$ ).

models indicates that they do not genuinely understand the *Sub-interval Property* of the atelic verbs (Bennett and Partee, 1978). Instead, they likely employ a shallow heuristic: “If Subject was V-ing, then Subject V-ed”, regardless of verbs in the Zero-shot setting. This aligns with findings by the NLI work McCoy et al. (2019), suggesting that the models might often rely on lexical overlap strategies rather than deep semantic compositionality.

**Contextual Neglect via Semantic Priming.** Perhaps the most alarming finding is the resistance to explicit negation in Group A (Interrupted Accomplishment). Despite the textually grounded interruption (e.g., “...but a storm destroyed it”), most models exhibit a stark performance degradation. This stands in sharp contrast to Group B (Interrupted Activity), where models maintain high accuracy. This performance gap is semantically revealing: for activities, an interruption does not negate the event’s occurrence (the *sub-interval property* ensures that “stopped running” entails “ran”), whereas for accomplishments, interruption strictly precludes the result.

The failure in Group A suggests that the models’ reasoning is driven by **Semantic Priming** rather

than compositional logic (Ettinger, 2020; Kassner and Schütze, 2020). We hypothesize that the co-occurrence of an agent (e.g., *carpenter*) and a telic verb (e.g., *building*) activates the “completed result” representation in the model’s high-dimensional space with such magnitude that it overrides the attention mechanism’s focus on the subsequent cancellation clause. This phenomenon exemplifies a failure of contextual fidelity, where static parametric knowledge priors overpower the dynamic input context (Mallen et al., 2023).

**The Outlier: Qwen2.5 shows exceptional performance.** Qwen2.5 stands as a notable exception, maintaining a balanced performance with a TBR of 0.53 and a healthy  $\Delta_{AA}$  of 0.44. Unlike its peers, Qwen demonstrates an emerging ability to suspend judgment in ambiguous contexts.

## 5.2 Trade-off: Bias Mitigation vs. Semantic Calibration in Advanced Prompts

Evaluating three advanced prompting strategies reveals a consistent trade-off between mitigating teleological bias and maintaining semantic calibration across aspectual classes.

**DAP: The Knowledge-Application Gap.** The DAP which introduces expert linguistic definition for the aspectual classes yields mixed results, exposing a critical gap between possessing linguistic knowledge and applying it. While supplying the linguistic rule improves Llama-3.1’s accuracy on Group C (0.02  $\rightarrow$  0.36), it remains ineffective for Mistral (0.06). This persistence of bias implies that the teleological prior is not a misunderstanding of the task definition but a robust feature of the model’s probability distribution. As noted by Dziri et al. (2023), compositional reasoning in Transformers often degrades when it conflicts with high-frequency statistical patterns (i.e., the reporting bias) seen during pre-training.

**CoT: Reasoning-Induced Uncertainty.** CoT prompting successfully reduces the TBR for most models, proving that allocating compute to intermediate reasoning helps suppress System-1 heuristics (Wei et al., 2022b). However, this comes at a cost. We observe a significant degradation in Group D (Activities), where accuracy drops substantially (e.g., Llama-3.1 0.98  $\rightarrow$  0.65). Qualitative analysis reveals a phenomenon of “over-reasoning” or unfaithful reasoning (Turpin et al., 2023). When forced to analyze “step-by-step”, models often hallucinate hypothetical interruptions for simple activities.<sup>3</sup> CoT appears to over-generalize the progressive-as-incomplete heuristic: when forced to reason step-by-step, models often impose hypothetical interruptions even on atelic verbs, inadvertently suppressing valid entailments.

**Counterfactuals: The Calibration Crisis.** The Counterfactual prompt represents the most aggressive intervention for arbitrary interruption. While highly effective at curing teleological bias (Group C Acc  $\rightarrow$  1.00), it induces an obvious calibration bias. Accuracy on Group D collapses to near zero for Llama-3.1. We show the example of Llama-3.1 in Figure 1. The models swing from being “naively optimistic” (Zero-shot) to “paranoidly skeptical” (Counterfactual). The Counterfactual clearly over-instructed the models to follow the imperfective patterns of the telic verbs in Group C, resulting in misunderstanding for the atelic verbs in Group D. The models failed to dynamically adjust their reasoning strategy based on

<sup>3</sup>See, e.g., the activity “running”, with model reasoning: “The premise describes a process in the present tense, indicating that the action of the athletes running on the track is ongoing or at least not explicitly completed.”

the (a)telicity of verbs, resulting in a wholesale rejection of progressive entailments. This highlights a lack of (aspectual) calibration (Kadavath et al., 2022; Kapoor et al., 2024): while the counterfactual prompt successfully suppresses teleological bias for telic verbs, models fail to modulate this skepticism based on verb class, applying it indiscriminately to atelic verbs as well. Detailed analyses and visualization for other models are provided in Appendix E.

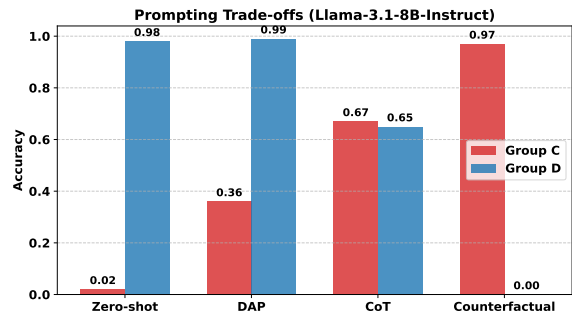


Figure 1: The trade-off in prompting (Llama-3.1). As prompts become more aggressive in targeting bias (Zero-shot  $\rightarrow$  Counterfactual), accuracy on Group C improves significantly (Pearson  $r = 1.00, p = 0.00$ ), while accuracy on Group D declines, though less consistently ( $r = -0.91, p = 0.09$ ). This trade-off illustrates the calibration bias.

**Summary on Different Prompts.** In summary, current LLMs struggle to balance the Imperfective Paradox. They oscillate between two extremes: **naive teleology** (Zero-shot) where everything is completed, and **paranoid skepticism** (Counterfactual) where nothing is ever certain (see the trends in Figure 1). This inability to decouple *telicity* from *entailment* highlights a critical gap in the true aspectual reasoning capabilities of LLMs.

### 5.3 Impact of Model Scale

To understand whether aspectual reasoning is a capability that improves linearly with scale or emerges as a distinct skill past a certain threshold, we evaluate the Qwen2.5 family, which showed better performance in §5.1, across five distinct parameter sizes: 1.5B, 7B, 14B, 32B, and 72B. The results, detailed in Table 5 (Appendix) and visualized in Figure 2, reveal that aspectual reasoning follows a distinct trajectory of emergence.

We observe a clear monotonic relationship between model scale and both evaluation metrics. Model size is strongly negatively correlated with Teleological Bias Rate (TBR;  $r = -0.93, p =$

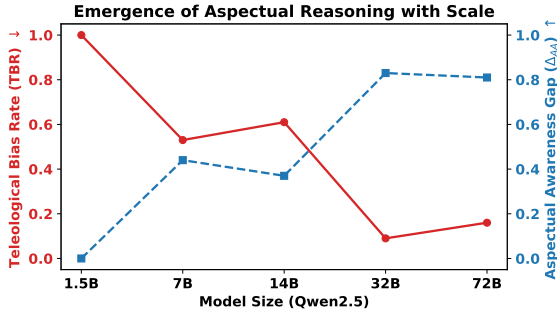


Figure 2: The emergence of aspectual reasoning capabilities across model scales. It shows a significant decrease in TBR and a corresponding increase in  $\Delta_{AA}$ .

0.02) and strongly positively correlated with the Aspectual Awareness Gap ( $\Delta_{AA}$ ;  $r = 0.94, p = 0.01$ ), indicating that increased model capacity systematically reduces teleological overgeneralization while enhancing aspect-sensitive reasoning.

The 1.5B model exhibits a complete inability to handle the Imperfective Paradox. With a TBR of 1.00 and a  $\Delta_{AA}$  of 0.00, it treats every progressive verb as an entailed completion. As model scale increases, a dramatic “phase shift” (Wei et al., 2022a) is observed, especially at the 32B level: Accuracy on Group C surges to 0.91, and the  $\Delta_{AA}$  peaks at 0.83. This suggests that robust aspectual distinction, the ability to selectively inhibit the entailment for telic verbs while maintaining it for atelic ones, is an emergent ability. Although the 72B model maintains strong performance,  $\Delta_{AA}$  decreases slightly to 0.81, suggesting a potential saturation or optimization bottleneck rather than further qualitative gains. Overall, these results indicate that aspectual understanding improves with scale and exhibits non-linear emergence characteristics rather than linear scaling.

#### 5.4 Fine-grained Analysis on Verb Semantics

Do all telic verb classes trigger the same degree of teleological bias? To answer this, we disaggregate the performance on Group A and Group C based on the four semantic sub-classes defined in §3: *Creation*, *Consumption*, *Change of State*, and *Motion to Goal*. Figure 3 visualizes the average accuracy across all evaluated models in the zero-shot setting. Detailed results for each model as well as for other prompt settings can be found in Appendix E.

**The Resilience of Motion vs. The Creation Penalty.** We observe a significant semantic divergence. Verbs of *Motion to Goal* (e.g., “arrive”, “en-

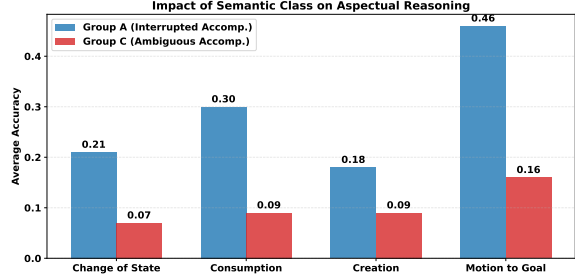


Figure 3: Average accuracy across models by semantic verb classes in Zero-shot. **Motion verbs** are consistently easier for models to reason about, while **Creation verbs** induce strong teleological bias (low accuracy in Group C) and resistance to contextual cancellation (lowest accuracy in Group A).

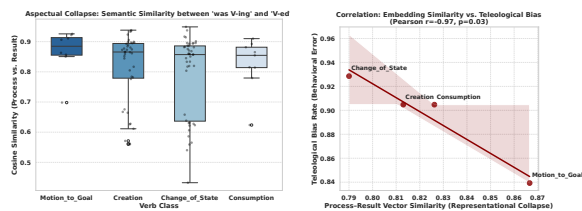
ter”) consistently yield the highest performance. In Group A, the average accuracy for *Motion to Goal* verbs is 46%, significantly outperforming *Creation* (18%) and *Change of State* (21%). The performance gap is substantial:  $\Delta_{Motion-creation} \approx +28\%$  in Group A and  $\approx +13\%$  in Group C.<sup>4</sup>

We hypothesize that this stems from the nature of the resultative state. For *Creation* verbs (e.g., “build”), the result involves the existential realization of an object. This explicit “coming into being” appears to act as a stronger attractor in the language model’s predictive distribution than a mere change in spatial coordinates involved in *Motion*. Consequently, models exhibit higher resistance to negation when an artifact’s existence is at stake.

**Representational Analysis: Encoding vs. Inference.** Does teleological bias reflect a failure to distinguish *process* from *result* in the model’s internal representations, or does it arise later, at the point of inference? In order to probe this, we measured the cosine similarity between contextual embeddings of progressive (e.g., “was building”) and perfective (e.g., “built”) verb phrases using bert-base-uncased (Devlin et al., 2019), treating representational similarity as a proxy for how well the model’s encoder separates the two aspectual forms.

Figure 4 reveals a striking inverse relationship (Pearson  $r = -0.97, p = 0.03$ ) between how similarly a model encodes the progressive and perfec-

<sup>4</sup>We additionally conducted a two-way repeated-measures ANOVA for Group A and Group C across models and semantic verb classes. The results revealed a significant main effect of verb class ( $F(3, 18) = 6.98, p < 0.01$ ). Crucially, we observed a significant interaction between group and verb class ( $F(3, 18) = 3.88, p < 0.05$ ), indicating that the impact of ambiguity varies across semantic classes.



(a) Cosine similarity (Process vs. Result).

(b) Similarity vs.  $TBR_C$  (Zero-shot).

Figure 4: Representational vs. Behavioral Divergence.

tive forms of a verb, and how strongly it hallucinates completion for that verb class. Crucially, this relationship runs in the opposite direction from what a simple confusion account would predict. *Motion to Goal* verbs show the highest representational similarity (median  $\approx 0.88$ ), meaning the model encodes “*was flying*” and “*flew*” as nearly identical; yet models reason about them most accurately (with  $TBR_C$  being the lowest). *Creation* verbs show slightly lower similarity (median  $\approx 0.85$ ), meaning the model’s encoder does distinguish “*was building*” from “*built*”; yet these verbs trigger higher hallucination rates (with  $TBR_C$  being higher).

This dissociation has a clear implication: the bias does not stem from the model failing to encode the aspectual distinction. If it did, we would expect high similarity to correlate with high error, but the opposite holds. Instead, even when the internal representation correctly separates process from result, the inference step is overridden by a strong prior that *Creation* events reach their goal. The model “knows” the difference between “*was building*” and “*built*”, but still predicts completion, because its training has instilled a robust expectation that construction projects succeed. Teleological Bias is therefore not a failure of *perception* (encoding), but a failure of *reasoning* (decoding): the model’s decoding process is captured by world-knowledge priors rather than guided by the logical content of the input.

## 6 Conclusion

In this work, we evaluated and discovered the **Teleological Bias** in LLMs using our diagnostic dataset, IMPERFECTIVENLI. Our findings indicate that current models fundamentally struggle to resolve the **Imperfective Paradox**, frequently conflating the *process* of a telic event with its *culmination*. We demonstrate that this bias is not merely a represen-

tational deficit but a decision-time failure: while internal embeddings often distinguish between process and result, strong probabilistic priors regarding goal attainment override these signals during inference. Furthermore, our investigation into prompting strategies reveals a critical calibration crisis, finding that while some strategies reduce overconfident or hallucinated predictions, they can also make models overly cautious, causing them to reject correct entailments. We conclude that solving aspectual reasoning requires moving beyond inference-time prompting toward aligning models to structurally respect the internal logical boundaries of event semantics. We hope that IMPERFECTIVENLI provides a foundation for future work on targeted alignment training, cross-lingual aspectual reasoning, and the broader challenge of teaching LLMs to respect formal semantic boundaries.

## 7 Limitations

While our work offers important insights into the imperfective paradox with the teleological bias of LLMs, we acknowledge several limitations that define the scope of our findings and point towards future research directions.

**Ecological Validity and Syntactic Diversity.** To strictly isolate the variable of aspectual telicity, we employed a template-based approach to construct IMPERFECTIVENLI. While this ensures high internal validity and precise minimal-pair comparisons, it sacrifices lexical and syntactic diversity. Real-world language usage often includes complex discourse markers, temporal adjuncts (e.g., “*for an hour*” vs. “*in an hour*”), and pragmatic cues that constrain interpretation (Zhang, 1995). Consequently, our controlled setup likely overestimates model competence relative to naturalistic settings; handling the imperfective paradox in messier, more complex narratives may prove considerably more challenging.

**Linguistic Specificity.** Our study focuses exclusively on English, a language that marks the progressive aspect periphrastically (via auxiliary “*be*” + “*-ing*”). However, aspect is a typologically diverse category. For instance, Slavic languages mark aspect morphologically (aspect prefixes), while Mandarin Chinese uses aspectual particles. Since LLM tokenizers and representations vary across languages, it remains an open question whether the teleological bias is a universal cognitive artifact

of Transformers or an artifact of English-centric pre-training data. Meanwhile, in human cognitive studies, research demonstrates that the conceptualization of event semantics differs fundamentally between aspect languages (those with grammaticalized aspect markers, e.g., English, Chinese, Russian) and non-aspect languages (e.g., German) (von Stutterheim et al., 2012; Flecken et al., 2014). Consequently, extending this investigation to a cross-lingual and typological framework represents a vital direction for future research.

### **Theoretical Linguistics vs. Human Judgments.**

Since our paper focuses on the theoretical concepts of imperfective paradox, we note potential divergence between theoretical labels and empirical human judgments. As the first study systematically investigating the Imperfective Paradox in LLMs, our primary objective was to characterize the teleological bias, the models’ tendency to hallucinate completion, based on established theoretical semantics. While we focus here on establishing this bias in models, we wish to note that exploring the fine-grained nuances of human perception in these contexts is a vital next step.

**Inference-Only Interventions.** Our mitigation strategies were limited to inference-time prompting (CoT, Counterfactuals). While we identified the performance trade-offs as a side effect of these prompts, we did not explore parameter-efficient fine-tuning (PEFT) or activation steering techniques to permanently alter the model’s internal aspectual representations. Future work should investigate whether the “dissociation” between representation and reasoning can be bridged through targeted alignment training rather than transient prompting.

**Scope of World Knowledge.** Finally, our semantic analysis focused on four broad classes of telic verbs. We did not model the granularity of “script knowledge” (Schank and Abelson, 1977; Lyu et al., 2021). For example, the probability of interruption varies by event type (e.g., “*building a house*” takes months and is prone to interruption, whereas “*drawing a circle*” takes seconds). Current models may rely on these latent temporal priors, which our uniform evaluation setup does not explicitly disentangle from grammatical aspect.

## **8 Ethical Considerations**

**Human Subjects for Data Evaluation.** For the human evaluation of sentence quality, we recruited external participants via the *Prolific* platform. Prior to the task, annotators were fully informed about the study’s scope and provided informed consent. We ensured that no personally identifiable information (PII) was collected. Evaluators were compensated at an hourly rate of £9, exceeding the platform’s recommended minimum and comparable to the local living wage.

**Artifact Usage.** Regarding the curated artifacts, we release the IMPERFECTIVENLI dataset under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). This license explicitly permits the distribution and use of the dataset for research purposes. We confirm that our use of these existing artifacts is consistent with their intended academic use, and our derivative dataset is released with the strict specification that it be used solely within research contexts, ensuring compatibility with the original access conditions. The dataset contains no harmful or offensive content.

**Use of AI Assistants.** As described in Section 3, we used Gemini to assist in the initial generation and curation of the dataset. To ensure data quality and mitigate potential machine artifacts, all generated instances underwent rigorous manual verification and refinement by three native English speakers (see validation details in Section 3 and Appendix C).

Regarding the manuscript and experimental code, we acknowledge the use of ChatGPT solely for grammatical correction, stylistic polishing, and assistance with preliminary coding routines. The authors verified all AI-assisted outputs and bear full responsibility for the accuracy and originality of the content presented in this work.

### **Acknowledgments**

This publication was supported by LMUexcellent, funded by the Federal Ministry of Research, Technology and Space (BMFTR) and the Free State of Bavaria under the Excellence Strategy of the German Federal Government and the Länder.

We also thank Max Müller-Eberstein for valuable discussions and insightful exchanges related to this project.

## References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, and 14 others. 2025. [Yi: Open foundation models by 01.ai](#). *Preprint*, arXiv:2403.04652.
- John Beavers. 2010. [The structure of lexical meaning: Why semantics really matters](#). *Language*, 86(4):821–864.
- John Beavers and Andrew Koontz-Garboden. 2012. [Manner and result in the roots of verbal meaning](#). *Linguistic Inquiry*, 43(3):331–369.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Michael Bennett and Barbara Hall Partee. 1978. *Toward the Logic of Tense and Aspect in English*. IU Linguistics Club. Indiana University Linguistics Club.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. [MTG: A benchmark suite for multilingual text generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States. Association for Computational Linguistics.
- Bernard Comrie. 1976. *Aspect*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.
- DeepSeek-AI. 2024. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- David R. Dowty. 1977. [Toward a semantic analysis of verb aspect and the english ‘imperfective’ progressive](#). *Linguistics and Philosophy*, 1(1):45–77.
- David R. Dowty. 1979. *Word Meaning and Montague Grammar*, 1 edition. Studies in Linguistics and Philosophy. Springer Dordrecht.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. [Faith and fate: Limits of transformers on compositionality](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 70293–70332. Curran Associates, Inc.
- Allyson Ettinger. 2020. [What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models](#). *Transactions of the Association for Computational Linguistics*, 8:34–48.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarín Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630(8017):625–630.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. [High agreement but low kappa: I. the problems of two paradoxes](#). *Journal of Clinical Epidemiology*, 43(6):543–549.
- Hana Filip. 2008. [Events and maximalization: The case of telicity and perfectivity](#). In Susan Deborah Rothstein, editor, *Theoretical and Crosslinguistic Approaches to the Semantics of Aspect*, pages 110–217. John Benjamins.
- Monique Flecken, Christiane von Steutterheim, and Mary Carroll. 2014. [Grammatical aspect influences motion event perception: findings from a cross-linguistic non-verbal recognition task](#). *Language and Cognition*, 6(1):45–78.
- Annemarie Friedrich and Danyana Gateva. 2017. [Classification of telicity using cross-linguistic annotation projection](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565, Copenhagen, Denmark. Association for Computational Linguistics.
- Annemarie Friedrich and Alexis Palmer. 2014. [Automatic prediction of aspectual class of verbs in context](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland. Association for Computational Linguistics.

- Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. [Situation entity types: automatic classification of clause-level aspect](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.
- Annemarie Friedrich and Manfred Pinkal. 2015. [Automatic recognition of habituais: a three-way classification of clausal aspect](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, Lisbon, Portugal. Association for Computational Linguistics.
- Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2023. [A kind introduction to lexical and grammatical aspect, with a survey of computational approaches](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 599–622, Dubrovnik, Croatia. Association for Computational Linguistics.
- Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Jonathan Gordon and Benjamin Van Durme. 2013. [Reporting bias and knowledge acquisition](#). In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13*, page 25–30, New York, NY, USA. Association for Computing Machinery.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- H. Paul Grice. 1975. [Logic and conversation](#). *Syntax and Semantics*, 3:41–58.
- Kilem Li Gwet. 2008. [Computing inter-rater reliability and its variance in the presence of high agreement](#). *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.
- Jürgen Hermes, Michael Richter, and Claes Neufind. 2015. [Automatic induction of german aspectual verb classes in a distributional framework](#). In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. [Large language models must be taught to know what they don't know](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 85932–85972. Curran Associates, Inc.
- Nora Kassner and Hinrich Schütze. 2020. [Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.
- Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. [Aspectuality across genre: A distributional semantics approach](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Elisabeth Leiss. 1992. *Die Verbalkategorien des Deutschen*. De Gruyter, Berlin, Boston.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.
- Qing Lyu, Li Zhang, and Chris Callison-Burch. 2021. [Goal-oriented script construction](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 184–200, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Bolei Ma. 2024. [Evaluating lexical aspect with large language models](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 123–131, Bangkok, Thailand. Association for Computational Linguistics.
- Alex Mullen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. [About time: Do transformers learn temporal verbal aspect?](#) In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland. Association for Computational Linguistics.
- Marc Moens and Mark Steedman. 1988. [Temporal ontology and temporal reference](#). *Computational Linguistics*, 14(2):15–28.
- Terence Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press.
- Katarzyna Pruś, Mark Steedman, and Adam Lopez. 2024. [Human temporal inferences go beyond aspectual class](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1923, St. Julian’s, Malta. Association for Computational Linguistics.
- Katarzyna Pruś, Mark Steedman, and Adam Lopez. 2025. [World knowledge resolves some aspectual ambiguity](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13207–13220, Vienna, Austria. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Anita Ramm, Sharid Loáiciga, Annemarie Friedrich, and Alexander Fraser. 2017. [Annotating tense, mood and voice for English, French and German](#). In *Proceedings of ACL 2017, System Demonstrations*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.
- Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*, 1 edition. Psychology Press, New York.
- Eric V. Siegel and Kathleen R. McKeown. 2000. [Learning methods to combine linguistic indicators: improving aspectual classification and revealing linguistic insights](#). *Computational Linguistics*, 26(4):595–627.
- Carlota S. Smith. 1997. *The Parameter of Aspect*, 2 edition. Studies in Linguistics and Philosophy. Springer Dordrecht.
- Steven E. Stemler. 2004. [A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability](#). *Practical Assessment, Research, and Evaluation*, 9(1):4.
- Leonard Talmy. 2000. *Toward a Cognitive Semantics, Volume 1: Concept Structuring Systems*. The MIT Press.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. [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, pages 74952–74965. Curran Associates, Inc.
- Zeno Vendler. 1957. [Verbs and times](#). *The Philosophical Review*, 66(2):143–160.
- Christiane von Steutterheim, Martin Andermann, Mary Carroll, Monique Flecken, and Barbara Schmiedtová. 2012. [How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data, and memory performance](#). *Linguistics*, 50(4):833–867.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*. Survey Certification.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le,

and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

ChengYan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Yun Xue, and Barbara Plank. 2025. [M-ABSA: A multilingual dataset for aspect-based sentiment analysis](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557, Suzhou, China. Association for Computational Linguistics.

Lihua Zhang. 1995. [A Contrastive Study of Aspectuality in German, English, and Chinese](#). Peter Lang Verlag, New York, United States of America.

## A Verbs

We show here the telic verbs and atelic verbs probed in the IMPERFECTIVENLI and in the main experiment:

**Atelic verbs:** ['run', 'swim', 'walk', 'push', 'chat', 'dance', 'sing', 'laugh', 'weep', 'shout', 'play', 'wander', 'discuss', 'wait', 'study', 'work', 'exercise', 'browse', 'float', 'glide', 'slide', 'roll', 'snore', 'dream', 'meditate', 'tremble', 'sweat', 'bleed', 'shine', 'vibrate', 'hum', 'roar', 'bark', 'stare', 'listen', 'watch', 'cuddle', 'fight', 'argue', 'wrestle', 'garden', 'shop', 'patrol', 'chase', 'jog', 'march', 'ski', 'surf', 'ride', 'whistle', 'smile', 'frown', 'scream', 'yell', 'whisper', 'mumble', 'talk', 'speak', 'gossip', 'tease', 'help', 'serve', 'rule', 'live', 'breathe', 'cough', 'yawn', 'blink', 'nod', 'clap', 'shiver', 'twitch', 'ache', 'itch', 'glow', 'sparkle', 'fade', 'rot', 'bloom', 'grow', 'shrink', 'spin', 'rotate', 'travel', 'commute', 'roam', 'drift', 'flow', 'rain', 'snow', 'blow', 'graze', 'hunt', 'fish', 'cook', 'smoke', 'sip', 'chew', 'rub', 'scratch']

**Telic verbs:**

Change of State: ['fix', 'fill', 'clean', 'repair', 'empty', 'dry', 'wash', 'iron', 'sharpen', 'inflate', 'freeze', 'melt', 'cure', 'renovate', 'demolish', 'dig', 'pack', 'unlock', 'memorize', 'learn', 'organize', 'erase', 'delete', 'install', 'wrap', 'open', 'close', 'rescue', 'save', 'mow', 'sweep', 'mop', 'polish', 'sand', 'crack', 'cut', 'slice', 'chop', 'boil', 'fry', 'roast', 'toast', 'decorate', 'furnish']

Consumption: ['eat', 'drink', 'read', 'burn', 'devour', 'spend', 'destroy', 'solve', 'peel']

Creation: ['build', 'paint', 'bake', 'write', 'knit', 'compose', 'draw', 'film', 'code', 'sew', 'sculpt', 'weave', 'brew', 'forge', 'design', 'prove', 'translate', 'construct', 'invent', 'produce', 'plan', 'calculate', 'measure', 'count', 'transcribe', 'download', 'upload', 'print', 'plant', 'harvest', 'mix', 'equip', 'establish', 'form', 'shape', 'assemble', 'gather', 'collect', 'pick']

Motion to Goal: ['climb', 'cross', 'hike', 'fly', 'mail', 'deliver', 'catch', 'locate']

## B Prompts

In this section, we present the detailed prompt templates used across our four experimental conditions. The placeholders {premise} and {hypothesis} are dynamically replaced with the corresponding instances from the dataset. The specific instructions and output formatting constraints for each strategy are shown below.

### P1: Strict Logic Prompt

**System Prompt:** You are a strict logician. Your task is to determine if a Hypothesis is necessarily true given a Premise.

- If the Hypothesis **MUST** be true based only on the Premise, output **"True"**.
- If the Hypothesis is contradicted by the Premise, output **"False"**.
- If the Hypothesis might be true but is not explicitly guaranteed by the Premise, output **"Unknown"**.

Do not use common sense assumptions. Only use the text provided.

**User Input:** Premise: {premise} Hypothesis: {hypothesis}

Please respond with **ONLY** one of the following options: "True", "False", or "Unknown".

### P2: Definition-Aware Prompt

**System Prompt:** ... [Standard Logic Instructions] ...  
**Important Linguistic Rule:** Note the distinction between activity and accomplishment verbs.

- For **goal-oriented actions** (e.g., "was building"), the progressive form does NOT imply completion.
- For **activities** (e.g., "was running"), the process implies the action occurred.

**User Input:** Premise: {premise} Hypothesis: {hypothesis}  
 Please respond with ONLY one of the following options: "True", "False", or "Unknown".

### P3: Chain-of-Thought Prompt

**System Prompt:** [Same as Strict Logic Prompt]  
**User Input:** Premise: {premise} Hypothesis: {hypothesis}

**Instruction:** First, analyze the temporal status of the event in the premise. Does the action have a defined endpoint? Was it completed? Then, provide your final label.

**Output Requirement:** You must output your response in strict **JSON format** containing exactly two keys:

- "reasoning": A string explaining your step-by-step logic.
- "label": The final judgment ("True", "False", or "Unknown").

Do not include markdown formatting (like “`json”). Return ONLY the JSON object.

**Example Format:**  
 {"reasoning": "The premise describes a process...",  
 "label": "True" / "False" / "Unknown" }

### P4: Counterfactual Simulation Prompt

**System Prompt:** [Same as Strict Logic Prompt]  
**User Input:** Premise: {premise} Hypothesis: {hypothesis}

**Instruction:** 1. First, list 3 possible real-world scenarios where the action in the premise occurs but the result in the hypothesis is NOT achieved (e.g., interruptions, failures). 2. Based on these possibilities, determine if the hypothesis is *necessarily* true.

**Output Requirement:** You must output your response in strict **JSON format** containing exactly two keys:

- "reasoning": A string explaining your step-by-step logic.
- "label": The final judgment ("True", "False", or "Unknown").

Do not include markdown formatting (like “`json”). Return ONLY the JSON object.

**Example Format:**  
 {"possible\_interruptions": ["..."],  
 "label": "True" / "False" / "Unknown" }

## C Human Evaluation and Results

To validate the quality of the constructed dataset, we conducted a human evaluation on the *Prolific* crowdsourcing platform. We recruited 3 native English speakers to evaluate the generated sentences.

Following established protocols in machine translation and text generation research (Chen et al., 2022; Wu et al., 2025), evaluators were instructed to rate each sentence on a 5-point Likert scale across three linguistic dimensions: *Grammar*, *Flu-*

*ency*, and *Adequacy*. To ensure high-quality feedback, evaluators were also encouraged to provide corrections for any sentence rated 3 or lower. The specific instructions provided to the annotators are detailed in the following box:

### Instructions Provided to Prolific Evaluators

**Task Overview:** Please evaluate the English sentences provided in the attached Excel file. For each sentence, select a score from 1 to 5 using the dropdown menu. Please do not copy-paste values.

#### Rating Criteria:

- **Grammar:** Is the sentence grammatically correct? (5 = Perfect, no errors; 1 = Major errors, grammatically broken)
- **Fluency:** Does the sentence sound natural to a native speaker? (5 = Very natural; 1 = Very awkward/unnatural)
- **Adequacy:** Is the meaning clear and logical? (5 = Perfectly clear; 1 = Confusing or nonsensical)

**Correction (Optional):** If you assign a score of 3 or lower in any category, please rewrite the sentence in the "Proposed Correction" column to improve its quality.

Metric	Mean Score	Std. Dev.	Agreement (%)
Grammar	4.94	0.25	99.5%
Fluency	4.79	0.51	95.5%
Adequacy	4.68	0.58	94.0%
<b>Average</b>	<b>4.80</b>	-	<b>96.3%</b>

Table 4: Human evaluation results. “Agreement” denotes the percentage of instances where the scores from the two evaluators differed by at most 1 point (on a 5-point scale).

The results of the human evaluation are summarized in Table 4. The generated dataset demonstrated exceptional quality across all linguistic dimensions. Specifically, the *Grammar* score reached a near-perfect mean of **4.94**, while *Fluency* and *Adequacy* achieved **4.79** and **4.68**, respectively. It is worth noting that due to the high concentration of scores at the top of the scale (ceiling effect), standard correlation metrics (e.g., Pearson’s  $r$  or Fleiss’  $\kappa$ ) are not applicable as they require variance to be meaningful (Feinstein and Cicchetti, 1990; Gwet, 2008). Consequently, instead of relying on variance-dependent metrics, we report the inter-annotator agreement rate, specifically calculating the percentage of sample pairs where the score difference between evaluators is within a tolerance of 1 (i.e., *Percent Adjacent Agreement*). This approach provides a more robust measure of consensus for skewed data (Stemler, 2004; Williamson et al., 2012). The high agreement rates (averag-

ing **96.3%**) confirm that the evaluators reached a strong consensus on the high quality of the data.

In addition, for the few low-scoring or unnatural sentences, evaluators provided corrections which were adopted after manual verification.

## D Implementation Details

All experiments were run on a single NVIDIA A100 (80GB) GPU. The models are open-weight and directly downloaded from the Huggingface Transformers (Wolf et al., 2020). The models are used solely for research purposes in accordance with their original terms, consistent with their intended use. And we do not redistribute any third-party model weights. We disable sampling and use greedy decoding to ensure deterministic outputs for all open LLMs by setting `do_sample=False`. All generations are capped at 512 tokens.

## E Additional Results

**Qwen2.5 with Different Scales.** Table 5 presents detailed results of *Qwen2.5* with different scales in Zero-shot. A sharp phase transition is observed between 14B and 32B, where the Aspectual Awareness Gap ( $\Delta_{AA}$ ) surges by 124% ( $0.37 \rightarrow 0.83$ ), significantly outperforming the linear scaling trend ( $r = 0.79$ ).

Size	Acc A	Acc B	Acc C	Acc D	TBR <sub>C</sub> (↓)	$\Delta_{AA}$ (↑)
1.5B	0.21	0.96	0.00	1.00	1.00	0.00
7B	0.20	0.98	0.47	0.97	0.53	0.44
14B	0.24	0.86	0.39	0.98	0.61	0.37
32B	0.53	0.90	0.91	0.92	0.09	0.83
72B	0.43	0.88	0.84	0.97	0.16	0.81

Table 5: Performance of the *Qwen2.5* family across different parameter scales (Zero-shot).

**Trade-offs of the Prompts across Models.** In this section, we extend the analysis of the trade-offs in prompting to the remaining five models. As illustrated in Figures 5–8, the trade-off is pervasive. Models such as Mistral-7B, Yi-1.5, Gemma-2, and GLM-4 exhibit strong, statistically significant inverse correlations. As the prompting strategy becomes more aggressive (Zero-shot  $\rightarrow$  Counterfactual), these models successfully correct the teleological bias in Group C (Pearson  $r > 0.90$ ) but suffer a catastrophic drop in performance on the atelic baseline in Group D (Pearson  $r < -0.79$ ). This confirms that the calibration crisis is a general issue in current LLMs: they struggle to isolate the intervention to only the target telic verbs.

Prompt	Model	“False” Rate
Zero-shot	Llama-3.1-8B-Instruct	0.00
	Mistral-7B-Instruct-v0.3	0.01
	Qwen2.5-7B-Instruct	0.00
	Yi-1.5-9B-Chat	0.00
	deepseek-llm-7b-chat	0.00
	gemma-2-9b-it	0.00
DAP	glm-4-9b-chat-hf	0.00
	Llama-3.1-8B-Instruct	0.19
	Mistral-7B-Instruct-v0.3	0.00
	Qwen2.5-7B-Instruct	0.02
	Yi-1.5-9B-Chat	0.00
	deepseek-llm-7b-chat	0.00
COT	gemma-2-9b-it	0.00
	glm-4-9b-chat-hf	0.00
	Llama-3.1-8B-Instruct	0.00
	Mistral-7B-Instruct-v0.3	0.00
	Qwen2.5-7B-Instruct	0.00
	Yi-1.5-9B-Chat	0.00
Counterfactual	deepseek-llm-7b-chat	0.00
	gemma-2-9b-it	0.00
	glm-4-9b-chat-hf	0.17
	Llama-3.1-8B-Instruct	0.33
	Mistral-7B-Instruct-v0.3	0.00
	Qwen2.5-7B-Instruct	0.00

Table 6: “False” rate (Group C) across different prompt settings and models.

DeepSeek-LLM (Figure 9) presents an outlier behavior. It shows resistance to the intervention, with no significant improvement on Group C ( $p = 0.74$ ) and consequently less degradation on Group D compared to others. This suggests a form of “reasoning rigidity” rather than the “over-correction” seen in other models.

**The Predictions of “False” in Group C.** Table 6 additionally reports the rate of “False” predictions across all experimental settings. It turns out that very few models in very few cases predicted “False”. This is reasonable as in such cases of Imperfective Paradox, the models would either select “Unsure” (the correct predictions for uncertainty) or “True” (the teleological bias). If models predicted “False”, it could be interpreted as indicating that the models “saw” the past progressive for accomplishments as non-completion, which suggests over-prediction. However, based on our observation with the following table, this is not the case.

**Detailed Results by Semantic Verb Class.** Tables 7-18 provide the detailed accuracy breakdown (Accuracy results for Group A and C) and bias

breakdown (TBR results for Group C) for each model across the four semantic sub-classes of telic verbs.

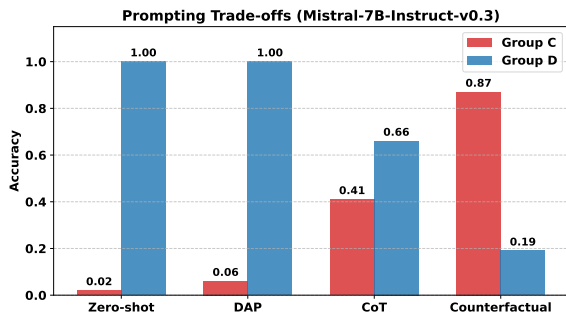


Figure 5: The Trade-off in Prompting (Mistral). The model shows a clear skepticism trap with strong correlations. (Group C:  $r = 0.95, p = 0.05$ ; Group D:  $r = -0.93, p = 0.07$ ).

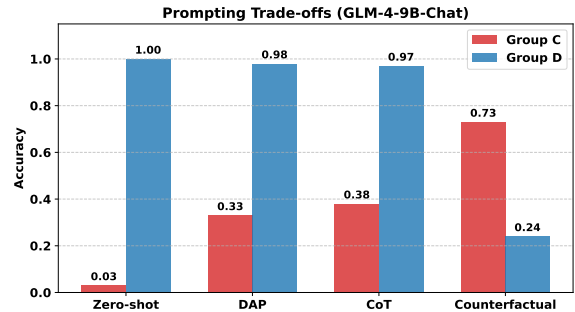


Figure 8: The Trade-off in Prompting (GLM-4). (Group C:  $r = 0.97, p = 0.03$ ; Group D:  $r = -0.79, p = 0.21$ ).

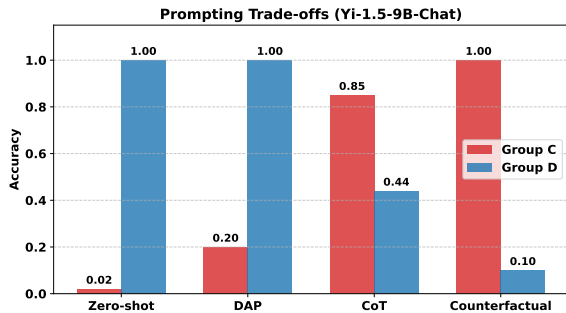


Figure 6: The Trade-off in Prompting (Yi-1.5). Similar to Llama, Yi exhibits a sharp trade-off. (Group C:  $r = 0.97, p = 0.03$ ; Group D:  $r = -0.95, p = 0.05$ ).

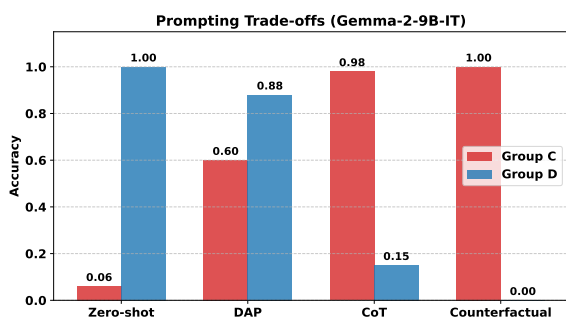


Figure 7: The Trade-off in Prompting (Gemma-2). The model achieves perfect correction on Group C but collapses to zero on Group D. (Group C:  $r = 0.94, p = 0.06$ ; Group D:  $r = -0.95, p = 0.05$ ).

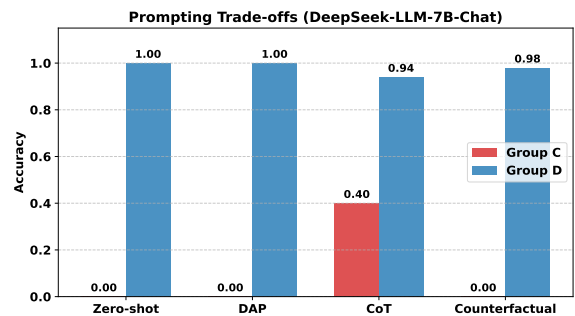


Figure 9: Resistance to Prompting (DeepSeek). Unlike other models, DeepSeek shows weak correlation, indicating rigidity against the counterfactual prompts. (Group C:  $r = 0.26, p = 0.74$ ; Group D:  $r = -0.55, p = 0.45$ ).

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.48	0.56	0.38	0.75
Mistral-7B-Instruct-v0.3	0.32	0.44	0.31	0.88
Qwen2.5-7B-Instruct	0.20	0.11	0.15	0.50
Yi-1.5-9B-Chat	0.34	0.67	0.23	0.62
deepseek-llm-7b-chat	0.05	0.00	0.05	0.00
gemma-2-9b-it	0.02	0.00	0.05	0.00
glm-4-9b-chat-hf	0.09	0.33	0.08	0.50

Table 7: Accuracy Breakdown by Verb Class in Zero-shot (Group A)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.02	0.00	0.03	0.00
Mistral-7B-Instruct-v0.3	0.00	0.00	0.03	0.12
Qwen2.5-7B-Instruct	0.43	0.33	0.51	0.62
Yi-1.5-9B-Chat	0.02	0.11	0.00	0.00
deepseek-llm-7b-chat	0.00	0.00	0.00	0.00
gemma-2-9b-it	0.02	0.11	0.05	0.25
glm-4-9b-chat-hf	0.00	0.11	0.03	0.12

Table 8: Accuracy Breakdown by Verb Class in Zero-shot (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.98	1.00	0.97	1.00
Mistral-7B-Instruct-v0.3	1.00	1.00	0.95	0.88
Qwen2.5-7B-Instruct	0.57	0.67	0.49	0.38
Yi-1.5-9B-Chat	0.98	0.89	1.00	1.00
deepseek-llm-7b-chat	1.00	1.00	1.00	1.00
gemma-2-9b-it	0.98	0.89	0.95	0.75
glm-4-9b-chat-hf	1.00	0.89	0.97	0.88

Table 9: Bias Breakdown by Verb Class in Zero-shot (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.66	0.56	0.49	1.00
Mistral-7B-Instruct-v0.3	0.55	0.67	0.44	1.00
Qwen2.5-7B-Instruct	0.34	0.33	0.28	1.00
Yi-1.5-9B-Chat	0.61	0.78	0.44	0.83
deepseek-llm-7b-chat	0.11	0.11	0.15	0.00
gemma-2-9b-it	0.09	0.00	0.05	0.33
glm-4-9b-chat-hf	0.02	0.33	0.05	0.33

Table 10: Accuracy Breakdown by Verb Class in DAP (Group A)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.30	0.11	0.54	0.17
Mistral-7B-Instruct-v0.3	0.07	0.00	0.05	0.00
Qwen2.5-7B-Instruct	0.84	0.78	0.95	1.00
Yi-1.5-9B-Chat	0.20	0.22	0.15	0.17
deepseek-llm-7b-chat	0.00	0.00	0.00	0.00
gemma-2-9b-it	0.61	0.56	0.56	0.67
glm-4-9b-chat-hf	0.34	0.44	0.28	0.17

Table 11: Accuracy Breakdown by Verb Class in DAP (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.48	0.78	0.33	0.50
Mistral-7B-Instruct-v0.3	0.93	1.00	0.95	1.00
Qwen2.5-7B-Instruct	0.14	0.22	0.03	0.00
Yi-1.5-9B-Chat	0.80	0.78	0.85	0.83
deepseek-llm-7b-chat	1.00	1.00	1.00	1.00
gemma-2-9b-it	0.39	0.44	0.44	0.33
glm-4-9b-chat-hf	0.66	0.56	0.72	0.83

Table 12: Bias Breakdown by Verb Class in DAP (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.14	0.22	0.13	0.33
Mistral-7B-Instruct-v0.3	0.18	0.22	0.23	0.50
Qwen2.5-7B-Instruct	0.30	0.11	0.28	0.67
Yi-1.5-9B-Chat	0.64	0.67	0.38	1.00
deepseek-llm-7b-chat	0.09	0.00	0.03	0.00
gemma-2-9b-it	0.18	0.00	0.08	0.50
glm-4-9b-chat-hf	0.20	0.44	0.21	0.33

Table 13: Accuracy Breakdown by Verb Class in CoT (Group A)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.64	1.00	0.64	0.50
Mistral-7B-Instruct-v0.3	0.36	0.78	0.38	0.33
Qwen2.5-7B-Instruct	0.93	1.00	1.00	1.00
Yi-1.5-9B-Chat	0.89	0.78	0.82	0.83
deepseek-llm-7b-chat	0.23	0.33	0.62	0.33
gemma-2-9b-it	1.00	1.00	0.97	0.83
glm-4-9b-chat-hf	0.30	0.44	0.46	0.33

Table 14: Accuracy Breakdown by Verb Class in CoT (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.36	0.00	0.36	0.50
Mistral-7B-Instruct-v0.3	0.61	0.22	0.56	0.67
Qwen2.5-7B-Instruct	0.07	0.00	0.00	0.00
Yi-1.5-9B-Chat	0.09	0.11	0.18	0.17
deepseek-llm-7b-chat	0.77	0.67	0.38	0.67
gemma-2-9b-it	0.00	0.00	0.03	0.17
glm-4-9b-chat-hf	0.70	0.56	0.54	0.67

Table 15: Bias Breakdown by Verb Class in CoT (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.27	0.22	0.44	0.67
Mistral-7B-Instruct-v0.3	0.30	0.33	0.15	0.67
Qwen2.5-7B-Instruct	0.07	0.11	0.05	0.17
Yi-1.5-9B-Chat	0.05	0.11	0.05	0.33
deepseek-llm-7b-chat	0.23	0.44	0.26	0.17
gemma-2-9b-it	0.02	0.00	0.05	0.00
glm-4-9b-chat-hf	0.25	0.44	0.08	0.50

Table 16: Accuracy Breakdown by Verb Class in Counterfactual (Group A)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	1.00	0.89	1.00	0.67
Mistral-7B-Instruct-v0.3	0.89	0.67	0.90	0.83
Qwen2.5-7B-Instruct	1.00	1.00	1.00	1.00
Yi-1.5-9B-Chat	1.00	1.00	1.00	1.00
deepseek-llm-7b-chat	0.00	0.00	0.00	0.00
gemma-2-9b-it	1.00	1.00	1.00	1.00
glm-4-9b-chat-hf	0.70	0.67	0.79	0.67

Table 17: Accuracy Breakdown by Verb Class in Counterfactual (Group C)

Models	Change of State	Consumption	Creation	Motion to Goal
Llama-3.1-8B-Instruct	0.00	0.00	0.00	0.00
Mistral-7B-Instruct-v0.3	0.11	0.33	0.05	0.17
Qwen2.5-7B-Instruct	0.00	0.00	0.00	0.00
Yi-1.5-9B-Chat	0.00	0.00	0.00	0.00
deepseek-llm-7b-chat	1.00	1.00	1.00	1.00
gemma-2-9b-it	0.00	0.00	0.00	0.00
glm-4-9b-chat-hf	0.20	0.33	0.15	0.17

Table 18: Bias Breakdown by Verb Class in Counterfactual (Group C)