Evaluating Humanities Theory Alignment in Large Language Models: Incremental Prompting and Statistical Assessment

Axel Pichler
Department of German Studies
University of Vienna
axel.pichler@univie.ac.at

Janis Pagel
Department of Digital Humanities
University of Cologne
janis.pagel@uni-koeln.de

Abstract

We propose a method to evaluate the extent to which an LLM's observable input–output behavior aligns with established theories in the humanities and cultural studies. We instantiate the framework on three humanities theories—Davidson's truth-conditional semantics, Lewis's truth in fiction, and Iser's concept of textual gaps—using a top-down, theory-driven black-box framework. Core assumptions of these theories are reconstructed into testable behavioral rules and assessed via controlled classification tasks with systematic prompt comparisons and significance testing. Our experiments show that theory-uninformed classification prompts generally outperform theory-enriched prompts in Lewis and Iser settings, while theory-informed prompts help in the Davidson task. Gemini Flash consistently achieves the highest scores across tasks and corpora, while the Iser gap detection task remains substantially harder than binary truth-conditional judgments. Statistical tests confirm robust prompt effects and the failure of basic prompts. However, model behavior under incremental theory exposure is unstable and architecture-dependent.

1 Introduction

In the Digital Humanities in general, and in Computational Literary Studies in particular, the question of how theories and concepts from the humanities and cultural studies can be modeled and operationalized has long played a central role (Moretti, 2014; Flanders and Jannidis, 2018). This question revolves around the problem of how cultural and textual phenomena can be datafied without losing core specificities. Here, we understand "datafication" as the translation of textual properties into formal, machine-readable representations, and "operationalization" as the development of procedures that, in multiple steps, map the extension of field-specific concepts onto indicators in texts, thereby enabling subsequent measurement (Pichler and Reiter, 2022; Jacke, 2025). Answering this question is complicated by the fact that there are no widely accepted conceptualizations of theory in the humanities and cultural studies in general, nor in literary studies in particular. While the notion of "concept" in literary studies appears to share the same facets as in the general philosophical debate on concepts (Margolis and Laurence, 2023), the notion of "theory" is characterized by nuances that repeatedly depart from an understanding of theory as explicit, ordered, and logically consistent systems of categories designed to describe and explain the phenomena within a given domain. For modeling literary theories and operationalizing their guiding concepts, it follows that reconstruction procedures are often necessary; their selection is guided by pragmatic criteria and geared toward the further algorithmic processing of their results.

With the rise of generative large language models (LLMs) and the so-called prompt-and-predict paradigm that characterizes their use (Liu et al., 2023), the impression arises that at least one of the central problems of modeling textual data and operationalizing text-analytical concepts can be partially addressed: Now that one can interact with the models via string manipulation, it appears less necessary to clarify in detail how one's data model relates to texts or how computational measurement methods relate to traditional concepts. One might simply articulate this relationship verbally (Halterman and Keith, 2025) without having to formulate formal-technical mapping rules—by which we mean explicit, transparent mappings between theoretical categories and algorithmic operations.

Tempting as this hope may sound, it is gradually becoming clear that it holds only to a limited extent, given that LLMs are not ideology-agnostic models but, depending on training methods and training data, incorporate a wide variety of biases. In this regard, we can roughly distinguish between

Proceedings of SIGHUM 2026, pages 280–294
March 28-29, 2026

(a) socio-cultural biases, (b) methodological or architectural biases, and (c) theoretical priors that arise from the distribution of the training data. It remains unclear to what extent such biases can be defined away through mere instruction in the course of prompt-and-predict usage. This raises, particularly for Computational Literary Studies, questions of interpretive validity and the reproducibility of LLM-based results.

Against this backdrop, we pose the following two questions:

RQ1. How can we test whether, and to what degree, a particular theory adheres to an LLM—that is, the extent to which an LLM acts in accordance with this theory without being explicitly prompted to do so? This question targets the theoretical bias of large language models and its empirical detectability.

RQ2. To what extent does an LLM's behavior change when it is provided—step by step—with increasing amounts of information about a theory? Here, we are interested in the models' conditionability and the stability of their outputs under incremental theory exposure.

In what follows, we propose a method to evaluate how closely an LLM's observable input–output behavior aligns with established theories in the humanities and cultural studies. For this purpose, we have selected three "theories" that we take to be operationalizable once they have been explicated to the level of precision required for the tests to be conducted: Donald Davidson's conception of meaning, David Lewis's conceptualization of truth in fiction, and Wolfgang Iser's *Leerstelle* ("gap").[1] Drawing on NLP work in behavioral evaluation, we adopt a theory-driven, top-down approach. Starting from a given theory, we reconstruct its testable claims and decision criteria to derive observable indicators of theory-conforming behavior. Based on these indicators, we design tasks and test datasets, then assess the degree to which an LLM's outputs match the expected decisions of a scholar following the reference theory. Our primary aim is to demonstrate a general pathway for answering our two research questions: derive testable indicators from suitably explicated theories and use them to evaluate LLM behavior. We selected Davidson, and Lewis first and foremost because their core claims admit precise operationalization into falsifiable decision criteria; second, because questions about states of affairs in narrated worlds (truth conditions) are central to literary interpretation and thus to addressing interpretive problems.

[1]As the debate on the operationalization of literary-theoretical concepts has shown, such explications may, on the one hand, risk narrowing the original concepts, but, on the other hand, if carried out with sufficient precision, they allow the resulting measurements to be reintegrated into humanities discourse.

2 Related Work

Behavioral Analysis of LLMs In line with prior work, we adopt the NLP tradition of behavioral testing (Beizer, 1995), which assesses system capabilities by validating input–output behavior without access to internal structure (i.e., black-box evaluation).

There are some attempts to test the behavior of LLMs and compare them to human behavior in similar settings. Wang et al. (2024) compare different values of self-attention heads and feed-forward layers of LLMs with human eye-tracking measures and find some correlations that are not present in non-transformer language models. Akata et al. (2025) let LLMs interact in settings derived from game theory and find LLMs to perform well in games where self-interest is involved and less well in games where coordination is required. Pichler and Pagel (2025) test LLMs via prompts with increasing amount of information on a re-conceptualization of a theory of focalization and find that the systems do not deviate from their pre-conceived concepts.

Use of LLMs in Computational Literary Studies There are multiple recent studies who have used LLMs in order to solve classification and/or information extraction tasks from Computational Literary Studies (Bamman et al., 2024; Hicke and Mimno, 2024; Konle et al., 2024; Pagel et al., 2024; Wu et al., 2024; Bamman et al., 2025; Gius et al., 2025; Graciotti et al., 2025; Guhr et al., 2025; Hicke et al., 2025b; Hicke and Mimno, 2025; Irfan and Ali, 2025; Jannidis et al., 2025; Klähn et al., 2025; Majumdar et al., 2025; Michel et al., 2025; Pichler et al., 2025; Tudor et al., 2025; Werner and Reiter, 2025).

Hicke et al. (2025a) use LLMs to carry out focalization annotations in a selection of novels by Stephen King. Interestingly, they found the models to be not very receptive to prompt variations and in particular, one of the best performing prompts

was one that did not contain any theoretical notions about the target concept (i.e. focalization) at all and they theorize that GPT-4o had utilized pre-exiting notions of focalization from its pre-training data (see Hicke et al., 2025a, p. 745).

3 Workflow

In the following, we describe the workflow that can be used to check whether an LLM produces outputs consistent with the decision criteria of a researcher working within a specific theory in the humanities or cultural studies. The workflow implements a form of behavioral alignment testing. Behavioral alignment testing is a method that compares the behavior of an LLM with that of an informed human being. Therefore, the three theories are converted into a format that allows their core assumptions to be tested. Specifically, we derive testable behavioral rules from the core assumption(s) of each respective theory. These behavioral rules are formulated as conditionals, predicting specific behaviors that align with the antecedent conditions, i.e. the conditions of the literary text. The individual steps of the workflow we use are as follows:

Identification of Core Assumptions The first step takes as its starting point the characteristics of humanities theories already outlined in the introduction–especially the fact that these are often not explicit, ordered, and logically consistent systems of categories. It involves the systematic identification of core assumptions or foundational principles underlying a given theoretical framework. These assumptions serve as the theoretical premises from which testable behavioral rules are derived. They define the central characteristics of the subject area and specify how these characteristics or statements about them are to be linked.

Formation of Behavioral Rules Based on the identified core assumptions, testable behavioral rules are formulated as conditional statements. We operationalize the consequence relation between a core assumption and its predicted consequence as logical entailment ($\models$). These rules predict how an individual is expected to behave under the given theoretical framework. They define the expected behavioral outcomes and specify the conditions under which deviations from these predictions may occur. This step ensures that predicted behavior is directly grounded in the theoretical premises and can be empirically tested.

Experimental Design An experimental setting is developed to test manually created sentences derived from the behavioral rules. LLMs are presented with these sentences as controlled input, followed by a classification task or a sentence completion task. In the classification task, we evaluate model predictions over a constrained label set (e.g., True, False, Undetermined). In the sentence completion task, the model is prompted to continue a given sentence, with completion evaluated via the same constrained label mapping. We compare three setups: (i) sentence completion task, (ii) theory-uninformed classification, and (iii) theory-enriched classification. The design isolates the influence of specific variables on the LLMs' interpretive or classificatory outputs and evaluates whether their responses align with the predictions derived from the theoretical premises.

4 Reference Theories and Operationalization

4.1 Davidson's Radical Interpretation

The first theory we reconstruct is Donald Davidson's theory of meaning as he developed it in particular in *Radical Interpretation* (Davidson, 1973). Davidson develops his theory based on the following questions: What knowledge could enable a listener to understand the utterances of a speaker of a language unknown to the listener? How could the listener acquire this knowledge?

Davidson's answer is guided by his central hypothesis "that a theory of truth, modified to apply to a natural language, can be used as a theory of interpretation" (Davidson, 1973, p. 189). This aligns with our aim of operational reconstruction insofar as Davidson's program already suggests a procedure: derive a compositional truth theory licenses interpretive competence. We adapt this procedure to account for LLM-specific constraints (non-agency, lack of beliefs), treating LLM outputs as simulations of a competent speaker's judgments.

Core Principle (Truth-Conditional Test): Given a factual text T that represents a situation K, and an assertion φ (attributed to a speaker S at time Z), we evaluate truth *relative to* T over a two-way label space by restricting items to cases where $K \models \varphi$ or $K \models \neg\varphi$. K contains only information textually licensed by T:

$$\mathrm{Truth}_T(\varphi) = \begin{cases} \text{True} & \text{if } K \models \varphi, \\ \text{False} & \text{if } K \models \neg\varphi. \end{cases}$$

By construction, items for which T does not fix the truth value of φ are excluded from the dataset.[2]

Interpretive Constraint (Charity/Coherence): Before evaluating $\mathrm{Truth}_T(\varphi)$, fix an interpretation function I (reference assignment, disambiguation, coreference, quantifier scope) that *maximizes coherence and charity* subject to T and publicly available evidence (Davidson, 1973). Charity constrains the choice of I; it is *not* itself a condition for truth. We then assess φ under I relative to T.

4.2 Lewis' Theory of Fiction

David Lewis, in his 1978 essay *Truth in Fiction* (Lewis, 1983), made the classical contribution to addressing the question of what holds true in the fictional world of a literary text. The core idea of his approach is that the fictional truth of a story depends on identifying the possible worlds where the events of the story take place (Köppe, 2014). Building on this idea, Lewis proposed the following three principles for determining fictional truth in stories:

1. **Principle 1**: What is explicitly described or narrated in the story is true in the fictional world.
2. **Principle 2**: What corresponds, in its factuality, to our "real world" is also true in the fictional world (the *reality principle*).
3. **Principle 3**: What corresponds, in its factuality, to collective belief worlds at the time the fictional story was created is also true in the fictional world (the *mutual belief principle*).

The principles[3] have been subject to criticism from various theoretical perspectives. Regarding the first principle, it has been argued that it is insufficient because it does not account for implicit truths. With respect to the second principle (the reality principle), Currie (1990) argued that it results in: (a) statements that are irrelevant to the plot but true in reality being considered true in the fictional world, and (b) in cases of unreliable narration, a bias toward interpretations that align with our reality. Following Lewis (1983), we operationalize three core principles of truth in fiction, each paired with a two-way behavioral labeling rule. The full formal definitions and interpretive constraints are provided in Appendix A.

4.3 Iser's Concept of the *Leerstelle*

Alongside Hans Robert Jauss, Wolfgang Iser is a central representative of reception theory in the German-speaking world. This approach places an idealized construct of the reader at the center of its literary-theoretical considerations. Drawing on Roman Ingarden's phenomenological theory, Iser understands literary comprehension as a dynamic process in which the reader continually draws on what has already been read while simultaneously forming expectations about what is to come (Sneis, 2018, pp. 144–161). Because texts never fully specify all aspects of a represented world, the reader supplies missing information—from visual completion, to the establishment of logical coherence. Central to this is the concept of the *gap* (German: *Leerstelle*).

As recent scholarship has emphasized (Willand, 2015), Iser's notion is formulated with considerable openness, which necessitates explication for practical application. In this paper, we use the following: A "gap" is a textually grounded indeterminacy—namely, a missing relation between explicitly articulated partial units (e.g., propositions, perspectives, or events) that prompts the reader to establish coherence through inferential reasoning and expectation formation, or to process provisional incoherences.

Core Principle (Iser/Gaps) Given a text T that represents a situation K, and an assertion φ (attributed to a speaker S at time Z), under a fixed interpretation I:

- φ is *True* in T if $K \models \varphi$,
- φ is *False* in T if $K \models \neg\varphi$,
- φ is *Not articulated (Gap)* if neither $K \models \varphi$ nor $K \models \neg\varphi$ holds, because T leaves a requi-

[2] In view of the reconstructed Core Principle, several limitations must be noted in relation to Davidson's theory of meaning. First, LLMs are not agents, they do not possess beliefs, and consequently cannot hold propositions to be true; we therefore assess simulated judgments. Second, the LLM–human interaction lacks the real-world situational and causal framework presupposed by radical interpretation (Davidson, 2001). Accordingly, our evaluation is conducted relative to the factual text T: we restrict the dataset to cases where $K \models \varphi$ or $K \models \neg\varphi$ and use a two-way label space {True, False}. Charity and coherence serve as constraints on interpretation (choice of reference, disambiguation) rather than as components of truth conditions. Thus, the test set examines to what extent LLM outputs approximate the communicative judgments of a competent speaker in Davidson's sense, relative to T.

[3] For Principles 2 and 3, we operationalized our unique closest world assumption by grounding Principle 2 in currently valid knowledge, while grounding Principle 3 in knowledge that was valid at the time of the publication of the literary texts.

site relation between explicit units unspecified under I.

Behavioral Rule (Three-way labeling) Under the fixed interpretation I, assign:

- *True* if $K \models \varphi$,
- *False* if $K \models \neg\varphi$,
- *Not articulated (Gap)* if neither entailment holds due to an indeterminacy (a missing relation among explicit units in T).

5 Experimental Setup

5.1 Data Annotation

We evaluate our approach using manually created test data with respect to the narrative worlds of two literary reference texts—Arthur Schnitzler's *Amerika* and Wolfgang Borchert's *Das Brot*—as well as selected excerpts from newspaper articles. The test instances consist of sentences specifically designed for the three theoretical frameworks and, like the reference texts, are formulated in German. In the case of Davidson and Lewis, each sentence is assigned a truth value (true/false), while for Iser an additional option, "not fully articulated," is available (for examples see Appendix C). The test datasets for Davidson and Lewis were created by one of the authors of this paper and annotated by another, whereas the test dataset for Iser was created collaboratively. We retained only instances with full annotator consensus to ensure data quality. Table 1 provides an overview of the resulting test datasets.[4]

File	Prin.	κ	N
Davidson	–	0.709	170
Lewis-Bor	P1	0.971	69
Lewis-Bor	P2	0.971	69
Lewis-Bor	P3	0.941	68
Lewis-Sch	P1	0.850	74
Lewis-Sch	P2	0.774	71
Lewis-Sch	P3	0.822	73
Iser	–	–	60

Table 1: Overview of the manually created test datasets, including inter-annotator agreement (Cohen's κ) and the number of rows with full annotator consensus.

5.2 Models

We evaluate four state-of-the-art large language models via the OpenRouter API: Gemini 2.0 Flash, Claude Sonnet 4.5, Qwen and Ministral. This diverse selection spans different model families, sizes, architectures, and training paradigms, allowing us to assess the generalizability of our findings across the current LLM landscape.

5.3 Prompt Design

We systematically compare multiple prompt templates varying in structure and theoretical grounding. Baseline prompts require the model to complete the input's last sentence directly with the label; classification prompts explicitly present the labels but not the underlying theory; theory-informed prompts incorporate explicit theoretical principles relevant to the classification task. All prompts include full textual context and request binary or, in the case of Iser, multiclass classification outputs; exemplary prompts for the Lewis framework are provided in Appendix D. As an example of a full-length context, Arthur Schnitzler's short story "Amerika" is included in Appendix B as a machine translation produced with GPT-5.2.

5.4 Evaluation Setup

Experiments comprise the Davidson truth-conditional task (10 short newspaper articles), six evaluations of Lewis's principles of truth in fiction across two texts (Schnitzler and Borchert, 3 principles each), and two Iser narrative gap detection tasks (Schnitzler and Borchert).

For each model-prompt-dataset combination, we conduct up to three independent runs – 3 for the Davidson and Iser, 2 for the Lewis data – with temperature 0.1 to balance consistency with minimal stochasticity while capturing any residual variation in model responses.

Classification Metrics For the Lewis and Davidson experiments (binary classification: true/false), we report F1-score (weighted), precision, recall, and accuracy. For the Iser experiment (3-class classification: true/false/not formulated), we additionally report F1-macro and F1-micro scores, alongside per-class precision, recall, and F1-scores to assess performance on the theoretically critical "not formulated" category representing Iser's textual gaps (*Leerstellen*).

Statistical Significance Testing To assess statistical significance, we employ a multi-tiered approach adapted to each experiment's classification

[4]All manually created reference data, the Jupyter notebooks used for experimentation, and the resulting evaluation tables are publicly available at the project's GitHub repository: https://github.com/AxPic/LLM-Theory-Alignment.

structure and number of runs: For binary classification (Lewis, Davidson), we use McNemar's test to compare prompt pairs on individual predictions, accounting for the dependency structure of identical test instances. For multi-class classification (Iser), we employ Cochran's Q test, which tests whether different prompts achieve systematically different success rates across the three classes.

All statistical tests use a significance level of $\alpha = 0.05$. This framework enables both granular instance-level comparisons and robust assessment of run-level consistency, while appropriately accounting for the paired nature of our experimental design and the distinct classification structures of each dataset.

6 Results

We report results from nine classification experiments across four large language models. We present overall performance (§6.1), prompt engineering effects (§6.2), and statistical significance (§6.3).

6.1 Overall Performance

Across all nine experiments and prompts, the models exhibited a stable performance ranking with only minimal variations, though absolute scores varied substantially by task type. Table 2 summarizes best performance across all experiments: Models achieved F1-scores ranging from 0.752 to 1.000, with substantial variation across tasks and prompt types. The F1 scores of the majority baseline are added for comparison, showing that on average, models easily outperform this simple baseline.

Gemini 3 Flash consistently outperformed all competitors and achieved the highest observed peak scores. Its performance remained robust across corpora and task formulations, indicating strong task-general capabilities.

Qwen 3 Next and *Mistral Ministral 14B* occupied intermediate positions across experiments. While neither model surpassed Gemini in any condition, both achieved moderate peak performance on selected tasks, with Qwen generally outperforming Mistral except in one of the Iser-settings.

Claude Sonnet 4.5 showed the weakest overall performance. Its best observed result was achieved on the Lewis-Borchert task with the theory-uninformed prompt (peak F1=0.869), while highlighting substantial brittleness with respect to task formalization and prompt style.

The following patterns emerged with regard to the individual tasks:

Binary Classification Tasks (Davidson, Lewis) The binary classification experiments achieved best F1-scores between 0.841 and 1.000, demonstrating high overall performance. The Davidson experiment, requiring truth-conditional evaluation against single contexts, reached F1=0.953 with Gemini 3 Flash using theory-enriched prompting. The Lewis experiments, evaluating three different principles of truth in fiction, showed solid performance: Principle 1 achieved perfect classification on Schnitzler (F1=1.000), though the reality principle (0.916) and the mutual belief principle (0.905) proved more challenging. On both test sets best performances were achieved in relation to Principle 1, which relies exclusively on statements explicitly articulated in the text, suggesting that the models – regardless of the theoretical enrichment of the prompts – have an easier time with this principle over the others.

Multi-Class Classification (Iser) The Iser experiments, which require three-way classification (*true*/*false*/*not formulated*), proved substantially more challenging than the binary settings. Across models, best performance reached $F1_{\text{weighted}} = 0.752$ on Schnitzler and 0.783 on Borchert, corresponding to an average deficit of approximately 0.15 F1 points relative to the binary tasks. These results suggest that Iser's concept of *Leerstellen* poses a fundamental challenge for current LLMs.

6.2 Prompt Engineering Effects

Figure 1 presents prompt effectiveness across all experiments. Contrary to initial hypotheses, theory-uninformed classification prompts consistently outperformed theory-enriched prompts in Lewis experiments (mean advantage: +0.16 F1 points) and Iser experiments (Schnitzler: +0.25, Borchert: +0.20), while theory prompts excelled only in Davidson contexts.

Prompt effectiveness varied dramatically in Davidson: `prompt_theory` achieved mean F1=0.778, `prompt_class` F1=0.833, while `prompt_basic` achieved rather low scores in comparison (F1=0.373).

Across all six Lewis evaluations (3 principles × 2 corpora), theory-uninformed classification prompts dominated: Schnitzler class=0.822 vs. theory=0.674 (+0.148); Borchert class=0.828 vs.

Experiment	Task Type	Best Model	Best Prompt	Peak F1	Maj. F1	N
Davidson	Binary	Gemini 3 Flash	theory	0.953	0.373	170
Lewis-Sch P1	Binary	Gemini 3 Flash	theory_1	1.000	0.379	74
Lewis-Sch P2	Binary	Gemini 3 Flash	class	0.916	0.423	71
Lewis-Sch P3	Binary	Gemini 3 Flash	theory_3	0.905	0.437	73
Lewis-Bor P1	Binary	Gemini 3 Flash	class	0.986	0.408	69
Lewis-Bor P2	Binary	Gemini 3 Flash	class	0.841	0.425	69
Lewis-Bor P3	Binary	Gemini 3 Flash	class	0.868	0.436	68
Iser-Sch (3-class)	Multi-class	Gemini 3 Flash	theory	0.752	0.176	60
Iser-Bor (3-class)	Multi-class	Gemini 3 Flash	theory	0.783	0.176	60
Mean (Binary)				**0.918**		
Mean (Multi-class)				**0.768**		

Table 2: Best performance per experiment. F1-scores represent weighted F1-scores.

theory=0.656 (+0.172). This pattern held consistently across principles with only one exception: Schnitzler Principle 1, where `prompt_theory_1` achieved perfect F1=1.000 (vs. class 0.879). Theory prompts showed high inter-model variance (SD=0.335–0.390 vs. 0.075–0.092 for theory-uninformed classification prompts), indicating sensitivity to model architecture and suggesting that formal notation does not universally improve classification.

For the Iser task, prompt effectiveness diverges sharply by model architecture rather than exhibiting a uniform advantage for either prompting strategy. Across both corpora, Claude consistently failed under theory-enriched prompting (Macro-F1 $\leq$ 0.02 on Schnitzler; 0.00 on Borchert) while performing substantially better with theory-uninformed classification prompts (0.41 and 0.37, respectively). Gemini 3 Flash exhibited the opposite pattern, achieving its strongest results under theory-enriched prompting on both Schnitzler (Macro-F1 = 0.75 vs. 0.54) and Borchert (0.78 vs. 0.64). Mistral and Qwen occupied an intermediate regime: both benefited from classification prompts, but theory-enriched prompting degraded collapsed performance (Schnitzler: Mistral 0.13, Qwen 0.42; Borchert: Mistral 0.00, Qwen 0.46). These results contradict the hypothesis that Iser-specific *Leerstellen* terminology would uniformly aid gap recognition.

Relevance of logical notation To test whether logical notation itself causes confusion, we evaluated a variant (`prompt_theory_no_symbol`) only on the Davidson-data in which symbols were replaced by natural-language equivalents ("$K \models \varphi$" $\rightarrow$ "situation logically supports φ", "iff" $\rightarrow$ "if and only if"). Table 3 shows divergent responses to symbol removal. Contrary to expectations, removing symbols *decreased* performance (F1 = 0.730 vs. 0.778), with significant degradation for Claude (-0.120, $p = 0.001$) and Mistral (-0.059, $p < 0.001$), while Gemini remained essentially unaffected (-0.002, $p = 1.0$). This counterintuitive result indicates that formal symbols provide *structural scaffolding* that is beneficial even to models struggling with formal reasoning: Claude's performance, though poor with symbols (0.507), heavily lowers without them (0.386).

Model	With Symbols	No Symbols	Δ	Signif.
Gemini 3 Flash	0.945	0.943	-0.002	p=1.00
Qwen 3 Next	0.840	0.827	-0.013	p=0.50
Mistral 14B	0.822	0.763	-0.059	p<0.001***
Claude Sonnet	0.507	0.386	-0.120	p=0.001***
Average	**0.778**	**0.730**	**-0.049**	—

Table 3: Symbol removal effects per model (Davidson theory prompts). Negative Δ indicates performance loss without symbols. Significance from McNemar tests comparing predictions.

Model-Prompt Interactions Table 4 shows model-specific prompt preferences averaged across all nine experiments. Gemini favored formal prompts marginally in Davidson but theory-uninformed classification prompts in Lewis/Iser, yielding near-parity overall (+0.01). Claude showed dramatic preference for theory-uninformed classification instructions (+0.20), while Qwen and Mistral favored theory-uninformed classification prompts moderately (+0.08, +0.09).

6.3 Statistical Significance

We assess statistical significance at the instance-level differences by comparing model predictions across prompts, using McNemar tests for binary tasks and Cochran's Q tests for multi-class Iser tasks. McNemar tests on binary classification revealed significant prompt differences in 71%

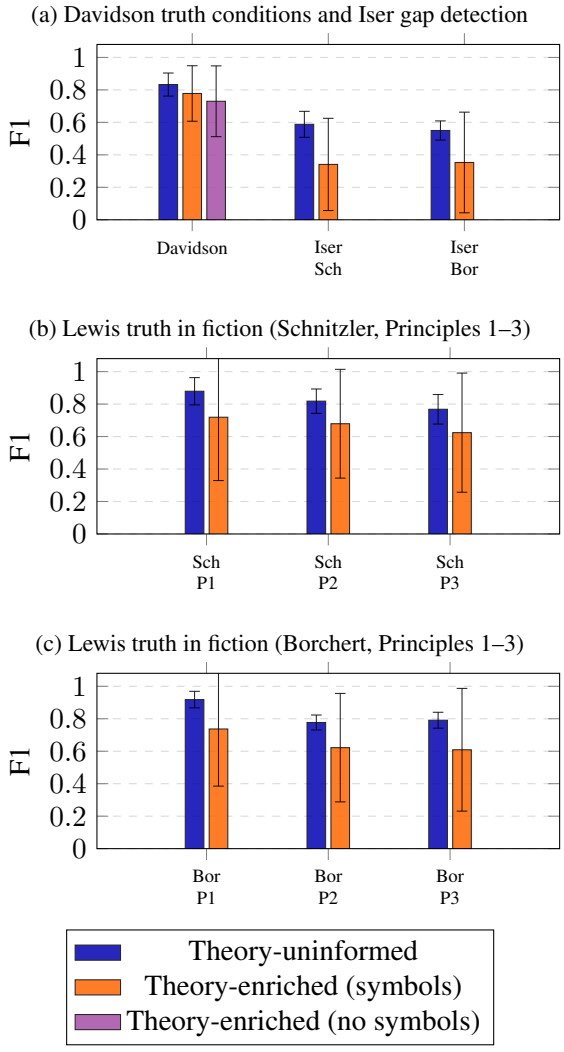

Figure 1: F1-scores by prompt type across tasks, split by theoretical framework and corpus. Error bars indicate the standard deviation across models and runs, with some bars extending beyond the theoretical F1 maximum of 1.0 due to high inter-model variance. Blue bars denote theory-uninformed classification prompts; orange bars denote theory-enriched prompts with formal notation; violet bars denote theory-enriched prompts without notation (Davidson only). Large error bars for theory-enriched prompts (e.g., Lewis-Sch P1) demonstrate architecture-dependent effectiveness, with Gemini achieving near-perfect scores while Claude fails completely on the same prompts.

Model	**Theory**	**Classification**	**Δ**
Gemini 3 Flash	0.850	0.861	+0.011
Claude Sonnet	0.365	0.565	+0.200
Qwen 3 Next	0.694	0.772	+0.078
Mistral 14B	0.619	0.708	+0.089

Table 4: Average F1-scores for theory-enriched vs. theory-uninformed prompts across all experiments. Δ indicates preference strength (positive values favor role-based prompts). Theory-enriched prompts substantially harm Claude (-0.20).

Experiment	**Signif.**	**Total**	**Rate**
Davidson	9	12	75%
Lewis-Sch P1	8	12	67%
Lewis-Sch P2	8	12	67%
Lewis-Sch P3	9	12	75%
Lewis-Bor P1	8	12	67%
Lewis-Bor P2	8	12	67%
Lewis-Bor P3	8	12	67%
Iser-Sch	8	12	67%
Iser-Bor	9	12	75%
Overall	**75**	**108**	**70%**

Table 5: Statistical significance summary. "Signif." = comparisons with p<0.05.

of model-experiment combinations. For multi-class Iser experiments, Cochran's Q tests showed 67% significance (Schnitzler) and 82% significance (Borchert), with Borchert's higher rate reflecting greater ambiguity and thus larger prompt-driven performance differences (Table 5).

Consistent patterns emerged: basic vs. theory-uninformed/theory-enriched comparisons achieved 100% significance (all p<.001), confirming the failure of basic prompts. Critically, theory vs. theory-uninformed classification comparisons showed significance in only 17% of cases, indicating theory-enriched prompting rarely yields statistically distinguishable improvements over simple theory-uninformed classification prompts.

7 Conclusion

Our experiments across three theoretical frameworks (Davidson truth conditions, Lewis's three principles of truth in fiction, and Iser's gaps) and multiple texts yield four main findings. First, theory-uninformed prompting consistently outperforms theory-enriched prompting across all tasks and models, indicating that the explicit incorporation of theoretical constraints does not reliably improve performance under the evaluated conditions. Second, formal notation provides structural

scaffolding: removing logical symbols from theory prompts leads to a universal performance decrease ($\Delta = -0.049$), with particularly severe effects for Claude ($\Delta = -0.120$), suggesting that symbolic structure anchors model behavior even when it is not semantically transparent. Third, Gemini 3 Flash performs best overall (rank #1 in 9/9 experiments) while exhibiting near-complete notation-agnosticism ($\Delta = -0.002$), pointing to architectural capacity for abstraction beyond surface form. Fourth, multi-class classification in the Iser gap task remains fundamentally challenging (best mean F1 = 0.768 vs. 0.924 for binary tasks, $\Delta = -0.16$), with uniformly low per-class performance (F1 = 0.27) indicating limited conceptual grounding.

Taken together, these results suggest that the extent to which a literary theory adheres to an LLM cannot be inferred from theory-enriched prompting alone, but is more reliably assessed through controlled comparisons with theory-uninformed baselines. Across tasks, models do not consistently act in accordance with the target theory when left unprompted; instead, apparent theory alignment emerges primarily when task structure aligns with strong surface regularities, such as binary truth-conditional judgments in the Davidson or Lewis settings. By contrast, tasks requiring abstract or meta-textual distinctions, such as Iser gap detection, show no evidence of latent theoretical bias, with overall weaker performance despite partial gains from theory-enriched prompting.

Incremental exposure to theoretical information does not lead to stable or monotonic improvements in model behavior. Rather, LLMs differ markedly in their conditionability: while some models (notably Gemini 3 Flash) remain robust under increasing theoretical and notational complexity, others (notably Claude) exhibit degradation when confronted with formalized theory, especially when logical notation is involved. The consistent performance drop observed after removing formal symbols further suggests that models might rely on notation as structural scaffolding rather than on semantic understanding of the underlying theory. Overall, theory prompting yields unstable and architecture-dependent effects, undermining assumptions of smooth theoretical internalization through stepwise exposure.

For the use of LLMs in CLS, these findings highlight the need for precise operationalizations of theoretical concepts and for annotation guidelines and reference data derived from these operationalizations. Such resources are essential for systematically evaluating model behavior on tasks grounded in literary theory. Only with such resources can the construct validity of model outputs be ensured. Moreover, by evaluating outputs against these reference data, we can determine a model's suitability for a given task irrespective of whether it is genuinely theory-aligned or merely exhibits surface-level conformity.

Declaration on Generative AI

We used GPT-5 and Claude Sonnet 4.5 for proofreading and code generation; all outputs were reviewed by the authors, who take full responsibility for the final content.

Limitations

Our study faces several limitations. First, we identify and operationalize only a single core principle per theory; ideally, multiple such core principles would be delineated and tested to probe theoretical coverage and robustness. Second, because our evaluation relies on black-box behavioral testing, we cannot determine whether a model is genuinely theory-aligned or merely exhibits surface-level conformity. Third, we evaluate on compact, manually constructed datasets (e.g., 60 Iser items per corpus; 68–74 per Lewis setting; 170 for Davidson), which constrains external validity and risks item-design artifacts. Fourth, reproducibility is limited: each model–prompt–dataset condition was run only two to three times (temperature 0.1), and results are explicitly characterized as indicative rather than conclusive. Fifth, the notation-manipulation result (performance drop after symbol removal) was tested only in the Davidson setting, limiting generalization of the "notation as scaffolding" interpretation to Lewis and Iser. Sixth, with respect to the basic prompts, which were designed to elicit a labeled continuation of an explicitly marked output sentence, we observe that instruction-tuned models no longer behave in line with the classical next-token prediction paradigm. Accordingly, this setup yields consistently lower scores across all evaluated metrics. Seventh, findings depend on model family and serving configuration—four instruction-tuned models were accessed via the OpenRouter API, and several effects were architecture-dependent and unstable under theory prompts. Finally, our task and prompt space is narrow: we compare basic, theory-uninformed classification, and theory-

enriched prompts (plus a notation variant), leaving broader process-oriented protocols outside the present scope.

Acknowledgments

We thank Janina Jacke, with whom the foundational groundwork for this study regarding the operationalization of the Davidson and Lewis theories has been collaboratively developed. We also thank Dominik Gerstorfer and Jonas Kuhn for their collaboration on an early version of the Davidson experiments. The second author has carried out this work under funding of the German Research Foundation (DFG, grant number 508319395).

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390.

David Bamman, Sabrina Baur, Mackenzie Hạnh Cramer, Anna Ho, and Tom McEnaney. 2025. Measuring the Stories in Contemporary Songs. *Anthology of Computers and the Humanities*, 3:820–844.

David Bamman, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. On Classification with Large Language Models in Cultural Analytics. In *CHR2024*, pages 494–527, Aarhus.

Boris Beizer. 1995. *Black Box Testing: Techniques for Functional Testing of Software and Systems*. John Wiley, New York.

Gregory Currie. 1990. *The Nature of Fiction*, 1 edition. Cambridge University Press.

Donald Davidson. 1973. Radical interpretation. *Dialectica*, 27(3/4):313–328.

Donald Davidson. 2001. *Epistemology Externalized*, 1 edition, pages 193–204. Oxford University PressOxford.

Julia Flanders and Fotis Jannidis, editors. 2018. *The Shape of Data in the Digital Humanities: Modeling Texts and Text-based Resources*, 1 edition. Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2019. | Series: Digital research in the arts and humanities.

Evelyn Gius, Stefanie Messner, and Axel Pichler. 2025. How are Literary Histories written? An LLM-based Analysis of Objects and Perspectives in German Literary History. *Anthology of Computers and the Humanities*, 3:1090–1107.

Arianna Graciotti, Franziska Pannach, Valentina Presutti, and Federico Pianzola. 2025. Llamas Don't Understand Fiction: Application and Evaluation of Large Language Models for Knowledge Extraction from Short Stories in English. *Anthology of Computers and the Humanities*, 3:4–32.

Svenja Guhr, Huijun Mao, and Fengyi Lin. 2025. Rethinking Scene Segmentation. Advancing Automated Detection of Scene Changes in Literary Texts. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 79–86, Albuquerque, New Mexico. Association for Computational Linguistics.

Andrew Halterman and Katherine A. Keith. 2025. What is a protest anyway? Codebook conceptualization is still a first-order concern in LLM-era classification. *arXiv preprint*. ArXiv:2510.03541 [cs].

Rebecca Hicke and David Mimno. 2024. [Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.

Rebecca M. M. Hicke, Yuri Bizzoni, Pascale Feldkamp, and Ross Deans Kristensen-McLachlan. 2025a. Says Who? Effective Zero-Shot Annotation of Focalization. *Anthology of Computers and the Humanities*, 3:739–755.

Rebecca M. M. Hicke, Brian W. Haggard, Mia Ferrante, Rayhan Khanna, and David Mimno. 2025b. Are You There God? Lightweight Narrative Annotation of Christian Fiction with LMs. *Anthology of Computers and the Humanities*, 3:1012–1035.

Rebecca M. M. Hicke and David Mimno. 2025. *Looking for the inner music* : Probing LLMs' understanding of literary style. *Computational Humanities Research*, 1:e3.

Saniya Irfan and Syed Juned Ali. 2025. QaLLM: An LLM-based NER Dataset Curation, Annotation and Evaluation in Historical Urdu Elegies. *Anthology of Computers and the Humanities*, 3:922–937.

Janina Jacke. 2025. Operationalization and interpretation dependence in computational literary studies. *JCLS*, 4(1).

Fotis Jannidis, Rabea Kleymann, Julian Schröter, and Heike Zinsmeister. 2025. Do Large Language Models Understand Literature? Case Studies and Probing Experiments on German Poetry. *JCLS*, 4(1).

Jannis Klähn, Janos Borst-Graetz, and Manuel Burghardt. 2025. From dictionaries to LLMs – an evaluation of sentiment analysis techniques for German language data. *Computational Humanities Research*, 1:e4.

Leonard Konle, Merten Kröncke, Fotis Jannidis, and Simone Winko. 2024. On the Unity of Literary Change. The Development of Emotions in German Poetry, Prose, and Drama between 1850 and 1920 as a Test Case. In *CHR2024*, pages 282–300, Aarhus.

Tilmann Köppe. 2014. 8. Fiktive Tatsachen. In Tobias Klauk and Tilmann Köppe, editors, *Fiktionalität*, pages 190–208. DE GRUYTER.

David Lewis. 1983. *Truth in Fiction*, 1 edition, pages 261–280. Oxford University PressNew York.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Pritha Majumdar, Franziska Pannach, Arianna Graciotti, and Johan Bos. 2025. '... like a needle in a haystack": Annotation and Classification of Comparative Statements. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 261–271, Albuquerque, New Mexico. Association for Computational Linguistics.

Eric Margolis and Stephen Laurence. 2023. Concepts. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2023 edition. Metaphysics Research Lab, Stanford University.

Gaspard Michel, Elena V. Epure, Romain Hennequin, and Christophe Cerisara. 2025. Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 742–755, Albuquerque, New Mexico. Association for Computational Linguistics.

Franco Moretti. 2014. "Operationalizing": or, the Function of Measurement in Modern Literary Theory. *The Journal of English Language and Literature*, 60:3–19.

Janis Pagel, Axel Pichler, and Nils Reiter. 2024. Evaluating In-Context Learning for Computational Literary Studies: A Case Study Based on the Automatic Recognition of Knowledge Transfer in German Drama. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 1–10, St. Julians, Malta. Association for Computational Linguistics.

Axel Pichler and Janis Pagel. 2025. Investigating Conceptual Plasticity: On Detecting a Re-Conceptualization of Focalization with Large Language Models. Proceedings of DH2025.

Axel Pichler, Janis Pagel, and Nils Reiter. 2025. Evaluating LLM-Prompting for Sequence Labeling Tasks in Computational Literary Studies. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 32–46, Albuquerque, New Mexico. Association for Computational Linguistics.

Axel Pichler and Nils Reiter. 2022. From concepts to texts and back: Operationalization as a core activity of digital humanities. *Journal of Cultural Analytics*, 7(4).

Jørgen Sneis. 2018. *Phänomenologie und Textinterpretation: Studien zur Theoriegeschichte und Methodik der Literaturwissenschaft*. De Gruyter.

Crina Tudor, Beata Megyesi, and Robert Östling. 2025. Prompting the Past: Exploring Zero-Shot Learning for Named Entity Recognition in Historical Texts Using Prompt-Answering LLMs. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226, Albuquerque, New Mexico. Association for Computational Linguistics.

Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024. Probing Large Language Models from a Human Behavioral Perspective. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 1–7, Torino, Italia. ELRA and ICCL.

Nicolas Werner and Nils Reiter. 2025. Between Woolf and Homer: An Explorative Approach to Intertextuality Detection using Large Language Models. *Anthology of Computers and the Humanities*, 3:382–435.

Marcus Willand. 2015. Isers impliziter Leser im praxeologischen Belastungstest: Ein literaturwissenschaftliches Konzept zwischen Theorie und Methode. In Andrea Albrecht, Lutz Danneberg, Olav Krämer, and Carlos Spoerhase, editors, *Theorien, Methoden und Praktiken des Interpretierens*, pages 237–270. DE GRUYTER.

Yaru Wu, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo. 2024. Perplexing Canon: A study on GPT-based perplexity of canonical and non-canonical literary works. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 172–184, St. Julians, Malta. Association for Computational Linguistics.

A Detailed Operationalization of Lewis' Principles

Core Principle 1 – Lewis (1983): "A sentence of the form 'In fiction F, φ' is true if and only if φ is true at every world where F is told as known fact rather than fiction."

Behavioral Rule 1 (two-way). Under a fixed interpretation I, label φ as True in F iff every world where F is told as fact makes φ true (including logical consequences under I); otherwise label it False.

Core Principle 2 – Lewis (1983): "A sentence of the form 'In fiction F, φ' is non-vacuously true if and only if some world where F is told as known fact and φ is true differs less from our actual world, on balance, than does any world where F is told as known fact and φ is not true."

Behavioral Rule 2 (two-way; unique closest world by data design). Under a fixed interpretation I, and given that the dataset guarantees a unique closest world to the actual world among the worlds where F is told as fact, label φ as True in F iff that closest world makes φ true; otherwise label it False.

Core Principle 3 – Lewis (1983): "A sentence of the form 'In fiction F, φ' is non-vacuously true if and only if whenever w is one of the collective belief worlds of the community of origin of F, then some world where F is told as known fact and φ is true differs less from w, on balance, than does any world where F is told as known fact and φ is not true."

Behavioral Rule 3 (two-way; unique closest worlds by data design). Under a fixed interpretation I, and given that the dataset guarantees for each relevant collective belief world w a unique closest world among the worlds where F is told as fact, label φ as True in F iff for every such w the unique closest world makes φ true; otherwise label it False.

Interpretive Constraint (Similarity/Community Beliefs) Before evaluating labels for sentences of the form "In fiction F, φ", we fix an interpretation function I (reference assignment, disambiguation, coreference, quantifier scope). In addition, we fix a similarity ordering to the actual world for F-as-fact worlds (and, for Core Principle 3, to each collective belief world w of the community of origin), with uniqueness ensured by data design. These constraints guide evaluation but are not themselves truth conditions. We then assess φ under I by quantifying over all F-as-fact worlds (Core Principle 1), or by selecting the unique closest such world to the actual world (Core Principle 2), or to each w (Core Principle 3).

B Example Reference Text: Arthur Schnitzler's Amerika

The ship docks; I set my foot upon the new continent...

The gray autumn morning overshadows sea and land; everything still sways beneath me; I still feel the restless motion of the waves... Out of the mist the city rises... Beside me, eyes wide open, alive, the crowd hurries on. They do not feel the foreignness; only the newness. I hear one or another whisper to himself: America—as if he wished to impress it firmly upon his mind that he is truly here now, so far away!...

I stand alone on the shore. It is not of the new America, from which I am to demand the happiness my homeland has denied me, that I think—I think of another.

I see that little room, I see it as clearly as if I had left it yesterday, not so many years ago. On the table the lamp with the green shade, the embroidered armchair in the corner. Engravings hang upon the wall; the pictures blur into shadow. Anna is with me. She lies at my feet, her curly head resting against my knee; I must bend down to look into her eyes.

We have stopped chatting; the evening moves on, and the room grows quiet. Outside it begins to rain; we hear the drops striking the windowpane, slow and heavy. She smiles, and I bend toward her mouth. I kiss her lips, her brow, her eyes, which she has closed. My fingers play with the fine golden hair that curls behind her ears. I brush it back and kiss that sweet white place of skin behind her ear. She looks up again and laughs. "Something new," she whispers, as if astonished. I press my lips firmly behind her ear. Then I say, smiling, "Yes, I've discovered something new!" She bursts out laughing, and like a child, delighted, she cries, "America!"

How droll it was then! So wild and foolish! I see her face before me, how it looked up at me with those roguish eyes, and how from her red lips the cry rang out: "America!" How we laughed then, and how the fragrance that rose from her curls and drifted over our America intoxicated me...

And that grand name remained. At first we would always cry it out when, of the countless kisses, one strayed behind the ear; then we whispered it—then we merely thought it; yet it always came to consciousness.

A flood of memories rises within me. Once we

saw a large ship depicted on a poster column and, stepping closer, read: “From Liverpool—Arriving New York—From Bremen—Arriving New York”... We burst out laughing in the middle of the street, and she declared quite loudly, while people stood about: “You know, we’re traveling to America today!” The people looked at her in astonishment; especially a young man with a blond mustache who was smiling as well. That annoyed me greatly, and I thought: yes, he would no doubt like to come along...

Then once we were sitting in the theater—I no longer remember what play—when someone on stage spoke of Columbus. It was a play in iambic verse, and I recall the line: “—and when Columbus stepped upon the deck...” Anna nudged my arm lightly with hers; I looked at her and understood her disdainful glance. Poor Columbus... as if he had discovered the true America! After the theater, when we were sitting in a wine tavern, we spoke at length of the good man who had thought so highly of his pitiful America. In truth, we rather pitied him. For a long time I could picture him only standing with a sorrowful gaze upon the shore of his new continent, oddly enough wearing a top hat and a very modern overcoat, shaking his head in disappointment. Once we drew him together on the marble tabletop of a café and kept inventing new details. She insisted he must be smoking a cigar; moreover, in our picture the great discoverer carried an umbrella, and his top hat was crushed—naturally—because of the mutineers. Thus Columbus became for us the most comical figure in all world history. How wild! How foolish!...

And now I stand in the midst of the great, cold city. I am in the false America and dream of my sweet, fragrant America over there... And how long ago it all was! Many, many years. A pain, a madness comes over me that something like that is lost beyond recall. That I do not even know where a message from me, where a letter might reach her—that I know nothing, nothing at all of her anymore...

My path leads me farther into the city, and my porter follows behind me. I pause for a moment, close my eyes, and through a strange, deceptive play of the senses I am enveloped by the same fragrance that on that evening drifted over me from Anna’s curls, when we discovered America...

C Example Reference Instances

C.1 Davidson: Truth-Conditional Statements

German	English	Label
Max Houven ist 42 Jahre alt.	Max Houven is 42 years old.	true
Max Houven ist 52 Jahre alt.	Max Houven is 52 years old.	false
Max Houven ist gestorben und Jacques Mouvet ist Belgier.	Max Houven has died and Jacques Mouvet is Belgian.	true

Table 6: Example reference instances for the Davidson truth-conditional task.

C.2 Lewis: Truth in Fiction (Schnitzler)

German	English	P1	P2	P3
Anna war einmal des Ich-Erzählers Geliebte.	Anna was once the first-person narrator’s lover.	true	true	true
Der Ich-Erzähler hasste den Geruch von Annas Haar.	The first-person narrator hated the smell of Anna’s hair.	false	false	false
Annas olfaktorisches Profil kann präzise bestimmt werden.	Anna’s olfactory profile can be precisely determined.	false	true	false
Der Kaiser von Österreich-Ungarn heißt Franz-Josef.	The Emperor of Austria-Hungary is named Franz Joseph.	false	false	true

Table 7: Example reference instances for the Lewis truth-in-fiction task (Schnitzler).

C.3 Iser: Gaps (Schnitzler)

German	English	Label
Anna hat Locken.	Anna has curly hair.	true
Anna hat glattes Haar.	Anna has straight hair.	false
Anna hat den Ich-Erzähler verlassen.	Anna left the first-person narrator.	not fully articulated

Table 8: Example reference instances for the Iser gap-detection task (Schnitzler).

D Prompt Templates for Lewis

D.1 Basic Prompt

```
### Instruction
Read the following text.

### Text
'''
{CONTEXT_TEXT}
'''

### Input
Sentence: '''{Satz}'''
```

```
### Output
In relation to the given text, the sentence -
    '''{Satz}''' - is [MASKED]
```

D.2 Classification Prompt (Theory-Uninformed)

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is **true**
    or **false** with respect to the provided
    context.

# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''

# Label
```

D.3 Theory Prompt: Lewis Principle 1

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is **true**
    or **false**
with respect to the provided context.

# Principle
Follow David Lewis's Core Principle 1: A
    sentence of the form "In fiction F, φ" is
    true if and only if φ is true in every
    world where F is told as known fact rather
    than fiction.

# Logical rules
- Interpret "and", "or", and "if...then"
    strictly in the formal-logical sense.
- Ignore rhetorical, poetic, or emotional
    aspects unless they affect truth conditions.

# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''

# Label
```

D.4 Theory Prompt: Lewis Principle 2

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is **true**
    or **false** with respect to the provided
    context.

# Principle
Follow David Lewis's Core Principle 2: A
    sentence of the form "In fiction F, φ" is
    non-vacuously true if and only if some world
     where F is told as known fact and φ is
    true differs less from our actual world, on
    balance, than does any world where F is told
     as known fact and φ is not true.

# Logical rules
- Interpret "and", "or", and "if...then"
    strictly in the formal-logical sense.
- Ignore rhetorical, poetic, or emotional
    aspects unless they affect truth conditions.

# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''

# Label
```

D.5 Theory Prompt: Lewis Principle 3

```
# Role
You are a literary scholar.

# Task
Classify whether the given sentence is **true**
    or **false** with respect to the provided
    context.

# Principle
Follow David Lewis's Core Principle 3: A
    sentence of the form "In fiction F, φ" is
    non-vacuously true if and only if whenever w
     is one of the collective belief worlds of
    the community of origin of F, then some
    world where F is told as known fact and φ
    is true differs less from w, on balance,
    than does any world where F is told as known
     fact and φ is not true.

# Logical rules
- Interpret "and", "or", and "if...then"
    strictly in the formal-logical sense.
- Ignore rhetorical, poetic, or emotional
    aspects unless they affect truth conditions.
```

```
# Output
Respond with exactly one label:
true
false

# Context
'''
{CONTEXT_TEXT}
'''

# Input
Sentence: '''{Satz}'''

# Label
```