Enhanced Noun-Noun Compound Interpretation through Textual Enrichment

Bingyang Ye, Jingxuan Tu, James Pustejovsky
Brandeis University
{byye, jxtu, jamesp}@brandeis.edu

Abstract

Interpreting Noun-Noun Compounds remains a persistent challenge for Large Language Models (LLMs) because the semantic relation between the modifier and the head is rarely stated explicitly. Recent benchmarks frame Noun-Noun Compound Interpretation as a multiple-choice question. While this setting allows LLMs to produce more controlled results, it still faces two key limitations: vague relation descriptions as options and the inability to handle polysemous compounds. We introduce a dual-faceted textual enrichment framework that augments prompts. Description enrichment paraphrases relations into event-oriented descriptions instantiated with the target compound to explicitly surface the hidden event connecting head and modifier. Conditioned context enrichment identifies polysemous compounds leveraging qualia-role binding and assigns each compound with condition cues for disambiguation. Our method yields consistently higher accuracy across three LLM families. These gains suggest that surfacing latent compositional structure and contextual constraint is a promising path toward deeper semantic understanding in language models. [1]

1 Introduction

Noun-Noun Compounds (NNCs) such as *"olive oil"* and *"court approval"* are ubiquitous in English and many other languages. Despite their surface brevity, they encode rich semantic relations. *Olive* is an ingredient of *oil*; a *court* may either grant an *approval* or be the purpose for which *approval* is sought. Automatic Noun-Noun Compound Interpretation (NNI) is crucial for downstream tasks such as information extraction, coreference resolution and question answering (Nakov, 2008; Dima and Hinrichs, 2015; Lang et al., 2022).

Interpreting NNCs presents a challenge for Large Language Models (LLMs), particularly when it comes to pinpointing the underlying relationship between the head noun and the modifier. For example, given an NNC *"carrot cake"*, the model is asked to provide a paraphrased description *"a cake made of carrots"*, making the semantic relation between the head *cake* and modifier *carrot* explicit. A recent benchmark (Rambelli et al., 2024) recasts NNI as a nine-way multiple-choice task in which each relation is rendered by a fixed template to reduce the variability of the generated descriptions. Example (1) shows the template for relation **Comp(osition)-R(eversed)**. An instantiated description with the target compound *"olive oil"* is ***oil*** *is composed of* ***olive***.

(1) **Comp-R** $\rightarrow n_2$ *is composed of* n_1

LLMs reach respectable performance in this setting, attaining up to nearly 60% accuracy (Rambelli et al., 2024). However, closer inspection reveals two fundamental bottlenecks:

Description Vagueness The template-based relation descriptions overlook the eventive connection between modifier and head, which is vital for accurate reasoning (Downing, 1977; Levi, 1978). The paraphrases either miss a verb connecting the modifier and head, or the verb is too generic and can be interpreted in many ways, which makes it challenging for LLMs to understand the relation correctly.

Polysemy Roughly 55% of the compounds in the benchmark admit more than one plausible reading (e.g. *court approval*). Without contextual constraints, both humans and models struggle to commit to a single label, causing noisy supervision and artificially capping performance.

We argue that these obstacles stem not from model capacity but from underspecified inputs. We

[1] The source code and dataset is available at https://github.com/brandeis-llc/NNC-Enrichment.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25908–25923
November 4-9, 2025

therefore propose a textual enrichment framework that injects the missing semantic and pragmatic information directly into the prompt, requiring no parameter updates. Our framework enhances NNI from two facets. The first one is relation description enrichment. Building on the NNC relation taxonomy from Tratz (2011), we use an LLM to compose the relation definition with the target compound, producing a concrete, event-oriented description as the candidate options for the multiple-choice task. The enriched descriptions reduce the vagueness in the original ones due to lack of an accurate eventive connection and help models better interpret the NNC relations.

The second facet is conditioned context enrichment, which addresses polysemy. Building on Generative Lexicon (GL) theory (Pustejovsky, 1995), we treat a noun's meaning as structured by four qualia roles: Formal, Constitutive, Agentive, and Telic. In GL, the composition of a compound arises when open roles in the qualia structure of one noun are filled, or bound, by the other noun, producing a coherent meaning. Polysemy occurs when different bindings yield multiple plausible interpretations of a compound. To capture this, our framework first detects candidate polysemous compounds through qualia-role binding. For each possible reading, we then introduce a condition cue—a minimal state variable that distinguishes one interpretation from another (e.g., TRASH_IN_BAG = yes/no for the compound trash bag). Finally, we generate a short context sentence that makes the condition explicit (e.g., "The hospital construction cannot proceed without court approval" instantiates the cue COURT_ALREADY_EXISTS = yes). These cues supply mutually exclusive contextual constraints, enabling models to choose the appropriate relation and improving performance further when paired with context sentences.

Our contributions are threefold: 1) We propose a textual enrichment framework that refines relation descriptions and conditions on contextual constraints. 2) We create a dataset that consists of both monosemous and polysemous NNCs with explicit condition cues, filling a gap in existing resources. 3) Empirical results show consistent improvements using both facets of the textual enrichment framework across open and closed LLMs.

The remainder of the paper is organized as follows: §2 surveys related work on compound interpretation and textual prompt engineering; §3 describes the benchmark we work on; §4 details the first facet of the framework for relation description paraphrasing; §5 introduces the second part of the framework for identifying polysemous compounds and generating condition cues based on the reading. Finally §6 concludes.

2 Related Work

2.1 Noun-Noun Compound Interpretation

NNI has traditionally been approached as a classification task, where compounds are assigned to predefined semantic relations (Kim and Baldwin, 2005; Tratz and Hovy, 2010). However, such classification frameworks are inherently limited due to their coarse granularity and inability to account for compound interpretations that require sophisticated explanations beyond short relation labels or N-grams patterns (Shwartz and Dagan, 2018). To address these limitations, recent work adopts paraphrase-based formulations, where the goal is to generate natural language paraphrases that explain the relation between the compound constituents (Hendrickx et al., 2013; Shwartz and Waterson, 2018). This shift not only enhances interpretability but also aligns with the broader trend toward *textual enrichment* and the goal of this paper, where models generate contextually rich, semantically grounded explanations beyond rigid label sets.

The recent development of language models has led to zero-shot and few-shot interpretations for NNCs. Ponkiya et al. (2020) shows that T5-based models have encoded the relevant knowledge to NNC during pre-training. Coil and Shwartz (2023) demonstrates that GPT-3 models reach near-perfect performance on existing benchmarks. Nonetheless, further analysis revealed that current large language models often rely on memorized patterns from training corpora, limiting their generalization to novel or ambiguous compounds (Rambelli et al., 2024). In this work, we investigate whether linguistically textual enrichment can boost LLMs' semantic comprehension capability on NNC, especially the ones with competing readings and novel compounds. To the best of our knowledge, this is the first attempt of modeling human-like interpretation on polysemous NNCs.

Psycholinguistic work (Schmidtke et al., 2016) shows that readers often default to a dominant interpretation of ambiguous compounds, with lower relational entropy leading to faster recognition. While this highlights human biases in compound processing, our focus is different: we make com-

Relation	Count
COMP(OSITION)-R(EVERSED)	85
CONT(AINMENT)-R(EVERSED)	54
LOCATION	107
PARTONOMY	16
PROD(UCTION)-R(EVERSED)	13
PRODUCTION	47
PURPOSE	270
TOPIC-R(EVERSED)	66
US(A)G(E)-R(EVERSED)	10

Table 1: Counts of each relation in LEXICALIZED.

peting readings explicit and test whether LLMs can disambiguate them when given contextual cues.

2.2 Textual Enrichment

Textual enrichment methods such as paraphrasing and contextual sentence generation have been used to improve model performance across NLP tasks such as classification, semantic similarity, and question answering (Choi et al., 2021; Elazar et al., 2022; Tu et al., 2022, 2024c). Classic methods range from heuristic-based lexical substitutions to back-translation. Wieting et al. (2017) back-translates paraphrase pairs for training sentence embeddings, enabling effective textual enrichment through diverse rephrasings. Wei and Zou (2019) proposes a set of simple augmentation techniques such as synonym replacement and random insertion to enrich text to improve model robustness. More recently, Choi et al. (2021) defines the task of sentence decontextualization that enables sentence understanding with enriched context from generative language models. Tu et al. (2023, 2024b) introduce the dense paraphrasing method, which makes hidden semantic information explicit across both textual and multimodal contexts. Zhao et al. (2025) leverages LLMs to enrich event mentions into full sentences, facilitating the annotation process for constructing the event coreference dataset. We are the first to apply textual enrichment to NNI by explicitly surfacing event structure and adding contextual constraints in the LLM prompt.

3 Data

We use the LEXICALIZED and NOVEL NNC datasets (Rambelli et al., 2024) to evaluate the effectiveness of our textual enrichment methods for NNI. LEXICALIZED set includes 668 NNCs that span over nine relations. We show the relation inventories and statistics of the set in Table 1. NOVEL set includes a set of 124 novel NNCs that can be used to examine the models' generalizability on semantic relations. Compounds from the NOVEL set are derived from the existing NNCs by replacing the head or the modifier with one of its hypernyms using WordNet 3.0 (Fellbaum, 2010). The interpretation of each NNC is formulated as a multiple-choice classification task. Each relation is paraphrased into a single, tightly controlled sentence description as the option. The LLMs are prompted to select the most appropriate description for each compound.

4 Enriching NNC Relation Descriptions

The original LEXICALIZED and NOVEL datasets adopt a semantic relation template (Pepper, 2020) to paraphrase compound relations into short descriptions, allowing for more controlled evaluation of NNI. Although these descriptions are designed to generalize across a wide range of compounds, their broadness often makes it difficult for both humans and models to accurately identify the intended relation.

Building on the findings from Rim et al. (2023) that implicit predicates introduce crucial semantics, we recast broad relations into verb-centered paraphrases for clearer NNC interpretation. Consider the compounds and relations in example (2). The descriptions of relation TOPIC-R for compound *art class* and PURPOSE for *travel agency* are vague and too generalized, and are challenging for models to interpret.

(2) TOPIC-R: a **class** that is about **art**.
PURPOSE: an **agency** intended for **travel**.

To alleviate the confusions stemming from the template-based relation descriptions, our enrichment strategy operationalizes the "verb-composition" theory from Downing (1977) and Levi (1978) by paraphrasing each multiple-choice option into an event-explicit description, reinstating the hidden verb and clarifying the relation.

4.1 Methodology

We frame the enrichment of NNC and its relation into a paraphrasing task, prompting LLMs to select the most appropriate event verb(s) based on the head and modifier types and relation definition, and generate the corresponding enriched relation description .

We ask the OpenAI `o3` model to paraphrase the relation based on a more well-defined relation taxonomy (Tratz, 2011) because it has a more concrete

Relation Definition (Topic-R)

a n_2 that discusses, depicts, expresses, explains, teaches, symbolizes, praises, celebrates, and/or contains info/data related to the activity related to n_1

Example

($n_1 = art,\ n_2 = class$)

a class that teaches art.

Figure 1: Topic-R relation definition from Tratz (2011) and an example instantiation.

explanation for each relation and each definition contains a list of verb candidates. Since Rambelli et al. (2024) has already generated a mapping from their nine-way relation inventory to the relation taxonomy, we directly retrieve the relation definitions and include them in the context of prompt instruction.

Figure 1 illustrates the definition of relation TOPIC-R instantiated with the target compound *art class*. The relation definition provides candidate event verbs which implicitly allow the composition of the compound. Compared to the original description *"a class that is about art"*, the enriched one for compound *art class* highlights the concrete event *teach*, and makes it clearer that the TOPIC-R relation pertains more to engaging in conveying information related to art, which facilitates models to predict it as the desired option. LLM prompts are detailed in Appendix A.2

4.2 Experiments

We evaluate the enriched relation descriptions on both LEXICALIZED and NOVEL sets. Following the experimental settings in Rambelli et al. (2024), we formulate the NNI task as a multiple-choice classification problem and use the same prompt for the in-context learning with LLMs. During inference, the order of the multiple-choice options is randomized. We adopt two open LLMs (`Llama-2-7B-chat-hf` and `Mistral-7B-Instruct-v0.2`) and one closed LLM (`GPT-4o`) in the experiment. We retain the hyper-parameter configuration of Rambelli et al. (2024) and treat their original, template-based prompt as our *baseline*.

4.3 Results

We compare the accuracy of the same model in selecting the correct relation from the original descriptions versus the newly enriched descriptions to assess the impact of the enrichment. Experiment results in Table 2 show that relation description enrichment leads to a substantial improvement in identifying the correct relation on LEXICALIZED on all models. Both `GPT-4o` and `Mistral` outperform `Llama-2` by a large margin with `GPT-4o` reaching the top score. Notably, each model reaches its peak in the 1-shot setting, which is in line with the results in baseline. This interesting finding suggests that adding extra in-context examples could bring noise and hurt model performance.

	0-shot		1-shot		3-shot	
Model	Base.	Enr.	Base.	Enr.	Base.	Enr.
`Llama-2`	12.7	**24.7**	19.6	**30.5**	19.5	**29.9**
`Mistral`	59.1	**70.8**	59.1	**72.8**	56.5	**70.1**
`GPT-4o`	69.6	**74.4**	65.8	**73.7**	65.0	**70.8**

Table 2: Accuracy on LEXICALIZED set with baseline (Base.) and enriched descriptions (Enr.) as options.

The results in Table 3 and 4 show improvements on all models on NOVEL set, demonstrating that our method can generalize well on novel compounds. Models all achieve best results in 3-shot setting, illustrating that in-context examples are important for models to understand compounds that they might have not seen before.

	0-shot		1-shot		3-shot	
Model	Base.	Enr.	Base.	Enr.	Base.	Enr.
`Llama-2`	12.6	**22.3**	15.6	**32.9**	16.5	**34.1**
`Mistral`	34.5	**40.2**	58.0	**60.7**	53.4	**62.6**
`GPT-4o`	59.6	**63.1**	60.2	**64.7**	61.3	**66.4**

Table 3: Accuracy on NOVEL set with the same head.

	0-shot		1-shot		3-shot	
Model	Base.	Enr.	Base.	Enr.	Base.	Enr.
`Llama-2`	10.3	**22.3**	14.6	**30.8**	21.2	**31.5**
`Mistral`	30.1	**34.2**	47.9	**50.6**	30.5	**51.3**
`GPT-4o`	51.4	**55.7**	52.0	**57.8**	52.8	**58.3**

Table 4: Accuracy on NOVEL set with the same modifier.

4.4 Analysis

We report the per relation F1 scores on `Mistral` with 1-shot example in Table 5. The largest gains appear for PURPOSE and TOPIC-R, whose baseline descriptions are especially underspecified because they omit the implicit event. Importantly, the labels most frequently mistaken for these two,

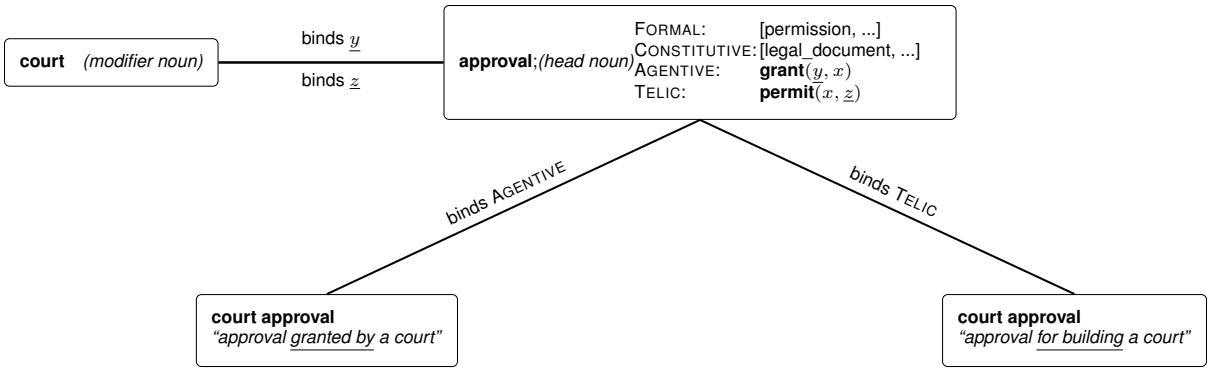

Figure 2: Multiple bindings for the compound *court approval*. The qualia structure of the head noun *approval* contains two unsaturated roles: AGENTIVE *grant*$(\underline{y}, x)$ and TELIC *permit*$(x, \underline{z})$. The modifier noun *court* can saturate either argument ($\underline{y}$ or $\underline{z}$), producing two context-dependent readings.

Relation	Base.	Enr.
COMP-R	74.9	81.6
CONT-R	43.0	48.9
PRODUCTION	62.5	62.7
PROD-R	35.2	46.5
LOCATION	65.0	66.1
USG-R	24.3	28.8
PURPOSE	54.6	72.3
TOPIC-R	62.7	74.3
PARTONOMY	50.0	55.1

Table 5: Per–relation F1 scores (%) of description enrichment on `Mistral` with 1-shot prompting.

i.e., PROD-R and CONT-R, also register sizeable improvements, indicating that enrichment reduces cross-class ambiguity rather than merely redistributing errors. The performance boost demonstrates that making the covert events overt aligns the prompt with long-standing linguistic theory and, as our experiments show, equips LLMs with the crucial semantic scaffold needed for accurate interpretation.

5 Enriching NNC Conditioned Contexts

We further discuss how textual enrichment handles polysemous NNCs which can admit more than one plausible reading through contextual constraints. In §4, we show that the NNC enriched descriptions lead to improvements on NNI. However, the task still suffers from inherent semantic ambiguity. The relationship between the head and modifier noun is often implicit and underspecified, allowing multiple plausible interpretations depending on context. Example (3) shows the two plausible readings of *"trash bag"*, the readings pivots on whether the bag is currently filled with trash or merely intended for that purpose.

(3) a. CONT-R: a **bag** that contains **trash**.
b. PURPOSE: a **bag** designed to hold **trash**.

Current NNC datasets usually assume a single correct semantic relation per compound. This simplification is flagged by the authors themselves as problematic because many compounds attract competing relational readings (Benjamin and Schmidtke, 2023; Rambelli et al., 2024). This observation motivates us to bridge the gap in current NNC datasets and address the issue of polysemy. We propose a conditioned context enrichment framework to detect NNCs that have more than one reading due to semantic ambiguity. We define a condition cue as a minimal, relation-agnostic state variable that distinguishes two plausible readings of a compound without encoding relation labels. And we provide condition cues to anchor compounds to these readings and thus guide the model toward the context-appropriate relations. Furthermore, context sentences are built upon these conditions to help models interpret different senses of the compounds.

5.1 Qualia Structure Binding for NNCs

The major challenge in interpreting polysemous NNCs comes from the identification of the implicit ambiguous relation between the modifier and the head. To understand the implicit relational composition of NNC, we leverage the qualia roles from Generative Lexicon (GL) theory (Pustejovsky, 1995), which has informed prior annotation schemes for compounds (Bouillon et al., 2012) and studies of semantic relation reliability (Yadav et al., 2017). GL is a lexical framework that treats a noun's meaning as a structured bundle of qualia roles, including TELIC, CONSTITUTIVE, AGEN-

Qualia role	Explanation	Example
FORMAL	What is it? What category does it belong to?	*container*
CONSTITUTIVE	What is it made of? What parts or materials compose it?	*plastic film*, fabric
AGENTIVE	How does it come into being? Who/what creates it?	*make*(x), sew(x)
TELIC	What is it for? What function or purpose does it serve?	*contain*$(x, \underline{y})$

Table 6: The four *qualia roles* that structure lexical meaning, illustrated with the noun ***bag***.

TIVE, and FORMAL. Table 6 shows the definition of qualia structure and its instantiation for word *bag*.

In GL, the composition of words is the combination of their qualia structures through the act of "binding". Binding occurs when an open variable (i.e., an unsatisfied slot) in one noun's qualia structure is filled by another noun. It is the mechanism that links the internal semantic requirements of one constituent to the referential content of the other, producing a coherent composite meaning. We assume that the semantic ambiguity in an NNC arises because different qualia roles of the head and the modifier are simultaneously compatible and can bind through different roles thus enabling more than one plausible semantic interpretations.

In Figure 2, the qualia structure of the head noun *approval* contains two independent variables that can be filled by the same modifier, thereby licensing two distinct bindings and, consequently, two readings. Its AGENTIVE role encodes the event $grant(y, approval)$, whose open slot y denotes the agent that issues the approval; its TELIC role encodes the purpose-oriented event $permit(approval, z)$, whose slot z denotes the object or activity for which the approval is sought. The noun *court* is semantically compatible with both variables: as an institution, it can act as the granting agent (*approval granted by a court*), and as a facility, it can be the very object requiring permission (*approval for building a court*). Because each binding completes a different event template in the qualia structure, the compound simultaneously supports an PRODUCTION interpretation and a PURPOSE interpretation. This dual satisfiability explains the compound's polysemy and illustrates how multiple open roles in a head noun, when matchable by the same modifier, naturally give rise to more than one relational reading.

5.2 Polyemous NNC Detection

We detect polysemous NNCs through an automatic enrichment framework powered by LLMs. We instruct the models to identify semantic ambiguity in NNCs through binding the generated qualia structures. We adopt the `o3` model using OpenAI API in the framework. We describe each step below.

Qualia Structure Generation We use the datasets in Rambelli et al. (2024) (LEXICALIZED and NOVEL sets) as the source datasets. Inspired by Tu et al. (2024a)'s automatic pipeline for augmenting meaning representations with sub-event structure, we cast this task as a structured enrichment problem: we generate qualia structures for heads/modifiers and apply qualia binding to expose and disambiguate their composition. Given a head or modifier of a compound, we generate its qualia structures. By the definitions of the qualia roles, the values for FORMAL and CONSTITUTIVE roles are a list of nominals while the values of AGENTIVE and TELIC are a list of verb phrases.

Relation Detection Given the generated qualia structures of the modifier m and head h, we instruct the model to follow the binding heuristics in Table 7 to automatically detect the composition relation. We prompt the OpenAI `o3` model to automatically detect the composition relation of compounds using qualia binding. Binding-based relation detection proceeds by first retrieving or automatically inducing the qualia structures of the modifier m and head h and then testing which binding heuristic in Table 7 is satisfied. For each compound we examine the qualia roles of both nouns and ask whether the modifier can fill one of the open slots of the head (or vice-versa) under the conditions listed in the rightmost column. $\otimes$ denotes successful binding.

For example, if m itself occurs in $h.C$, which is a list of constituents of h or is semantically similar to a constituent of h above a threshold τ, we register the binding $m \otimes h.C$ and label the compound COMP-R. Consider the compound *chicken soup*. The head noun *soup* has a CONSTITUTIVE role listing its possible ingredients (e.g., vegetables, meat, broth). The modifier *chicken* directly matches one of these constituents. Formally, we

Relation	Binding Role	Qualia Binding
COMP-R	$m \otimes h.\mathsf{C}$	$m \in h.\mathsf{C} \ \vee \ \mathrm{sim}(m,x) > \tau, x \in h.\mathsf{C}$
CONT-R	$m \otimes h.\mathsf{T}$	$\exists v \in h.\mathsf{T} : \mathsf{fills}\big(m, \rho(v)\big), \rho \in \{\textit{Theme}, \textit{Patient}\}, v \in \mathsf{CONTAIN}$
LOCATION	$h.\mathsf{F} \otimes m.\mathsf{F}$	$m.\mathsf{F} \cap \mathsf{PLACE} \wedge \mathrm{located\_at} \notin h.\mathsf{F}$
PARTONOMY	$h \otimes m.\mathsf{C}$	$h \in n.\mathsf{C} \ \vee \ \mathrm{sim}(h,x) > \tau, x \in m.\mathsf{C}$
PRODUCTION	$h \otimes m.\mathsf{A}$	$\exists v \in m.\mathsf{A} : \mathsf{fills}\big(h, \mathrm{agent}(v)\big)$
PROD-R	$m \otimes h.\mathsf{A}$	$\exists v \in h.\mathsf{A} : \mathsf{fills}\big(m, \mathrm{agent}(v)\big)$
PURPOSE	$m \otimes h.\mathsf{T}$	$\exists v \in h.\mathsf{T} : \mathsf{fills}\big(m, \rho(v)\big)$
TOPIC-R	$m \otimes h.\mathsf{T}$	$\exists v \in h.\mathsf{T} : \mathsf{fills}\big(m, \rho(v)\big), \rho = \textit{Topic}$
USG-R	$m \otimes h.\mathsf{T}$	$\exists v \in h.\mathsf{T} : \mathsf{fills}\big(m, \rho(v)\big), \rho = \textit{Instrument}$

Table 7: Compound relations and the corresponding binding types used to detect each relation. *h* refers to head, *m* refers to modifier. F, C, T, A refer to FORMAL, CONSTITUTIVE, TELIC, AGENTIVE roles respectively. $m \otimes h.\mathsf{C}$ means the modifier binds to the CONSTITUTIVE role of the head.

detect that m = chicken $h.C$ = {vegetables, meat, broth, ... }, satisfying the binding condition $m \otimes h.C$. This triggers the COMP-R label, since the modifier specifies what the head is composed of. Intuitively, the meaning is "soup composed of chicken." If we find a containment verb v in $h.T$ such that v assigns the modifier to a Theme or Patient role, we assign CONT-R relation. LOCATION is detected when the Formal role of the head denotes a PLACE and the modifier is situated at that location, while PARTONOMY is triggered when h is part of the $m.C$ or highly similar to one of its parts. If an event in $h.A$ takes m as its agent, we obtain PRODUCTION, whereas the reverse binding yields PROD-R. PURPOSE-, TOPIC-, and USAGE-oriented relations are all variations of m binds to $h.T$, differentiated by the thematic role that the modifier fills.

By iterating over these binding tests and returning the satisfied binding results, the system maps qualia evidence to a single compound relation; if multiple bindings fire, the compound is flagged as polysemous. We limit the number of plausible readings to two as there is no compound that can admit more than two readings in the working dataset.

Condition Cues To anchor the compound to a concrete condition and thus guide the model toward the context-appropriate relation, we enrich the dataset by providing a semantic condition for each reading of the compound. Given an NNC together with its annotated relation(s), we instruct the LLM to propose a set of *condition cues*. A condition cue is essentially a neutral, minimal state variable, expressed as a key-value pair. The key is a noun phrase describing the salient features of the compound or pinpointing the differences between two readings without encoding explicit grammatical roles or overtly signaling any particular relation. The value is either binary or categorical. Typical variables include IN_CURRENT_USE (yes/no) for CONT-R readings, PURPOSE_FULFILLED (yes/no) for PURPOSE readings.

Human adjudication To assess the reliability of the automatically assigned relations and condition cues, we engage human annotators in a post-hoc verification step. Each predicted relation is presented in context, and annotators only need to remove relations that are not reasonable or add missing relations. Adjudication is done by two graduate students in a U.S.-based university with Computational Linguistics backgrounds. We reach an inter-agreement score of 90.3%. The errors mostly stem from o3 picking the wrong argument type for the binding. Manual review requires revision of *15.6%* of the predicted relations, affecting *10.1%* of the compounds, which corresponds to an end-to-end pipeline accuracy of *84.4%*. The high accuracy from the automated pipeline also shows the usefulness of the qualia structure. This pipeline greatly reduce the workload on human annotators to detect and label polysemous NNCs. We also apply human validation to ensure that there is no data leakage in NNI inputs. Two PhD students are asked to check if the condition cues and the context sentences contain relation-suggesting phrases. If they do, then they are regenerated until they meet the criterion. About 25.7% of the generated condition-cues and 12.1% of the context sentences need regeneration.

Our framework eventually outputs a dataset CONDITIONED-LEXICALIZED (C-LEX) that aug-

Label	# examples			Ratio (%)		
	Mono.	Poly.	Total	Mono.	Poly.	Total
CONT-R	14	65	79	1.4	6.3	7.6
COMP-R	63	25	88	6.1	2.4	8.5
PURPOSE	136	283	419	13.2	27.4	40.6
PARTONOMY	7	41	48	0.8	4.0	4.6
USG-R	5	16	21	0.5	1.5	2.0
TOPIC-R	14	73	87	1.4	7.1	8.4
PRODUCTION	9	132	141	0.9	12.8	13.6
LOCATION	54	58	112	5.2	5.6	10.8
PROD-R	1	37	38	0.1	3.6	3.6
Totals	**303**	**730**	**1033**	**29.3**	**70.7**	**100.0**
Compound	303	365	668	45.4	54.6	100.0

Table 8: Statistics of C-LEX. Compound refers to the statistics of NNC instances.

ments compounds in LEXICALIZED with polysemy and corresponding condition cues. The resulting corpus-level statistics are reported in Table 8. We also create a conditioned dataset for NOVEL, namely CONDITIONED-NOVEL (C-NOVEL) [2].

5.3 Experiment

We adopt the same models as in §4.2 on C-LEX and C-NOVEL and their subsets to evaluate the effectiveness of our approach. Each compound c is associated with a set of relations $\langle r_1, ..., r_n \rangle$ and their condition cues $\langle \sigma_1, ..., \sigma_n \rangle$ obtained in §5.2. We describe our experimental settings below.

Condition Cue Only (*Cnd-Cue*) For every tuple $\langle c, ; r_k, ; \sigma_k \rangle$, we embed σ_k in the prompt, serving as an explicit semantic anchoring help LLMs predict the correct relation, especially differentiating readings of polysemous NNCs. The candidate relations are the original relation options. The model must output a single relation label.

Conditioned Context Sentence (*Cnd-Sent*) This experiment supplements the prompt with a synthetic sentence s_k automatically generated from σ_k using OpenAI `o3` model, e.g. "He tied the *trash bag* and left it at the curb." All other instructions are unchanged.

Plus Enriched Descriptions (*+Enr-Desc*) On top of the *Cnd-Sent* setting, we replace the original relation descriptions with the enriched descriptions.

[2]Novel dataset only replaces head nouns or modifier nouns of the lexicalized compounds with their hypernyms and we do not see any relational change due to the replacement. Therefore the statistics of C-Novel is exactly the same as C-Lex.

5.4 Results

We conduct evaluations on two levels: (i) **compound-level**, which considers pairs of readings for polysemous compounds, and requires correct predictions across both readings for the compound to be considered accurately classified; and (ii) **reading-level**, where performance is measured individually for each compound reading.

The compound-level evaluation results are shown in Table 9. Despite this stricter requirement, the enriched prompts still deliver consistent gains compare to the *baseline* where the models are run on data with no enrichment. The improvement indicates that the models are not merely matching isolated descriptions but are internalizing a coherent compositional representation that generalizes across a compound's full meaning space.

The relative gains are most pronounced for the lightweight, open-source models `Llama-2` and `Mistral`. Enrichment lifts their compound-level accuracy by more than 9 points, whereas `GPT-4o` shows a smaller but still reliable boost. This disparity indicates that once surfacing the implicit condition, smaller models no longer need to infer it and can instead devote their limited capacity to relational reasoning. Our strategy offers a parameter-efficient alternative to model scaling, allowing resource-friendly LLMs to approach the interpretive performance of much larger systems.

Table 10 reports reading-level accuracy on C-LEX. Enriching the prompts with conditioned cues raises performance for all three LLM families by 5-25 points. This suggests that our enrichment improves a model's ability to map a particular reading to the appropriate relation. For example, "*trash bag*" is often labeled as of PURPOSE relation (*a bag intended for trash*, while its gold label is CONT-R (*a bag contains trash)* because both relations are plausible and the model would randomly pick one from the two relations. Now with the enhancement of condition cue IN_CURRENT_USE (yes/no), models are able to distinguish the two readings and predict both relations correctly. To pinpoint where the gains arise, we examine the results on monosemous and polysemous subsets and trace the corrected readings under enrichment. On average, 54.8 % of these predictions come from polysemous compounds, while 45.2 % originate from monosemous ones. The near-even split shows that our condition cues uplift both compound types, yet the modest tilt toward poly-

Setting	Llama-2	Mistral	GPT-4o
Baseline	12.7 / 19.6 / 19.5	59.1 / 59.1 / 56.5	69.0 / 65.8 / 65.0
Cnd-Cue	35.5 / 30.1 / 27.8	65.0 / 68.1 / 68.3	69.3 / 70.6 / 72.0
Cnd-Sent	38.9 / 33.7 / 30.0	65.6 / 69.3 / 69.7	69.4 / 70.6 / 72.3
+Enr-Desc	**41.5** / 35.5 / 31.7	65.9 / 69.4 / **70.4**	69.9 / 71.6 / **72.4**

Table 9: Compound-level evaluation results (%) on C-LEX; each cell shows *0-shot / 1-shot / 3-shot* accuracy.

Setting	Llama-2	Mistral	GPT-4o
Baseline	12.7 / 19.6 / 19.5	59.1 / 59.1 / 56.5	69.0 / 65.8 / 65.0
Cnd-Cue	38.3 / 35.7 / 33.3	69.3 / 71.4 / 72.1	74.2 / 75.5 / 76.5
Cnd-Sent	40.8 / 37.6 / 34.9	69.7 / 72.0 / 72.9	74.9 / 77.1 / 77.8
+Enr-Desc	**44.4** / 39.3 / 40.1	70.4 / 72.1 / **73.3**	75.2 / 77.8 / **78.2**

Table 10: Reading-level evaluation results (%) on C-LEX; each cell shows *0-shot / 1-shot / 3-shot* accuracy.

semy suggests that condition cues are especially valuable when multiple readings compete. On both levels, results in 3-shot example settings generally output the best performance except for Llama model. This indicates that these examples can instruct models to predict the correct relation. We also see similar trends of improvement by adding context sentence and enriched descriptions, proving the effectiveness of these enrichment.

We further extend the enrichment to the novel compound set C-NOVEL. Results show that enrichment boosts reading-level accuracy by 7–10 points and compound-level accuracy by 6–9 points across all models. Because these gains occur on likely unseen lexical items, they show that the condition cues confer a generalizable grasp of compositional semantics rather than task-specific memorization, enabling the method to scale effectively as the inventory of compounds grows. We provide more detailed results in Appendix A.3.

5.5 Analysis

We investigate the effectiveness of each component in our conditioned context enrichment framework. We also add the enriched description options as an extra component from §4.2.

Condition Cues To evaluate the effectiveness of condition cue, we create a subset that only keeps the gold reading in the original dataset for each compound. As a result, we can evaluate on the same $\langle c, ; r_k, \rangle$ pairs as in LEXICALIZED. We have seen consistent improvements across all models, suggesting that enriching with conditions can enhance the NNI task.

Per-relation analysis (Appendix A.3) shows F1 gains for seven of the nine labels once condition cues are added. The most pronounced improvements appear for commonly conflated pairs such as CONT-R and PURPOSE, confirming that the cues help the model decide between readings like those available to *trash bag*. F1 drops only for PARTONOMY and PROD-R. The only two relation that have seen degraded performances are PARTONOMY and PROD-R. This is because more than double the number of new readings are introduced for these two relations. It also reveals that disambiguating these two relations from other competing relations are rather challenging.

Context Sentence Various settings of experiments on all sets of our data show that supplementing the condition cue with a context sentence yields an additional rise in accuracy. The improvement suggests that a discourse-level cue, phrased in natural language, helps the model ground the abstract state constraint in a concrete scenario. A per-relation breakdown (Appendix A.3) pinpoints the largest gains in CONT-R and PARTONOMY. Consider *mail box*, which admits (i) a PURPOSE reading *"a box intended to hold mail"* and (ii) a CONT-R reading *"a box that currently contains mail."* The condition cue MAIL_DELIVERED = YES proved ambiguous, and the model selected the PURPOSE relation. Augmenting the prompt with the sentence *"Our front-porch mail box is crammed with letters and catalogs waiting to be sorted."* explicitly depicts the box in active use, disambiguates the reading, and leads the model to the correct CONT-R label. This pattern illustrates why context sentences, by instantiating the condition cue in a realistic setting, offer an effective contextual constraint for accurate relation selection.

6 Conclusion

In this paper, we propose a textual-enrichment framework to enhance the NNI task by surfacing compound-specific events and providing semantic condition constraints to compounds with competing readings. Comprehensive experiments show that LLMs consistently benefit from these enrichment, yielding improvements both at the granular reading level and at the more demanding holistic compound level, and can extend to novel compounds. Beyond boosting accuracy, our method offers a parameter-efficient alternative to scaling and yields structured artifacts, i.e., eventful paraphrases and condition cues, that can be reused for explanation, retrieval, and consistency checking.

Limitations

Thus far we have only applied the enrichment framework to the LEXICALIZED set, whose modest size facilitated rapid iteration and oracle comparison. Given our framework is fully automated, the system can generate event-explicit descriptions and conditions cues at scale. This scalability opens the door to enriching much larger resources, e.g., Tratz (2011). In future work we will exploit this property to release high-coverage enriched NNC datasets that can serve as more demanding test beds for compositional semantics in LLMs.

Another limitation of our current implementation is that it handles at most two alternative readings per compound. This binary design reflects the practical observation that triple- or higher-order ambiguities are rare in the datasets we studied, yet it also hard-codes an upper bound on the framework's expressive power. Extending the framework to $n > 2$ readings would require a richer pool of orthogonal state variables and a more flexible selection mechanism, and would offer a sharper test of an LLM's capacity to navigate densely polysemous compounds. Investigating such multi-way disambiguation remains an important direction for future work.

Finally, our study is limited to English compounds. While English NNCs provide a rich testbed, cross-linguistic variation in compounding is well documented, and it remains unclear whether our enrichment framework would transfer directly to languages with different compounding strategies (e.g., German, Chinese). We leave multilingual extensions and evaluations for future work.

References

Shaina Benjamin and Daniel Schmidtke. 2023. Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory & Cognition*, 51(5):1170–1197.

Pierrette Bouillon, Elisabetta Jezek, Chiara Melloni, and Aurélie Picton. 2012. Annotating qualia relations in Italian and French complex nominals. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1527–1532, Istanbul, Turkey. European Language Resources Association (ELRA).

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Albert Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

Corina Dima and Erhard Hinrichs. 2015. Automatic noun compound interpretation using deep neural networks and word embeddings. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 173–183, London, UK. Association for Computational Linguistics.

Pamela Downing. 1977. On the creation and use of english compound nouns. *Language*, pages 810–842.

Yanai Elazar, Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2022. Text-based NP enrichment. *Transactions of the Association for Computational Linguistics*, 10:764–784.

Christiane Fellbaum. 2010. *WordNet*, pages 231–243. Springer Netherlands, Dordrecht.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Inga Lang, Lonneke Plas, Malvina Nissim, and Albert Gatt. 2022. Visually grounded interpretation of noun-noun compounds in English. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 23–35, Dublin, Ireland. Association for Computational Linguistics.

Judith N Levi. 1978. *The syntax and semantics of complex nominals*. Academic Press, New York.

Preslav Nakov. 2008. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Artificial Intelligence: Methodology, Systems, and Applications: 13th International Conference, AIMSA 2008, Varna, Bulgaria, September 4-6, 2008. Proceedings 13*, pages 103–117. Springer.

Steve Pepper. 2020. The bourquifier: An application for applying the hatcher-bourque classification (version 3)[ms excel].

Girishkumar Ponkiya, Rudra Murthy, Pushpak Bhattacharyya, and Girish Palshikar. 2020. Looking inside noun compounds: Unsupervised prepositional and free paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

4313–4323, Online. Association for Computational Linguistics.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Giulia Rambelli, Emmanuele Chersoni, Claudia Collacciani, and Marianna Bolognesi. 2024. Can large language models interpret noun-noun compounds? a linguistically-motivated study on lexicalized and novel compounds. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11823–11835, Bangkok, Thailand. Association for Computational Linguistics.

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Daniel Schmidtke, Victor Kuperman, Christina L Gagné, and Thomas L Spalding. 2016. Competition between conceptual relations affects compound recognition: The role of entropy. *Psychonomic bulletin & review*, 23(2):556–570.

Vered Shwartz and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211, Melbourne, Australia. Association for Computational Linguistics.

Vered Shwartz and Chris Waterson. 2018. Olive oil is made *of* olives, baby oil is made *for* babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 218–224, New Orleans, Louisiana. Association for Computational Linguistics.

Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-03.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Uppsala, Sweden. Association for Computational Linguistics.

Jingxuan Tu, Timothy Obiso, Bingyang Ye, Kyeongmin Rim, Keer Xu, Liulu Yue, Susan Windisch Brown, Martha Palmer, and James Pustejovsky. 2024a. GLAMR: Augmenting AMR with GL-VerbNet event structure. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7746–7759, Torino, Italia. ELRA and ICCL.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 39–49, Nancy, France. Association for Computational Linguistics.

Jingxuan Tu, Kyeongmin Rim, and James Pustejovsky. 2022. Competence-based question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1521–1533, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. 2024b. Dense paraphrasing for multimodal dialogue interpretation. *Frontiers in artificial intelligence*, 7:1479905.

Jingxuan Tu, Keer Xu, Liulu Yue, Bingyang Ye, Kyeongmin Rim, and James Pustejovsky. 2024c. Linguistically conditioned semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1161–1172, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Prabha Yadav, Elisabetta Jezek, Pierrette Bouillon, Tiffany J Callahan, Michael Bada, Lawrence E Hunter, and K Bretonnel Cohen. 2017. Semantic relations in compound nouns: Perspectives from inter-annotator agreement. *Studies in health technology and informatics*, 245:644.

Jin Zhao, Jingxuan Tu, Bingyang Ye, Xinrui Hu, Nianwen Xue, and James Pustejovsky. 2025. Beyond benchmarks: Building a richer cross-document event coreference dataset with decontextualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3499–3513, Albuquerque, New Mexico. Association for Computational Linguistics.

	CONT.	COMP.	PURP.	PART.	USG.	TOP.	PROD.	PROD-R.	LOC.
CONT.	–	4	49	0	4	1	7	0	0
COMP.		–	12	4	1	2	0	2	0
PURP.			–	20	4	45	111	15	27
PART.				–	0	4	0	0	13
USG.					–	1	1	0	5
TOP.						–	1	12	1
PROD.							–	1	5
PROD-R.							–	–	7
LOC.									–

Table 11: Frequency of relation co-occurrence in the polysemous compounds in C-LEX.

A Appendix

A.1 Statistics

Table 11 shows the number of relation co-occurrences for polysemous compounds in C-LEX.

A.2 Prompts

In this section, we demonstrate various prompts that we use to either generate intermediate artifacts or probe LLMs on NNI task. Figure 3 illustrates the prompt examples to generate enriched relation descriptions. The two variations are for relation with one definition and two definitions respectively. Figure 4 shows an example prompt for NNI task with the enriched description options. The model is asked to select the most appropriate descriptions. Figure 5 shows the prompt to automatically generate qualia structures for a single noun. Figure 6 shows the prompt for automatically extracting possible composition relation by checking if qualia structure binding is possible. Each relation is assigned with a set of rules that model will check if the binding rules are satisfied and admit the corresponding relation. Figure 7 and 8 show the prompts that instruct LLM to generate condition cues for compounds given its reading(s). Figure 9 and 10 show the prompts for converting the abstract condition cues into discourse-level, contextualized sentences.

A.3 Results

In this section, we include results evaluated against different components of our enrichment framework. Table 12 and 13 show the results on C-NOVEL with different experiment settings, suggesting the generalizability of our enrichment framework. Table 14 shows the evaluation results on all the readings in LEXICALIZED, which is essentially a subset of C-LEX. Table 15 shows the relation-wise evaluation results of the condition enrichment framework.

Setting	Llama-2	Mistral	GPT-4o
Baseline	12.6 / 15.6 / 16.5	34.5 / 58.0 / 53.4	59.6 / 60.2 / 61.3
Cnd-Cue	30.5 / 33.1 / 34.8	40.0 / 63.9 / 64.5	63.4 / 66.6 / 68.5
Cnd-Sent	31.7 / 33.3 / 34.1	39.7 / 65.4 / 65.8	65.0 / 66.8 / 68.5
+Enr-Desc	36.5 / 39.0 / 41.6	45.1 / 67.1 / 68.7	67.1 / 68.4 / 69.2

Table 12: Compound-level evaluation results (%) on C-NOVEL-SAMEHEAD; each cell shows *0-shot / 1-shot / 3-shot* accuracy.

Setting	Llama-2	Mistral	GPT-4o
Baseline	10.3 / 14.6 / 21.2	30.1 / 47.9 / 30.5	51.4 / 52.0 / 52.8
Cnd-Cue	26.3 / 31.3 / 32.6	37.5 / 56.4 / 57.8	56.9 / 59.3 / 59.4
Cnd-Sent	27.1 / 27.9 / 30.4	36.9 / 57.1 / 58.5	57.6 / 60.7 / 61.2
+Enr-Desc	30.5 / 31.3 / 31.8	42.3 / 60.0 / 62.7	60.9 / 61.6 / 62.0

Table 13: Compound-level evaluation results (%) on C-NOVEL-SAMEMOD; each cell shows *0-shot / 1-shot / 3-shot* accuracy.

A.4 Model Details

We follow the model hyperparameters in Rambelli et al. (2024). All artifacts for enrichment (e.g., enriched descriptions, condition cues, context sentences, etc.) are generated by `o3` model using OpenAI API. `Llama-2-7B-chat-hf`, `Mistral-7B-Instruct-v0.2` and `GPT-4o` are used to do NNI task. The open-source LLMs, i.e., `Llama-2-7B` and `Mistral-7B` are run on NVIDIA RTX A6000 with these hyperparameters:

```
Temperature: 0,
do_sample: False,
top-k: 10,
top-p: 5,
max-tokens: 50,
frequency and presence penalty: 0
```

One Definition Prompt (PURPOSE)

```
Given n2 = {n2}, n1 = {n1}, choose the most
appropriate words in the definition and generate
the sentence. If there are slashes, choose
only one word from the options. Do not add any
word except what is already in the definition and
n1 and n2:

a n2 that is designed to perform(s)/engage(s)
in/finance(s) the activity related to n1.

Only return the sentence
```

Multi-Definition Prompt (CONT-R)

```
Given n2 = {n2}, n1 = {n1}, choose the most
appropriate definition based on the word type of
{n1} and generate the sentence with the most
appropriate words. If there are slashes, choose
only one word from the options. Do not add any
word except what is already in the definition and
n1 and n2:

Definition 1: a n2 that physically contain(s)/
hold(s)/define bound(s) of n1(s).
Definition 2: a n2 that much/many n1(s) live
in/at/on.

Only return the sentence.
```

Example Instantiations (`n1 = trash, n2 = bag`)

PURPOSE	CONT-R
*a **bag** is designed to engage in the activity related to **trash**.*	*a **bag** that physically holds **trash**.*

Figure 3: Two prompts for LLM based on the number of relation definitions (top) and their concrete instantiations (bottom) for compound *trash bag*. Other relations can be obtained by replacing the relation definition.

Setting	Llama-2	Mistral	GPT-4o
Baseline	12.7 / 19.6 / 19.5	59.1 / 59.1 / 56.5	69.0 / 65.8 / 65.0
Cnd-Cue	39.1 / 33.4 / 31.6	74.7 / 75.6 / 77.0	80.2 / 80.4 / 80.4
Cnd-Sent	42.1 / 39.8 / 34.1	75.0 / 77.2 / 77.4	80.7 / 80.7 / 81.1
+Enr-Desc	44.9 / 39.9 / 35.4	75.6 / 78.2 / 77.9	80.8 / 81.2 / 81.8

Table 14: Evaluation results (%) on all readings in LEXICALIZED; each cell shows *0-shot / 1-shot / 3-shot* accuracy.

Relation	Baseline	Cnd-Cue	Cnd-Sent	+Enr-Desc
COMP-R	73.6	83.7	83.7	83.7
CONT-R	43.0	70.1	78.5	78.5
PRODUCTION	60.6	69.7	72.4	76.0
PROD-R	34.7	24.2	24.2	24.2
LOCATION	65.0	77.5	76.3	76.3
USG-R	25.1	31.3	35.3	35.3
PURPOSE	50.8	75.9	75.9	82.0
TOPIC-R	62.7	70.6	73.0	73.0
PARTONOMY	52.1	42.5	49.0	49.0

Table 15: Per–relation F1 scores (%) of conditioned context enrichment on `Mistral-7B-Instruct` with 3-shot prompting.

```
Question:
Which of the following is the most
likely description of "trash bag"?

[1] A bag that is made of trash where
    trash is one of the primary
    ingredients that make up bag.

[2] A bag that is a part of trash.

[3] A bag that uses trash to perform
    where trash is the tool.

[4] A bag that trash creates.

[5] An trash that bag creates.

[6] A bag that physically holds trash.

[7] A bag that contains a plan about trash.

[8] Trash is the location where bag is at.

[9] A bag that is designed to perform
    the activity related to trash.

Answer:
```

Figure 4: Multiple-choice prompt for the primary LLM (example for *trash bag*).

```
You are an expert in lexical semantics,
specifically Generative Lexicon (GL) theory.
Your task is to generate the qualia structure
for a given English noun, focusing on its most
common or contextually relevant meaning.
Provide the most salient values for each role:

- FORMAL: {FORMAL ROLE Definition}
- CONSTITUTIVE: {CONSTITUTIVE ROLE Definition}
- TELIC: {TELIC ROLE Definition}
- AGENTIVE: {AGENTIVE ROLE Definition}

If a specific qualia role genuinely does not apply
or cannot be determined for the noun's identified
meaning, output an empty list \texttt{[]} for that
role's value. DO NOT return an empty string for
the entire output if other roles can be filled.

Input Noun: {noun}
```

Figure 5: Prompt for eliciting the qualia structure of an English noun.

```
### System
You are a linguistically trained analyst specialised in Generative-Lexicon (GL) qualia roles.

### User
Determine whether the MODIFIER can plausibly act as the AGENT / CREATOR of the HEAD noun,
i.e. perform the AGENTIVE actions that bring the HEAD into existence.

---

- HEAD noun & AGENTIVE inventory
HEAD = {head_noun}
HEAD.AGENTIVE = {agentive_role}

- MODIFIER candidate
MODIFIER = {modifier}

---

#### Decision procedure
1. Accept the clause only if it is grammatical and semantically natural
2. Mark an AGENTIVE value compatible if the clause passes step 1.
```

Figure 6: Chain-of-thought prompt for relation PROD-R.

```
### System
You are an expert in lexical semantics. You are given a compound word and two relation labels
referencing possible head-modifier relations.
Each relation is a reading plus an explanatory sentence.
Your goal is to produce short, neutral state variables to disambiguate the two readings.
- Do NOT use words that reveal the labels.
- DO NOT reveal grammatical structure (e.g., agent/patient).
- ONLY use semantic information.
- Express each condition as KEY = value, where KEY is a
  noun phrase in CAPS and value is "yes"/"no" or a short noun.
- Return only the JSON specified below.
- Do NOT include any extra text or explanation.

### User
Compound: {compound}

Relation labels and descriptions:
{labels_senses}

### Task
1. Choose state variable(s) that differentiate the readings.
2. Default to one variable; use two only if needed.
```

Figure 7: Prompt for generating state-variable-based condition cues for polysemous compounds.

```
### System
You are an expert in lexical semantics.
Given a compound and its reading, output one short, relation-agnostic state variable whose values
represent the reading.
- The state variable must encode ONLY semantic world-knowledge,
  NOT surface grammar, argument structure, or words that betray the relation label
  (e.g., avoid “IS_HELD”, “IS_MADE_OF”, “IS_AGENT”).
- Use the format STATE_NAME = value
- STATE_NAME: UPPER-CASE noun phrase (<= 3 words)
- value: “yes”/“no” or noun phrase (<= 3 words)
- The state variable must align with the reading.

### User
Compound: {compound}

Reading (label: definition)
{labels_sense}
```

Figure 8: Prompt for generating state-variable-based condition cues for monosemous compounds.

```
### System
You are a careful contextual writer.
Your job is to craft ONE natural English sentence for each sense of a
polysemous noun–noun compound.
The sentence must
- include the compound verbatim;
- make the provided state assignments true;
- stay neutral—do not mention the state-variable names or any relation
label (e.g. “usage”, “containment”).

Return the sentences in the JSON format shown at the end and nothing else.

### User
Compound  : {compound}

Reading A : {reading_a}
Relation for Reading A: {relation_a}
Condition A   : {state_variable_a}

Reading B : {reading_b}
Relation for Reading B: {relation_b}
Condition B   : {state_variable_b}

### Task
1. Write one sentence for each reading such that the state variables
   are true and the meaning of the compound corresponds to the reading.
2. Try to make the sentence natural and fluent.
3. The sentence for Reading A should differentiate the compound from Reading B.
4. Keep each sentence <= 25 words; avoid technical jargon.
```

Figure 9: Prompt for generating conditioned context sentences for each reading of a polysemous compound.

```
### System
You are a careful contextual writer.
Your job is to craft ONE natural English sentence for to represent the sense of a noun–noun compound.
The sentence must
- include the compound verbatim;
- align with the provided state assignment;
- stay neutral—do not mention any relation label (e.g. “usage”, “containment”).

Return the sentence in the JSON format shown at the end and nothing
else.

### User
Compound  : {compound}

Reading: {reading}
Relation for Reading A: {relation}
State: {state}

### Task
1. Write one sentence for each reading such that the state variable
is true and the meaning of the compound coresponds to the reading.
2. Try to make the sentence natural and fluent.
3. Keep each sentence <= 25 words; avoid technical jargon.
```

Figure 10: Prompt for generating conditioned context sentences for each reading of a monosemous compound.