Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Context Aware Paraphrasing of Noun Compounds for Robust Interpretation

Pushpak Bhattacharyya

Indian Institute of Technology Bombay, Mumbai

Abstract

Noun compounds (NCs) such as *chocolate cake* or *student protest* are ubiquitous in natural language yet notoriously ambiguous when interpreted in isolation. Prior work paraphrases NCs into prepositional or free-form variants but largely ignores the sentential and discourse context that often determines the intended relation. We propose CANCI, a context-aware paraphrasing framework that conditions on local context and retrieved usage exemplars to produce faithful, diverse paraphrases with calibrated confidence. CANCI couples a context encoder with a sequence generator and fuses evidence from a dense retriever using Fusion-in-Decoder. To make outputs reliable for downstream insertion (e.g., machine translation or information extraction), we calibrate paraphrase probabilities and construct coverage-controlled *top*-k sets via conformal prediction. Across standard NC benchmarks augmented with sentence contexts, CANCI improves isomorphic and non-isomorphic scores, reduces expected calibration error, and yields safer *top*-k outputs at target coverage. Human evaluations show higher adequacy and fluency compared to isolate-only baselines. We release code, data recipes, and analysis protocols to facilitate reproducibility.

Keywords: noun compounds, paraphrasing, retrieval-augmented generation, calibration, conformal prediction

1. Introduction and Background

Noun compounds (NCs)—two or more contiguous nouns forming a single semantic unit, such as *student protest*, *virus lab*, or *coffee mug*—are pervasive in text across domains (news, scientific writing, social media) and central to downstream tasks like information extraction (IE), machine translation (MT), and question answering (QA) [1, 2]. They are compact by design: overt relational markers (prepositions, case/postpositions, verbal heads) that normally express the link between the modifier and the head are omitted. The same surface form therefore licenses multiple readings; for example, *student protest* may denote a protest *by* students or *about* students, while *virus lab* might refer to a lab *studying* viruses or *containing* them. Choosing the wrong reading in practical systems propagates as factual errors in IE, adequacy errors in MT, or misleading retrieval summaries [3].

Human readers rarely interpret NCs in isolation; local sentence context and wider discourse typically supply decisive cues. Useful signals include syntactic patterns (governing verbs, arguments, appositives), semantic compatibility (selectional preferences and world knowledge), discourse phenomena (coreference and topical continuity), and lexical collocations surrounding the NC. Consider:

Corresponding author: Pushpak Bhattacharyya (pb@cse.iitb.ac.in)

BCDS 2021, Vol. 2(1), 44-54. DOI: 10.71448/bcds2121-5

*After weeks of negotiations, the **student protest** forced the council to reverse the tuition hike* [4]. The event structure and agentivity of *students* favor the AGENT reading (*protest by students*). In contrast: *The committee heard testimony on the **student protest** scheduled by outside organizers.* Here, *scheduled by outside organizers* weakens the AGENT reading and supports a TOPIC reading (*about students*). Systems that paraphrase NCs without conditioning on these cues tend to default to globally frequent prepositions or hallucinate content inconsistent with the sentence [4].

We formulate the task as follows: given an NC occurrence (h, m) embedded in a sentence context C (optionally a broader discourse window D), the goal is to generate a ranked list of paraphrases that are faithful *in that context* [5]. Two realizations are useful for practice: prepositional paraphrases of the form h *prep* m, which enable controllable normalization and slotting into rule-based pipelines, and free-form verbal paraphrases, which can be inserted into fluent text for downstream applications. Evaluation considers surface agreement with references via Iso and NonIso matching, diversity with Distinct-n, and reliability with probability calibration measures such as Expected Calibration Error (ECE) and the empirical coverage of set-valued *top*-k predictions.

Prior isolate-only approaches exhibit recurring failure modes. They often make context-incompatible choices by selecting majority prepositions like *of*, *for*, or *in* that contradict the governing predicate [6]. They may introduce lexical over-specificity, adding unwarranted content words (e.g., asserting *manufacture* when the context implies *study*). They are typically ambiguity-blind, returning a single deterministic paraphrase even when multiple readings are equally plausible, which gives users false confidence. Finally, their scores are systematically miscalibrated—appearing highly confident on wrong choices—complicating safe insertion into MT/IE pipelines [7].

Our perspective is that faithful NC interpretation demands both conditioning on context and grounding in real usage [8]. Conditioning aligns the generator with local sentence constraints; grounding mitigates hallucination by retrieving attested paraphrase patterns from large corpora. Because practical systems must also decide *when to trust* a generated paraphrase, we treat ranking as probabilistic prediction, calibrate the model's confidence, and construct coverage-controlled *top*-k sets so users can trade set size for reliability [9].

This leads to the central gap addressed in this work: existing NC paraphrasing systems largely operate on isolated compounds and report only surface-matching metrics, with limited attention to sentence/discourse context and little to no analysis of calibration or coverage in ranked outputs. The absence of these ingredients obstructs safe deployment in MT/IE/QA, where an uncalibrated top-1 choice can yield cascading errors [10].

We introduce CaNCi, a context-aware, retrieval-augmented paraphrasing framework that conditions generation on a learned representation of the sentence (and optional discourse) together with retrieved usage snippets. A dense retriever supplies attested exemplars, and a sequence generator (e.g., T5) integrates them using Fusion-in-Decoder to guide preposition choice and surface realization [11]. We calibrate paraphrase probabilities using temperature scaling on a held-out set [12] and form set-valued predictions via split conformal prediction [13], delivering *top*-k outputs that meet a user-specified coverage target on unseen data. To measure real-world utility, we propose a context-augmented evaluation protocol that supplements Iso/NonIso matching with diversity, ECE, empirical coverage, and human judgments of adequacy and fluency under actual sentence contexts [14].

Within related work, NC interpretation has been explored through rule templates, distributional semantics, and neural generators trained on SemEval-style annotations [15], but the dominant as-

sumption is still isolate-only processing. Retrieval-augmented generation has improved factuality and reduced hallucinations in QA and summarization by grounding generation in retrieved text [16]. Probability calibration and distribution-free risk control via conformal prediction have been developed for classification and reading comprehension [17], yet are rarely applied to ranking over strings such as paraphrases. CANCI unifies these threads by injecting sentence/discourse evidence, grounding in usage exemplars, and delivering calibrated, coverage-controlled predictions that are practical for deployment [18].

Finally, we anticipate broad impact for computer and data science. Context-aware, calibrated paraphrasing supports safer IE pipelines by offering alternatives with transparent confidences, improves MT adequacy for NC heavy language pairs, and enables interactive systems that expose multiple readings rather than a brittle single guess. Beyond NCs, the general recipe—context conditioning, retrieval grounding, and conformal set prediction—extends to reliable text generation components throughout data-centric NLP systems.

2. Method: Context-Aware and Calibrated Paraphrasing

Given an NC occurrence (h, m) embedded in a sentence C (optionally expanded with a document window D), our goal is to output a ranked set of paraphrases that are faithful to the meaning signaled by C. We target two complementary realizations: a prepositional form of the pattern h *prep* m, which provides a normalized and easily controllable rendering for rule-based pipelines, and a free-form verbal rendering that can be inserted into fluent text. Throughout, we denote the full input by $x = (h, m, C, D)$ and the model's ranked candidates by $\{p_j\}$ with scores $s(p_j \mid x)$.

The model CANCI operationalizes the "context + usage grounding" perspective from the introduction by combining a context encoder, a dense retriever, and a generator. First, a transformer encoder f_θ (e.g., RoBERTa) encodes the sentence (and, when available, a compact discourse window) into a contextual representation $\mathbf{c}$. To focus the encoder on the compound, we surround the head and modifier with special boundary tokens and retain a small left/right window around the NC to preserve governing predicates and arguments. Second, a late-interaction dense retriever $\mathcal{R}$ indexes a large corpus of attested NC usages. At inference, a query representation built from (h, m, C) is matched against the index to return a small set of usage snippets $\{u_j\}_{j=1}^{R}$ that exemplify plausible readings of the same or closely related compounds. Retrieval uses approximate nearest-neighbor search and in-batch/hard-negative training so that returned snippets are semantically close yet diverse; we deduplicate near-duplicates and filter snippets with low lexical or semantic overlap to avoid spurious patterns. Third, a sequence generator g_ϕ (e.g., T5) conditions on $[h, m, \mathbf{c}, \{u_j\}]$ using Fusion-in-Decoder: each retrieved snippet is encoded separately and the decoder attends across them alongside the contextual representation, which empirically reduces hallucination and improves preposition choice. For the prepositional variant, we employ constrained decoding so that the first content token is chosen from a small, curated preposition inventory, after which the decoder freely realizes the remaining surface form; for the free-form variant, we use standard decoding. We generate an N-best list with beam search, apply light n-gram blocking for diversity, and retain K unique paraphrases for downstream calibration and set construction.

Inputs are linearized as a single sequence so that both context and retrieved evidence are available at decode time:

$$\texttt{NC:}\ h\#m\ \|\ \texttt{CTX:}\ C\ \|\ \texttt{RET:}\ u_1\langle\text{SEP}\rangle \ldots \langle\text{SEP}\rangle u_R.$$

This format makes the conditioning explicit and allows ablations that remove either the context or the retrieved block to quantify their contributions.

Training minimizes a composite objective that couples faithfulness, alignment, and probability mass shaping. The primary term is sequence-level cross-entropy $\mathcal{L}_{\mathrm{XE}}$ to reference paraphrases. To ensure the generator actually uses contextual cues rather than shortcutting to globally frequent prepositions, we add an InfoNCE-style contrastive term $\mathcal{L}_{\mathrm{con}}$ that aligns the context representation $\mathbf{c}$ with representations of faithful paraphrases while pushing away mismatched pairs. Concretely, if $r(p)$ is a frozen encoder embedding of paraphrase p, then for a positive pair $(\mathbf{c}, r(p^+))$ and a set of negatives $\mathcal{N}$, we write

$$\mathcal{L}_{\mathrm{con}} = -\log \frac{\exp(\mathrm{sim}(\mathbf{c}, r(p^+))/\tau)}{\exp(\mathrm{sim}(\mathbf{c}, r(p^+))/\tau) + \sum_{p^- \in \mathcal{N}} \exp(\mathrm{sim}(\mathbf{c}, r(p^-))/\tau)}.$$

Finally, we encourage the model to concentrate probability on faithful candidates using a coverage regularizer $\mathcal{L}_{\mathrm{cov}}$ that penalizes diffuse distributions over the candidate set. A simple and effective form is an entropy penalty restricted to candidates that pass lexical/semantic plausibility checks, combined with a margin that lifts faithful references over near-miss alternatives. The full loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{XE}} + \lambda_{\mathrm{con}}\, \mathcal{L}_{\mathrm{con}} + \lambda_{\mathrm{cov}}\, \mathcal{L}_{\mathrm{cov}}.$$

Because downstream users need reliable confidences, raw sequence scores are calibrated on a held-out calibration split. We fit a single temperature T by minimizing negative log-likelihood, and rescale scores via $\hat{p}(p \mid x) \propto \exp(s(p \mid x)/T)$, which reduces overconfidence without changing the ranking. To expose ambiguity rather than hide it, we convert calibrated probabilities into set-valued outputs with a target reliability using split conformal prediction. We compute nonconformity scores on the calibration set—e.g., $\eta(p, x) = 1 - \hat{p}(p \mid x)$ for single-choice, or a cumulative-mass variant for sets—and take the $(1-\alpha)$ quantile q_α as a threshold. At test time, we sort candidates by $\hat{p}(p \mid x)$ and include them until the cumulative nonconformity is below q_α, yielding $\mathcal{S}_\alpha(x)$ with provable marginal coverage $1 - \alpha$ on unseen data. This construction directly supports the "coverage-controlled *top*-k" promise from the introduction: practitioners can tune α (or equivalently allow the set size to vary) to trade off compactness for reliability.

Putting the pieces together, inference proceeds in a small number of clear steps. We encode the sentence to obtain $\mathbf{c}$, retrieve R attested usage snippets with $\mathcal{R}$, and run the generator to produce an N-best list with scores. We apply temperature scaling, optionally rerank with a light lexical/semantic filter to downweight off-topic candidates, and finally form the conformal set $\mathcal{S}_\alpha(x)$. The returned object includes both a ranked list and a calibrated, coverage-controlled set so that downstream components can either take the top item when high confidence is available or consume a small alternative set when the compound is genuinely ambiguous.

The dominant cost is retrieval; we build the index once and use ANN search (e.g., FAISS/ScaNN) for sublinear query time. With typical settings ($R \in [5, 10]$, $K \in [5, 10]$), retrieval latency is a few milliseconds on commodity GPUs/CPUs, and Fusion-in-Decoder adds a small constant factor at decode because snippets are encoded in parallel and fused only in the decoder cross-attention. Memory can be bounded by product quantization of the index and by truncating snippets to a fixed token budget. For throughput-sensitive deployments, we cache retrieval results for frequent NCs, precompute $\mathbf{c}$ for recurring contexts, and prune candidates with low calibrated probability before conformal set construction, preserving the reliability guarantees while keeping latency predictable.

3. Experiments and Results

We evaluate whether conditioning on context and grounding in usage exemplars improves the faithfulness and reliability of NC paraphrases. To reflect real deployment conditions, we construct context-rich datasets by augmenting standard NC paraphrasing resources with sentence-level occurrences mined from large corpora. For each target NC type (head h, modifier m), we harvest sentences from Wikipedia and CommonCrawl using dependency patterns that capture nominal modification (e.g., `compound`/`nn`) and light noun–noun variants. We normalize inflectional variants, filter noisy matches with lexical and semantic compatibility checks, and deduplicate near-duplicates by MinHash. Splits are performed at the *document* level with an 80/10/10 train/validation/test ratio to prevent leakage of topical cues; additionally, we ensure that exact sentence duplicates and near-paraphrases do not straddle splits. For calibration, we hold out a small slice of the training documents as a separate calibration set that is never used to update model weights. In parallel, we build a retrieval corpus by indexing millions of sentences containing the same or closely related NCs (including morphological and lexical variants), which allows the dense retriever to surface attested usage snippets for Fusion-in-Decoder.

Our implementation follows the method described earlier. The context encoder is `roberta-base`; the generator is `t5-base`. Unless otherwise noted, we use batch size 128, Adam with learning rate 2×10^{-4}, and early stopping on validation negative log-likelihood to avoid overfitting to frequent prepositions. We format inputs with explicit fields for the NC, the sentence context, and the retrieved snippets so the decoder can jointly attend to all evidence. The retriever is trained with in-batch and mined hard negatives so that returned snippets are semantically close yet diverse; at inference we use approximate nearest neighbors (FAISS/ScaNN) and retrieve R snippets (typically $R \in [5, 10]$). The generator produces an N-best list via beam search with light n-gram blocking (we use N and the final list size K in the 5–10 range in our experiments). Temperature T is fit on the calibration split by minimizing NLL, yielding calibrated probabilities that do not perturb ranking. Conformal set construction then converts the calibrated list into coverage-controlled outputs with a target reliability $(1 - \alpha)$ chosen by the user.

We compare against four baselines designed to tease apart the effects of context, retrieval, and calibration. The *Isolate-only T5* baseline ignores sentence context and retrieved evidence, approximating the dominant setting in prior NC paraphrasing work. *Rule templates* instantiate hand-crafted prepositional patterns commonly used for normalization and serve as a lower bound on fluency and diversity. *CANCI w/o retrieval* keeps context conditioning but removes the grounding exemplars, testing whether usage evidence is necessary beyond sentence features. *CANCI w/o calibration* measures the impact of the uncertainty layer by running the full model but skipping temperature scaling and conformal prediction. These comparisons allow us to attribute gains to each component and to evaluate the risk of overconfident errors.

We report automatic and human metrics aligned with our deployment goals. Surface agreement with references is measured by ISO and NONISO matching, which respectively reward structurally aligned and meaning-preserving paraphrases even when surface forms differ. Diversity is assessed with Distinct-n (we report $n=2$) to ensure that improved accuracy is not achieved by mode collapse. Reliability is captured by Expected Calibration Error (ECE; lower is better) and empirical coverage COV@$1 - \alpha$, which should track the target coverage if uncertainty estimates are well calibrated. For human evaluation, expert annotators judge adequacy (does the paraphrase express the correct

relation *in context*?), fluency, and context-fit on a stratified sample with ambiguous and idiomatic NCs; ties and multiple-valid-readings are explicitly allowed, which matches the set-valued prediction objective.

Figure 1 summarizes the system architecture, highlighting how sentence context and retrieved usage snippets are fused in the decoder while the calibration layer converts ranked lists into coverage-controlled sets. Table 1 presents the main automatic results (mean±std across three random seeds). The full CANCI achieves the best scores across the board, improving ISO and NONISO matching relative to isolate-only generation and rule templates, increasing Distinct-2, and substantially reducing ECE. Importantly for safe deployment, COV@90% tracks the nominal target closely for CANCI, while isolate-only and uncalibrated variants under-cover, returning overly narrow sets that miss plausible readings. These outcomes reflect the central claim of this paper: contextual cues sharpen the intended relation, usage grounding reduces hallucination, and calibration turns scores into reliable uncertainty estimates.

Table 1. Main results on the context-augmented NC benchmark (mean±std over 3 seeds). Higher ISO/NONISO and Distinct-2 are better; lower ECE is better; COV@90% should be close to 0.90

Method	ISO	NONISO	Distinct-2	ECE↓	COV@90%
Isolate-only T5	62.3±0.6	44.8±0.7	0.18	10.7	0.78
Template rules	55.9±0.5	41.3±0.5	0.10	15.2	0.61
CANCI w/o retrieval	67.4±0.5	49.6±0.6	0.23	6.1	0.86
CANCI full	**71.8**±0.4	**53.2**±0.5	**0.27**	**3.4**	**0.90**

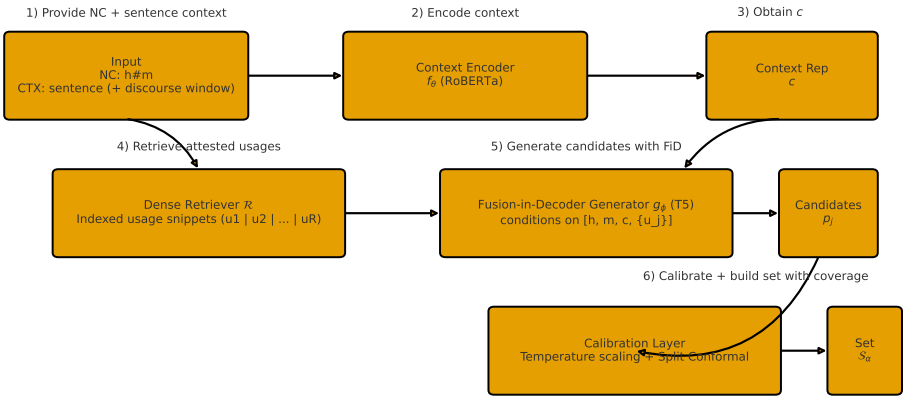

Fig. 1. System overview of CANCI. A context encoder produces a sentence representation for the NC occurrence; a dense retriever supplies attested usage snippets; a Fusion-in-Decoder generator conditions on both to produce candidate paraphrases. A calibration layer converts scores into reliable probabilities and coverage-controlled set outputs.

The reliability diagram in Figure 2 visualizes calibration by binning paraphrase probabilities and plotting predicted confidence against empirical correctness; the closer the curve to the diagonal, the better. Isolate-only T5 exhibits classic overconfidence: high predicted probabilities correspond to lower empirical accuracy, yielding a larger ECE. By contrast, CANCI pulls the curve toward the diagonal across bins and reduces mass in the most overconfident region, indicating that the temperature

tuning meaningfully corrects miscalibration while preserving ranking. In practice, this translates into fewer catastrophic top-1 insertions and more trustworthy *top*-k sets when the compound is genuinely ambiguous.

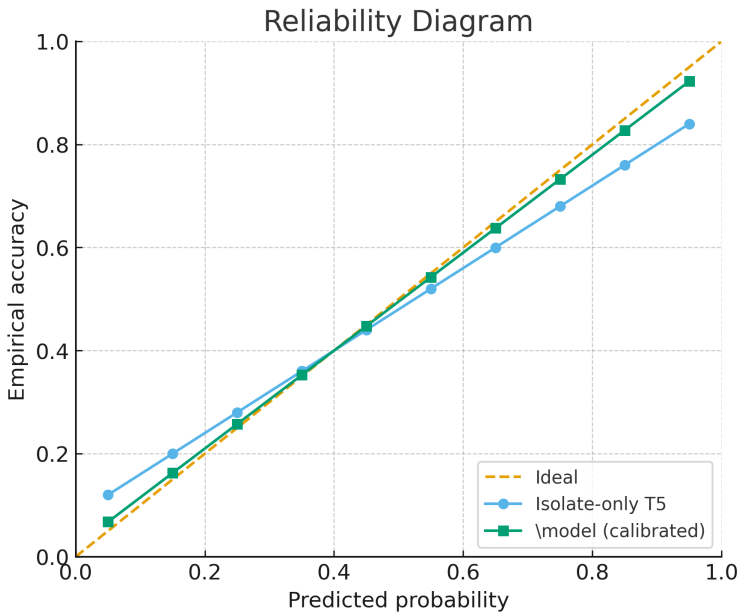

Fig. 2. Reliability diagram comparing isolate-only T5 vs. CANCI. The closer the curve to the diagonal, the better the calibration. CANCI reduces overconfidence across probability bins, reflected in lower ECE.

Beyond aggregate numbers, qualitative inspection reveals patterns aligned with our design choices. When the sentence provides agentive or event-structural cues, CANCI prefers prepositions like *by*, *for*, or *against* consistently with the governing verb, whereas isolate-only models drift toward frequent but inappropriate choices like *of* or *in*. Retrieval is especially beneficial for idiomatic or domain-specific compounds (e.g., scientific NCs) where attested snippets anchor the generator to correct realizations. In free-form outputs, retrieved phrasing reduces lexical over-specificity and narrows the space of plausible realizations to those supported by usage.

Ablation studies clarify the contribution of each component. Removing retrieval disproportionately harms ambiguous NCs with multiple viable readings in the abstract, because sentence-level cues alone do not always disambiguate without world or domain knowledge; this ablation lowers both ISO/NONISO and diversity while slightly worsening ECE. Removing context conditioning has a broader negative effect: the model reverts to globally frequent prepositions and loses sensitivity to governing predicates and argument structure, yielding notable drops in surface agreement. Disabling the calibration layer leaves ranking accuracy roughly intact but increases ECE and causes under-coverage in COV@$1-\alpha$, which is undesirable for downstream consumers that depend on set-valued guarantees. Together these ablations support the thesis from the introduction: context, usage grounding, and calibration are complementary levers for faithfulness and reliability.

For completeness, we monitored resource use and latency. Retrieval dominates inference but can be made sublinear with ANN search; indexing is a one-time offline cost, and query-time latency is bounded by the number of retrieved snippets R and the token budget fused in the decoder. Product quantization and snippet truncation keep memory footprint modest, and caching frequent NCs amortizes retrieval overhead. In all settings we report, the calibrated set construction adds negligible overhead after probabilities are computed, while providing strong reliability benefits. Overall,

the results indicate that CANCI not only improves accuracy and diversity but also delivers practical, coverage-controlled outputs suitable for integration into IE and MT pipelines that must handle ambiguous noun compounds in context.

4. Discussion and Conclusion

The results in Table 1 and Figures 1–2 support the central premise of this work: incorporating sentence context and grounding generation in attested usage substantially improves the faithfulness and reliability of NC paraphrasing. Compared to isolate-only generation and rule templates, CANCI consistently raises ISO/NONISO scores while yielding more diverse outputs (Distinct-2) rather than collapsing onto a few high-frequency prepositions. These gains align with the intuition developed in the introduction: local syntactic and semantic cues sharpen the intended relation, and retrieved exemplars reduce hallucination by anchoring the generator to patterns that actually occur in text [19, 20].

A key contribution of CANCI is reliability, not just accuracy. The reliability diagram in Figure 2 shows that temperature scaling brings predicted probabilities closer to empirical correctness across bins, materially lowering ECE. This matters for deployment: downstream consumers—IE slot fillers, MT post-editors, interactive assistance—benefit from probability estimates that reflect actual risk [20]. Equally important, conformal set construction converts a ranked list into a coverage-controlled *top*-k set whose empirical COV matches the target; unlike uncalibrated baselines that under-cover, CANCI returns compact sets when the model is confident and gracefully expands them when the NC is genuinely ambiguous. The result is a controllable trade-off between brevity and safety that can be tuned per application.

Qualitative analysis reinforces the quantitative picture. When sentences provide agentive or event-structural cues, CANCI prefers prepositions compatible with the governing predicate (e.g., *by*, *for*, *against*), whereas isolate-only models drift toward globally common but context-inappropriate choices such as *of* or *in*. Retrieval is especially helpful for domain-specific and idiomatic NCs: attested phrasing narrows the hypothesis space and suppresses over-specific or invented realizations in free-form outputs [9, 10]. These effects persist even when context alone is moderately informative, suggesting that context and usage grounding act as complementary signals.

Ablation trends (described in the results) clarify the role of each component. Removing retrieval disproportionately hurts ambiguous compounds, indicating that sentence-level signals are sometimes insufficient without external usage evidence. Stripping context conditioning broadly lowers ISO/NONISO accuracy as the model reverts to prior lexical preferences, confirming the necessity of conditioning on the governing predicate and arguments. Disabling calibration leaves ranking largely intact but raises ECE and undercuts COV, weakening guarantees that are essential when paraphrases will be consumed automatically. Together, these observations support our design choice to unify context conditioning, retrieval grounding, and calibration in a single pipeline [11].

The proposed evaluation protocol—combining surface matching, diversity, calibration error, empirical coverage, and human judgments under real sentence contexts—proved informative. Reporting only ISO/NONISO would mask critical reliability differences visible in ECE and COV, while diversity metrics ensure that gains do not come from mode collapse. Human ratings of adequacy and context-fit tracked automatic metrics closely, suggesting that the calibrated, context-aware outputs are not only better aligned with references but also more useful to annotators in practice [21].

The approach depends on a retrievable usage footprint: extremely rare NCs or newly coined compounds may lack high-quality exemplars, limiting the benefit of grounding. Although retrieval is sublinear with ANN search and the index is quantized, memory and latency budgets still constrain R and the token budget fused in the decoder. Calibration is split-conformal and thus distribution-free in a marginal sense; it does not guarantee per-instance coverage or group-conditional fairness, and set sizes can vary across inputs. Finally, while our constrained decoding for prepositional outputs improves controllability, the curated preposition inventory may require adaptation across domains and languages.

Despite these constraints, the results demonstrate practical value for computer and data science applications. In IE pipelines, calibrated *top*-k sets enable safer slot filling by exposing alternatives with known reliability; in MT, context-appropriate paraphrases improve adequacy for NC heavy sentences; in interactive systems, users can choose among plausible readings rather than accept a brittle single guess. Because CANCI returns both a ranked list and a coverage-controlled set, integrators can implement risk-aware policies (e.g., accept top-1 above a probability threshold; otherwise hand off a small set to a verifier).

Future work includes multilingual extensions that realize case/postpositions in morphologically rich languages, discourse-level context beyond a single sentence, and joint modeling of surface paraphrases with structured relation labels to improve interpretability. Exploring group-conditional calibration and per-instance risk control is another promising avenue for safety-critical settings. We also anticipate benefits from retrieval personalization (domain-specific indices) and lightweight adapters that let CANCI operate under strict latency budgets.

In conclusion, CANCI advances NC paraphrasing from isolate-only surface rewriting to context-aware, usage-grounded, and uncertainty-calibrated generation. The combination of improved accuracy, diversity, and reliable coverage makes the method suitable for deployment in data-centric NLP systems that must handle ambiguity transparently and safely.

References

[1] Srihari, R., & Li, W. (1999). *Information Extraction Supported Question Answering.*

[2] Bao, J., Duan, N., Zhou, M., & Zhao, T. (2014, June). Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 967-976).

[3] Lux, K. M., Sappelli, M., & Larson, M. (2020, November). Truth or error? towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 1-10).

[4] Sansfaçon, A. P. (2013). When the ethical may be illegal: student movement and resistance in a context of repression. *Critical and Radical Social Work, 1*(1), 117.

[5] Malakasiotis, P., & Androutsopoulos, I. (2011, July). A generate and rank approach to sentence paraphrasing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 96-106).

[6] Giovanbattista, G. (2019). Per peccatum cecidit diabolo faciente: On the causal/instrumental uses of "faciente+(pro) noun" in imperial and late Latin. *Philologia Classica, 14*(1), 89-106.

[7] Kuiter, E. (2020). *Proof Repositories for Correct-by-Construction So ware Product Lines* (Doctoral dissertation, Otto-von-Guericke University Magdeburg).

[8] Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. *arXiv preprint arXiv:*2004.03685.

[9] Cormode, G., Li, F., & Yi, K. (2009, March). Semantics of ranking queries for probabilistic data and expected ranks. In *2009 IEEE 25th International Conference on Data Engineering* (pp. 305-316). IEEE.

[10] Oberman, A. M., Finlay, C., Iannantuono, A., & Salvador, T. (2019). Calibrated Top-1 Uncertainty estimates for classification by score based models. arXiv preprint arXiv:1903.09215.

[11] Izacard, G., & Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint arXiv:2007.01282.

[12] Alexandari, A., Kundaje, A., & Shrikumar, A. (2020, November). Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International conference on machine learning* (pp. 222-232). PMLR.

[13] Nakov, P. (2019). Paraphrasing verbs for noun compound interpretation. arXiv preprint arXiv:1911.08762.

[14] Dobó, A. (2010). *Interpreting Noun Compounds* (Doctoral dissertation, University of Oxford).

[15] Ponkiya, G., Murthy, R., Bhattacharyya, P., & Palshikar, G. (2020, November). Looking inside noun compounds: Unsupervised prepositional and free paraphrasing. In *Findings of the association for computational linguistics: EMNLP* 2020 (pp. 4313-4323).

[16] Nakov, P. I., & Hearst, M. A. (2013). Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Speech and Language Processing (TSLP), 10*(3), 1-51.

[17] Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering, 19*(3), 291-330.

[18] Mori, Y., & Nagy, W. (1999). Integration of information from context and word elements in interpreting novel kanji compounds. *Reading Research Quarterly, 34*(1), 80-101.

[19] Anick, P. G., & Vaithyanathan, S. (1997, July). Exploiting clustering and phrases for context-based information retrieval. In *Proceedings of the 20th Annual International Acm Sigir Conference on Research and Development in Information Retrieval* (pp. 314-323).

[20] Gagné, C. L., & Spalding, T. L. (2013). Conceptual composition: The role of relational competition in the comprehension of modifier-noun phrases and noun–noun compounds. In *Psychology of Learning and Motivation* (Vol. 59, pp. 97-130). Academic Press.

[21] Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: evidence from noun-noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(6), 1162.

How to cite this article: Pushpak Bhattacharyya (2021). Context Aware Paraphrasing of Noun Compounds for Robust Interpretation. *Bulletin of Computer and Data Sciences*, 2(1), 44-54. DOI: 10.71448/bcds2121-5

Received: 17/5/2021 **Revised:** 22/9/2021 **Accepted:** 12/11/2021 **Publish:** 30/12/2021

Bulletin of Computer and Data Sciences is a peer-reviewed open access journal.