ASK BEFORE YOU SUMMARIZE: NLI-GUIDED UNCERTAINTY AND CLARIFICATION-AWARE ABSTRACTIVE SUMMARIZATION

Anders Vestrum
UC Berkeley
3041972833

Mihail Dimitrov
UC Berkeley
3037621433

Noah Lund Syrdal
UC Berkeley
3041928386

ABSTRACT

We study a clarification-aware abstractive summarization system that can either produce a summary immediately or ask exactly one targeted clarification question when the source document appears ambiguous. Our implementation uses disagreement across multiple sampled summaries as an uncertainty signal: we first generate diverse candidate summaries with an open summarization model, build an NLI-based semantic graph over those candidates, and compute a global uncertainty score to decide whether to commit or clarify. When uncertainty is high, we localize the least-supported summary sentence with sentence-level NLI scoring, generate a binary clarification question about that unstable claim, and optionally regenerate the summary conditioned on the resolved interpretation.

To support evaluation, we construct an ambiguity-sensitive summarization artifact from CNN/DailyMail with binary clarification questions, gold options, and source-grounded evidence. We evaluate the system with document entailment as a faithfulness proxy, along with threshold calibration and ablations on the gating and localization components. In our latest completed evaluation, the full pipeline improves over the greedy baseline on a 40-example test slice, with a mean entailment gain of $+0.0791$ and a paired permutation p-value of 0.0001. The full variant also outperforms both the entropy-gate and no-localization ablations. These results suggest that the current system is strongest when it combines improved output selection with selective clarification.

1 INTRODUCTION

Abstractive summarization systems can produce fluent outputs even when the source document itself admits multiple plausible readings. In these cases, the problem is not only hallucination in the usual sense, but premature commitment to one interpretation of an entity reference, time expression, numerical scope, or causal relation. Existing faithfulness metrics are useful for detecting unsupported claims after the fact, but they do not tell the system when to pause and ask for clarification before finalizing a summary.

Our project studies a simple interactive alternative: generate several candidate summaries, estimate whether their disagreement indicates genuine uncertainty, and ask exactly one binary clarification question only when that uncertainty is high. The emphasis is on selective interaction rather than open-ended dialogue. If the document appears unambiguous, the system should commit immediately; if the document appears ambiguous, the system should ask one targeted question and then produce a repaired summary.

Our key contribution is a selective clarification framework that combines NLI-based uncertainty estimation with targeted question generation for abstractive summarization. Empirically, we show that clarification improves faithfulness when applied selectively rather than uniformly. At the same time, our results reveal an important nuance: most of the current gain comes from improved output selection, with clarification providing an additional but smaller benefit on the hardest cases.

2 Related Work

Faithfulness remains a central challenge in abstractive summarization. Maynez et al. (2020) show that abstractive systems frequently introduce unsupported content, motivating evaluation methods that go beyond lexical overlap. Among these, NLI-based consistency scoring has proved especially useful: Laban et al. (2022) show that decomposing summaries into smaller units substantially improves inconsistency detection, which directly motivates our use of sentence-level NLI signals. As a complementary perspective, QAGS evaluates factuality via question answering over the source and summary (Wang et al., 2020), reinforcing the broader view that summary evaluation should focus on source support rather than surface similarity alone.

On the uncertainty side, Kuhn et al. (2023) propose semantic uncertainty as a way to estimate confidence over meanings rather than raw strings. Their core insight is that multiple generations can differ lexically while still expressing the same proposition, so uncertainty estimation should depend on semantic disagreement rather than token-level variation alone. Our implementation is motivated by the same principle, although we operationalize it with an NLI graph over sampled summaries rather than the exact semantic-entropy formulation.

Finally, prior work on clarification questions argues that a useful question is one whose answer materially improves the downstream response (Rao & Daumé III, 2018). More recent work shows that uncertainty can help determine when clarification is worthwhile (Testoni & Fernández, 2024; Zhang & Choi, 2025). These ideas motivate our selective policy: ask only when the sampled summaries disagree enough that a clarification step is likely to help.

3 Dataset

The artifact at `data/ambiguity_prompts.jsonl` contains 60 examples (20 development, 40 test) drawn from the CNN/DailyMail test split (Hermann et al., 2015; Nallapati et al., 2016). Each record includes the source article, reference summary, a binary clarification question, two answer options, the gold option, a gold answer, and a source-grounded evidence quote.

Construction is semi-automatic. For each article (minimum 120 words, seed 42), we prompt `gpt-4o-mini` (temperature 0) to generate a binary A/B clarification question targeting entity, temporal, numerical, or causal ambiguity. We then prompt the same model to resolve the question against the article and return a structured JSON response containing the gold option, a concise answer, and a supporting evidence span drawn from the source text. Examples where either the question fails strict binary-format parsing or the resolution returns "Unknown" are discarded. The resulting split consists of 20 development items used for threshold calibration and 40 test items used for all reported evaluations.

The current artifact is intentionally small and serves as a proof-of-concept evaluation testbed. Scaling to a substantially larger ambiguity-aware benchmark is a primary goal for the final project phase, as the current 40-example test slice limits the statistical power of downstream evaluations.

4 Method

Our implemented pipeline has four stages, illustrated below.

Stage 0: Multi-sample summary generation. Given a source document D, we sample N candidate summaries $\mathcal{S} = \{s_1, \ldots, s_N\}$ from the summarizer (parameters θ) via nucleus sampling with top-p truncation (Holtzman et al., 2020). Here $p_\theta(w_t \mid w_{<t}, D)$ is the model's probability for the t-th output token w_t given the preceding tokens $w_{<t}$ and document D. Each candidate is scored by its length-normalized log-probability,

$$\text{score}(s_i) = \frac{1}{|s_i|} \sum_{t=1}^{|s_i|} \log p_\theta(w_t \mid w_{<t}, D),$$

where $|s_i|$ is the number of tokens in s_i. Semantic deduplication then removes near-duplicate candidates: any candidate whose cosine similarity (computed with `all-MiniLM-L6-v2` on ℓ_2-

normalized embeddings) to any already-kept candidate exceeds 0.86 is discarded. This ensures that the retained pool $\mathcal{S}$ reflects genuine semantic diversity rather than surface variation.

Stage 1: Global uncertainty gate. We convert the raw scores into a proper probability distribution over candidates via softmax,

$$\tilde{p}_i = \frac{\exp(\text{score}(s_i))}{\sum_{k=1}^{N} \exp(\text{score}(s_k))},$$

so $\tilde{p}_i$ is the normalized generation weight of s_i, reflecting how likely the model considered that candidate relative to the others. We write $\text{ent}(a \to b) \in [0, 1]$ for the NLI entailment probability that hypothesis b is entailed by premise a, as assigned by the cross-encoder. The bidirectional entailment between two candidate summaries s_i and s_j is

$$w_{ij} = \tfrac{1}{2}\big[\text{ent}(s_i \to s_j) + \text{ent}(s_j \to s_i)\big] \in [0, 1],$$

which is high when the two summaries make the same claims and low when they diverge. Setting $\mu_{ij} = \tilde{p}_i \cdot \tilde{p}_j$ as the joint generation weight of the pair, the global uncertainty score is

$$U_{\text{global}} = -\sum_{i<j} \mu_{ij}\, w_{ij}.$$

U_{global} is bounded above by zero. It becomes increasingly negative as high-confidence candidates agree more strongly, and approaches zero as they diverge semantically. The system outputs directly when $U_{\text{global}} < \tau$, indicating sufficient consensus, and triggers clarification when $U_{\text{global}} \geq \tau$.

The threshold τ is selected by a grid search over $\tau \in [-0.45, 0.05]$ (step 0.02) on the 20 held-out development examples, maximizing expected EntScore subject to an ask-rate constraint of $[0.10, 0.40]$. If no threshold falls within the feasible band, the procedure falls back to the best non-zero-ask-rate point. Calibration yields $\tau = -0.19$ for the current pipeline.

Stage 2: Sentence-level localization. When clarification is triggered, we identify the riskiest summary claim rather than asking about the document as a whole. Let $\text{sent}(x)$ denote the set of sentences in a text x, and let $c_k^{(i)}$ be the k-th sentence of summary s_i. The source support of a summary sentence is defined as

$$\sup(c_k^{(i)}, D) = \max_{d \,\in\, \text{sent}(D)} \text{ent}(d \to c_k^{(i)}),$$

the strongest entailment from any source sentence d to the claim $c_k^{(i)}$. A value near 1 means the claim is well-grounded in the source; a value near 0 means it is unsupported or potentially hallucinated. The per-sentence risk, weighted by each candidate's generation probability, aggregates this across all sampled summaries:

$$\text{risk}(k) = \sum_i \tilde{p}_i \cdot \big(1 - \sup(c_k^{(i)}, D)\big).$$

The clarification hotspot is the sentence position $k^* = \arg\max_k \text{risk}(k)$ with the highest weighted lack of source support. This localization step ensures that the clarification question targets the specific claim that is most likely to be wrong across the high-confidence candidates, rather than a randomly chosen or superficially salient part of the summary.

Stage 3: Clarification and repair. From the hotspot sentence c_{k^*}, a binary A/B clarification question is generated by prompting `gpt-4o-mini` (temperature 0) with the hotspot sentence and a 2000-token prefix of D. The prompt enforces a strict three-line format (one question line, option A, option B) and a post-processing step guarantees format compliance. The question is then resolved against D by a second zero-temperature prompt that instructs the model to answer using only the source document, returning "Unknown" if the document is uninformative. A repaired summary is regenerated by the base summarizer conditioned on the resolved answer. Writing $\mathcal{S}' = \mathcal{S} \cup \{s_{\text{repair}}\}$ for the extended candidate pool, the final output is selected as

$$s^* = \arg\max_{s \in \mathcal{S}'} \frac{1}{|\text{sent}(s)|} \sum_{c \,\in\, \text{sent}(s)} \sup(c, D),$$

the candidate whose sentences are best supported by the source on average. Crucially, this selection rule applies whether or not clarification was triggered: even in direct-output cases the system picks the best-supported candidate from $\mathcal{S}$ rather than defaulting to the greedy decode.

5 BASELINES AND EVALUATION

5.1 BASELINE SETUP

All summarization variants use `facebook/bart-large-cnn` (Lewis et al., 2020) as the base summarizer to keep comparisons controlled across methods. The full pipeline samples $N = 8$ candidate summaries with temperature 1.0 and top-p 0.88. All NLI-based components, including the uncertainty graph, sentence-level support scoring, and the $\mathrm{EntScore}$ evaluation metric, use the cross-encoder `cross-encoder/nli-deberta-v3-large` (He et al., 2021).

5.2 BASELINES

We compare the full pipeline against three baselines:

1. **Greedy BART**: a single greedy decode from `facebook/bart-large-cnn` with no uncertainty gate and no clarification. This is the most direct comparison, isolating the effect of the entire pipeline.
2. **One-pass sampled**: a single nucleus-sampled decode (temperature 0.9, top-p 0.9) with no gate. This controls for whether sampling alone, without the multi-candidate selection and NLI scoring, accounts for any gains.
3. **One-question (no localization)**: a three-step `gpt-4o-mini` pipeline that always asks one clarification question, without NLI-based uncertainty gating or sentence-level localization. The model proposes a binary question, answers it from the document, and generates a summary conditioned on the answer, all at temperature 0.3. This baseline isolates the value of selective clarification over always asking.

5.3 EVALUATION METRICS

Our primary metric is document entailment. Using the same notation as in the method, we define

$$\mathrm{EntScore}(s, D) = \frac{1}{|\mathrm{sent}(s)|} \sum_{c \in \mathrm{sent}(s)} \max_{d \in \mathrm{sent}(D)} \mathrm{ent}(d \to c),$$

the average per-sentence source support. This is also the quantity maximized by the Stage 3 selection rule, making the evaluation directly aligned with the optimization objective. We additionally report ask rate, 95% bootstrap confidence intervals, and paired permutation p-values for statistical significance.

The ablations isolate the contribution of each pipeline component:

1. **Full**: graph-based uncertainty gate + sentence-level localization + clarification and repair.
2. **Entropy gate**: replace the NLI graph gate with an entropy-of-weights gate $H = -\sum_i \tilde{p}_i \log \tilde{p}_i$ over the generation distribution. This tests whether the graph-based uncertainty signal is necessary or whether a simpler diversity measure suffices.
3. **No localization**: keep the graph gate but replace targeted k^* localization with the one-question baseline. This tests whether sentence-level localization adds value beyond gated clarification alone.

6 PRELIMINARY RESULTS

Calibration on the 20 development examples yields $\tau = -0.19$. On the 40-example test slice, the full pipeline triggers clarification on 6 examples (ask rate 0.15) and outputs directly on the remaining 34.

The overall mean $\mathrm{EntScore}$ for the full pipeline is 0.4888, compared to 0.4097 for the greedy BART baseline, a mean delta of $+0.0791$. The 95% bootstrap confidence interval for the delta is $[0.0384, 0.1266]$ and the paired permutation p-value is 0.0001, indicating a statistically robust improvement.

The per-decision breakdown sheds light on the source of the gain. On the 6 examples where the system asks a clarification question, the mean delta relative to the greedy baseline is $+0.0510$. On the 34 examples where it outputs directly, the mean delta is $+0.0841$. This reveals that the stronger output-selection policy, picking the best-supported candidate from the sampled pool, accounts for the larger share of the improvement, while targeted clarification provides an additional but smaller benefit on the hardest cases.

Method	Mean entailment	Ask rate	Mean delta vs. baseline
Greedy BART baseline	0.4097	0.000	0.0000
Full pipeline	0.4888	0.150	+0.0791
Entropy-gate ablation	0.4832	0.000	+0.0735
No-localization ablation	0.4531	0.125	+0.0434

Table 1: Results on the 40-example test slice. The full pipeline is best overall. The entropy-gate ablation, which never asks a clarification question, narrows the gap considerably, showing that improved output selection accounts for a large share of the gain. The no-localization ablation trails further, confirming that targeted sentence-level localization matters when clarification is used.

The full pipeline outperforms both ablations (Table 1). The entropy-gate ablation achieves 0.4832 without ever asking, demonstrating that the NLI-based graph gate and best-candidate selection together are already a strong direct-output policy. The no-localization ablation (0.4531) falls substantially further behind, showing that indiscriminate clarification without localization is less effective than either the full system or the entropy-gate variant. Together, these results suggest that the two components play complementary roles: the selection policy drives the bulk of the gain, while targeted localization makes clarification worthwhile when it is triggered.

7 Conclusion

We implemented a clarification-aware summarization pipeline that uses multi-sample disagreement, NLI-based uncertainty estimation, sentence-level localization, and one binary clarification question. In the latest evaluation, the system improves over the greedy baseline and outperforms both ablations. The clearest takeaway is that clarification helps, but most of the present gain comes from better output selection; selective clarification is therefore useful, but currently secondary. Future work should focus on making clarification more valuable, expanding the ambiguity dataset beyond the current small artifact, and replacing the position-based alignment in Stage 2 with semantic clustering of sentences across candidates, removing the assumption that sentence position is a reliable proxy for semantic role.

References

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023. doi: 10.48550/arXiv.2302.09664. URL `https://arxiv.org/abs/2302.09664`.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL `https://aclanthology.org/2022.tacl-1.10/`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880. Association for Computational Linguistics, 2020.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290. Association for Computational Linguistics, 2016.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2737–2746, 2018. doi: 10.18653/v1/P18-1255. URL https://aclanthology.org/P18-1255/.

Alberto Testoni and Raquel Fernández. Asking the right question at the right time: Human and model uncertainty guidance to ask clarification questions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 258–275, 2024. doi: 10.18653/v1/2024.eacl-long.16. URL https://aclanthology.org/2024.eacl-long.16/.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, 2020. doi: 10.18653/v1/2020.acl-main.450. URL https://aclanthology.org/2020.acl-main.450/.

Michael JQ Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5541–5558, 2025. doi: 10.18653/v1/2025.findings-naacl.306. URL https://aclanthology.org/2025.findings-naacl.306/.