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

Anders Vestrum
3041972833

Mihail Dimitrov
3037621433

Noah Lund Syrdal
3041928386

ABSTRACT

We study a clarification-aware abstractive summarization system that can either produce a summary immediately or enter a one-question clarification branch when sampled summaries suggest ambiguity. 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 **AmbigSum**, an ambiguity-sensitive summarization artifact derived from CNN/DailyMail, with binary clarification questions, gold options, and source-grounded evidence. AmbigSum contains 500 examples split into 150 development items and 350 test items. We evaluate the system with document entailment as a faithfulness proxy, along with threshold calibration and ablations on the gating and localization components. On the 350-example test split, the full pipeline achieves a mean document-entailment score of 0.5670 with an ask rate of 0.4171, improving over a greedy BART baseline (0.4982, $\Delta = +0.0688$, 95% CI $[0.0551, 0.0833]$, $p < 0.001$) and roughly comparable to the entropy-gate ablation (0.5640) while improving over the no-localization ablation (0.5247). Decomposing the gain by gate decision, examples flagged for clarification show a mean delta of $+0.0959$ over baseline ($n = 146$, 95% CI $[0.0714, 0.1225]$), versus $+0.0495$ on direct-output examples ($n = 204$, 95% CI $[0.0352, 0.0652]$). These results suggest that targeted clarification provides larger per-example gains where the gate triggers it, while output selection contributes substantially across all examples.

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 selective alternative: generate several candidate summaries, estimate whether their disagreement indicates genuine uncertainty, and enter a one-question clarification branch only when that uncertainty is high. The emphasis is on selective clarification 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 binary question and then produce a repaired summary conditioned on the resolved interpretation. Figure 2 sketches the resulting two-branch policy, and Figure 3 shows a worked unrepaired-vs-repaired example from our artifact.

Our key contribution is a selective clarification framework that combines NLI-based uncertainty estimation with targeted question generation for abstractive summarization. The project makes three concrete contributions: an end-to-end implementation of a summarize-now versus ask-once policy; AmbigSum, a CNN/DailyMail evaluation artifact with binary clarification questions and source-grounded evidence; and an ablation study separating semantic-graph gating, entropy-based gating, and targeted localization. Empirically, the full graph-gated system performs best among the implemented variants, while also showing that output selection is already a strong part of the pipeline. This makes the result useful even where clarification is not yet dominant: it identifies which parts of the system are working and where future work should focus.

2 Related Work

Faithfulness remains a central challenge in abstractive summarization. Maynez et al. (2020) show that abstractive systems frequently introduce unsupported content even when surface-level metrics like ROUGE remain high; we adopt their framing that intrinsic and extrinsic hallucinations are the failure mode worth measuring, but we depart from their human-centric protocol by automating the signal at sentence level. Laban et al. (2022) make the case that document-summary entailment becomes reliable only after granular decomposition, and we borrow their per-sentence aggregation directly when computing $\sup(c, D)$; we do not, however, adopt their convolutional aggregator, choosing instead a simpler max-over-source aggregator to keep the localization signal interpretable per claim. FENICE (Scirè et al., 2024) extends this with explicit claim extraction and per-claim verification, and our hotspot-localization stage shares the same claim-level granularity, but we operate over the model's own summary sentences rather than running a separate claim extractor, since our goal is to localize a question target rather than produce a final factuality judgment. QAGS (Wang et al., 2020) pursues the same source-grounding principle through question answering rather than NLI; we keep its emphasis on source support but not its QA pipeline, because our clarification stage already requires a question-generation budget and we prefer to avoid two QG stacks.

On the uncertainty side, Kuhn et al. (2023) argue that confidence should be estimated over meanings rather than strings, using bidirectional NLI to cluster paraphrastic samples before computing entropy. We borrow the bidirectional-entailment idea verbatim in defining w_{ij}, but we deliberately avoid the explicit clustering step: our pipeline weights pairs by μ_{ij} instead of collapsing them into discrete meaning clusters, which keeps the gate differentiable in τ and avoids brittle cluster-boundary decisions on small candidate pools. Aichberger et al. (2024) push semantic-diverse decoding further by steering generations toward semantically distinct outputs, a direction we leave to future work since BART nucleus sampling already produces enough disagreement at $N{=}8$ to populate the graph. Nguyen et al. (2025) show that pairwise semantic similarity is a competitive uncertainty estimator without an entropy formulation, which is the closest comparable to our U_{global} score; the difference is that we use the NLI cross-encoder directly rather than a learned similarity, trading some flexibility for a single-model design. Chen et al. (2025) construct a semantic graph over generations for hallucination detection, which is methodologically close to our setup; the difference is that they target post-hoc detection on QA outputs while we use the graph as a gating signal for whether to ask a clarification question, so the same structural primitive serves a different decision.

Prior work on clarification questions argues that a useful question is one whose answer materially improves the downstream response (Rao & Daumé III, 2018); we inherit this utility-driven framing but instantiate it with an offline gold answer rather than a live user, which is a deliberate restriction so the evaluation can isolate the gating and localization decisions from question-quality and user-modeling effects. Testoni & Fernández (2024) show that uncertainty signals correlate with when clarification is worthwhile in dialog, which is the design choice we make at the gate. Chen et al. (2024) and Zhang & Choi (2025) explore richer LLM-driven clarification strategies including multi-turn ambiguity resolution; we deliberately restrict ourselves to a single binary question because our goal is to attribute gains to gating and localization specifically, not to optimize the clarification interface itself.

Our pipeline can also be situated within the broader neuro-symbolic tradition of combining neural generators with structured verifiers. Where prior work has used explicit logical forms or external solvers as the symbolic layer (Pan et al., 2023; Lyu et al., 2023), we instantiate the symbolic layer as an NLI-induced semantic graph: edges encode bidirectional entailment w_{ij}, and the gate decision

DATA EXAMPLE

ARTICLE

"Elizabeth Warren says she is relieved Dzhokhar Tsarnaev was found guilty, but does not believe he should be executed."

QUESTION

Did Elizabeth Warren support the death penalty for Dzhokhar Tsarnaev?

REFERENCE CONTEXT

Tsarnaev was found guilty on all 30 charges. The case moved to the penalty phase. Warren opposed the death penalty.

OPTIONS

A) Yes, she believed he should be executed.
B) No, she opposed the death penalty.

GOLD ANSWER

Answer: B
No, she opposed the death penalty.

SOURCE EVIDENCE

"I don't support the death penalty... he should spend his life in jail... he should die in prison."

Figure 1: Example item from the ambiguity-sensitive CNN/DailyMail artifact (`ambig_0004`). Each item pairs a news article with one binary clarification question, two plausible options, a gold answer, and source-grounded evidence.

$U_{\text{global}} \geq \tau$ is a discrete rule over that graph. Sentence-level localization via $\arg\max_{i,j} \text{risk}(i, j)$ and the penalized-entailment keep rule are similarly structured decisions over neural scores. This places our system in the same design family as verifier-guided generation (Cobbe et al., 2021; Madaan et al., 2023), but specialized to a clarification-vs-commit policy rather than post-hoc answer correction.

3 Dataset Artifact

3.1 Dataset

AmbigSum is the evaluation artifact we built for this study. It contains 500 examples split into 150 development items and 350 test items. Each record includes the source article, reference summary, a binary clarification question, two answer options, the gold option, a gold answer, and source-grounded evidence. Figure 1 shows a representative item from the artifact, illustrating how the binary question and gold evidence anchor the offline clarification protocol. In this project, clarification is evaluated in an offline setting: the answer to the clarification question is resolved using the artifact's gold option and source-grounded evidence, rather than through a live user. This makes the setup controlled and repeatable, but it also means that the results should be interpreted as clarification-aware summarization under an artifact-based protocol rather than as a deployed interactive system.

3.2 Construction Procedure

The artifact is built from CNN/DailyMail rather than from synthetic toy passages so that the source documents preserve the structure and length distribution of realistic news summarization inputs (Hermann et al., 2015; Nallapati et al., 2016). We shuffle the available articles, skip very short documents, and generate one binary clarification question per retained article. The question format is deliberately restrictive: each item must contain one question, two answer options, a gold option, a resolved answer, and an evidence span from the source. This makes the artifact compatible with the project's ask-once setting, where the system is not allowed to carry out an open-ended conversation.

The resulting examples are not meant to be a comprehensive ambiguity benchmark. Instead, they serve as a controlled evaluation artifact for the specific behavior studied here: deciding whether un-

certainty warrants one additional clarification step before finalizing a summary. The source-evidence field is important because it keeps the generated clarification answer tied to the document rather than to free-form model preference. The development split is used only for threshold calibration, and the 350-example test split is held out for the reported full and ablation comparisons.

4 METHOD

Our implemented pipeline has four stages, summarized in Figure 2.

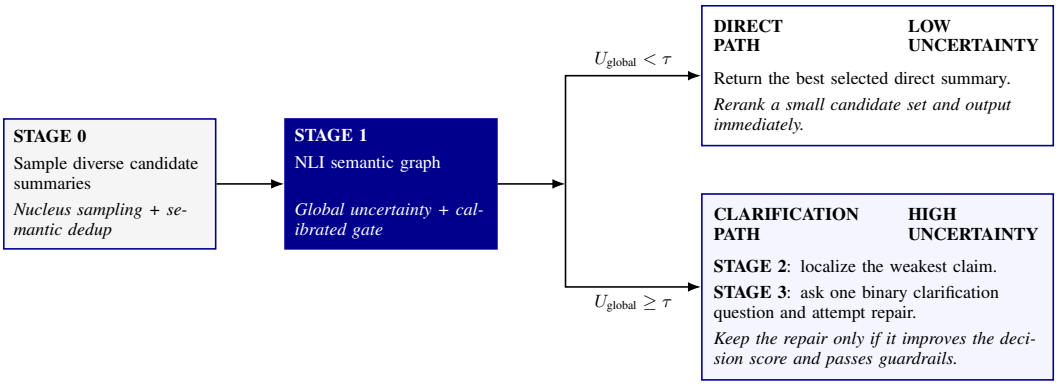

Figure 2: Selective clarification pipeline. Stage 0 samples diverse candidate summaries. Stage 1 builds an NLI-based semantic graph and applies a calibrated uncertainty gate τ. When $U_{\text{global}} < \tau$, the system returns the best direct summary immediately. When $U_{\text{global}} \geq \tau$, it enters the clarification path: localize the weakest claim, ask one binary clarification question, attempt repair, and keep the repaired summary only if it improves the decision score and satisfies guardrails.

Stage 0: Multi-sample summary generation. Given a source document D, we sample N candidate summaries $\mathcal{S} = \{s_1, \ldots, s_N\}$ from `facebook/bart-large-cnn` (Lewis et al., 2020) 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. We apply semantic deduplication to remove candidates that are too close under embedding similarity. The purpose is not simply to maximize diversity, but to avoid treating trivial paraphrases as independent evidence of uncertainty. If sampling produces too few distinct summaries, the implementation relaxes the deduplication constraint enough to retain a minimally useful candidate pool. This multi-sample construction is motivated by semantic-uncertainty work, which argues that confidence should be estimated over meanings rather than individual strings (Kuhn et al., 2023). It is also consistent with more recent work that uses semantically diverse generations as a practical uncertainty signal (Aichberger et al., 2024).

Stage 1: Global uncertainty gate. We convert the raw scores into a normalized weighting over candidates via a softmax,

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

where $\tilde{p}_i$ is treated as a relative confidence weight rather than a proper posterior over candidates: length-normalized log-probabilities are not summable into a true distribution, so we use $\tilde{p}_i$ only as a heuristic weight that emphasizes higher-likelihood candidates when comparing pairs. 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. We use NLI-based scoring because it provides a semantic notion of agreement between summaries, allowing the system to compare meaning rather than surface overlap. The bidirectional entailment between two candidate summaries s_i and s_j is then

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

Setting $\mu_{ij} = \tilde{p}_i \cdot \tilde{p}_j$ as the joint weight of the pair (so pairs of high-likelihood candidates contribute more), we define the global uncertainty score as negative weighted semantic agreement:

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

In practice, U_{global} captures agreement in meaning space rather than surface form. When candidate summaries express similar propositions, pairwise entailment scores are high, leading to a more negative low-uncertainty value. Conversely, when candidates diverge semantically, entailment scores drop and U_{global} approaches zero, signaling ambiguity. The system triggers clarification when $U_{\text{global}} \geq \tau$ and outputs directly otherwise, where the threshold τ is selected by calibration on held-out development examples. This follows the same broad motivation as semantic uncertainty (Kuhn et al., 2023), but instantiates it with pairwise NLI agreement. It is also close in spirit to newer uncertainty estimators that explicitly exploit pairwise semantic similarity (Nguyen et al., 2025).

Stage 2: Sentence-level localization. When clarification is triggered, we identify the riskiest summary claim. Let $\mathrm{sent}(x)$ denote the set of sentences in a text x, and let $c_j^{(i)}$ be the j-th sentence of summary s_i. The source support of a single summary sentence is

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

i.e. the strongest entailment from any source sentence $d \in \mathrm{sent}(D)$ to the claim $c_j^{(i)}$. This sentence-level decomposition is directly motivated by NLI-based factuality work such as SummaC (Laban et al., 2022) and more recent claim-level factuality evaluation methods such as FENICE (Scirè et al., 2024). A value near 1 means the claim is well-supported; a value near 0 means it is not. The implementation scores each retained candidate sentence independently, ignoring very short fragments. The per-sentence risk, weighted by generation probability, is

$$\mathrm{risk}(i, j) = \tilde{p}_i \cdot \big(1 - \sup(c_j^{(i)}, D)\big),$$

and the clarification hotspot is the candidate sentence $(i^*, j^*) = \arg\max_{i,j} \mathrm{risk}(i, j)$ with the highest weighted lack of support. This treats each summary sentence independently and selects the globally highest-risk claim across all sampled summaries.

Stage 3: Clarification and repair. From the hotspot sentence $c_{j^*}^{(i^*)}$, the system generates one binary clarification question, resolves the question using the artifact's gold option and source-grounded evidence, and optionally regenerates one or more repaired summaries conditioned on the resolved fact. This selective ask-once design is aligned with prior work on clarification utility and uncertainty-aware questioning (Rao & Daumé III, 2018; Testoni & Fernández, 2024; Chen et al., 2024; Zhang & Choi, 2025). Before question generation, however, a clarification policy gate may still skip asking if hotspot support is at least σ_{ask} or if the number of retained candidates is below $N_{\min}^{\text{clarify}}$. Writing $\mathcal{S}' = \mathcal{S} \cup \{s_{\text{repair}}\}$ for the extended candidate pool, the idealized final output is

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

i.e. the candidate whose sentences are best supported by the source on average. The implementation approximates this idealized rule with a pragmatic selector: it compares a greedy BART summary, a one-pass sampled BART summary, the highest-scoring sampled Stage 0 candidate, and any repaired candidate that is generated. Direct-output selection is therefore already a reranking problem, not a single-decoder baseline. A repaired summary is accepted only under a conservative keep rule based on penalized entailment and guardrails; Figure 3 shows a concrete unrepaired-vs-repaired comparison on a document where a primary cause is ambiguous between two readings. For any summary s, let $p_e(s)$ be the mean sentence-level document-entailment probability and $p_c(s)$ the maximum sentence-level document-contradiction probability, both computed by the NLI cross-encoder over summary sentences against the source document. We score summaries as $\mathrm{pen}(s) = p_e(s) - \lambda p_c(s)$ and keep the repair only if $\Delta\mathrm{pen} \geq \delta_{\text{pen}}$, semantic preservation exceeds η_{pres}, and the repaired-to-original length ratio lies in $[\rho_{\min}, \rho_{\max}]$; otherwise, we keep the direct output.

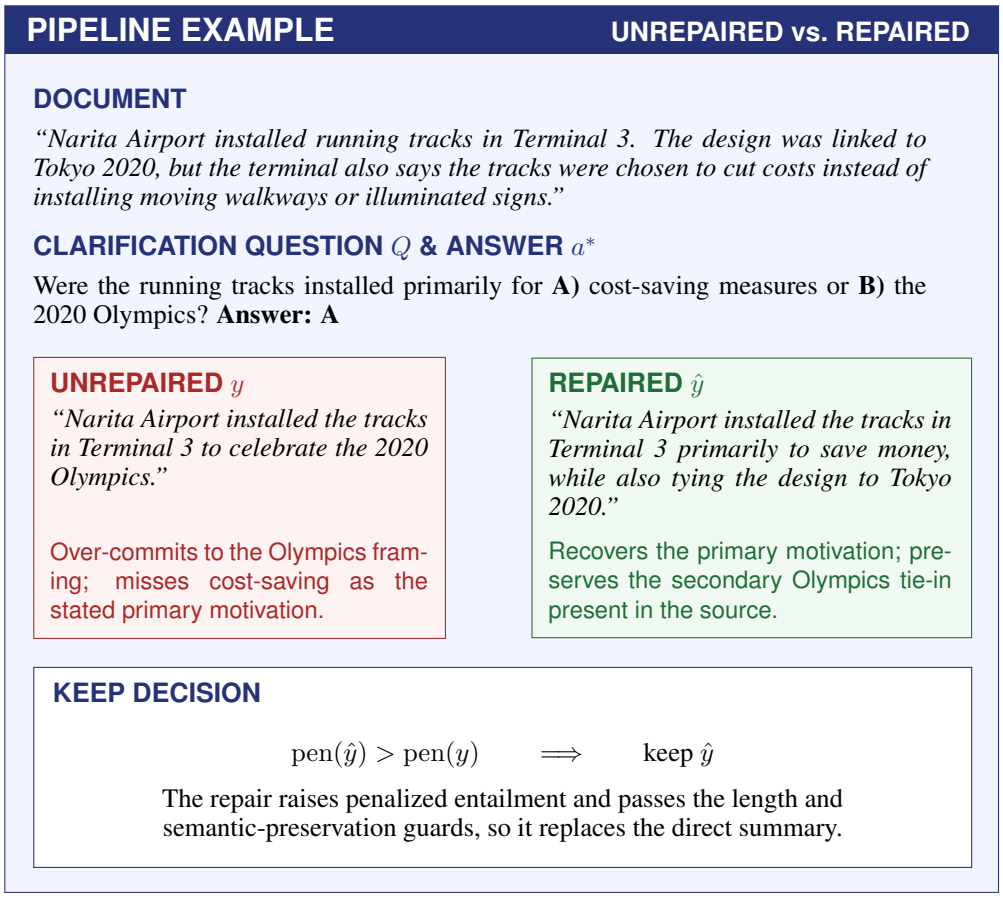

Figure 3: Side-by-side pipeline walkthrough on dataset item `ambig_0005`. The unrepaired summary y over-commits to one reading of an ambiguous document. After the binary clarification question Q is resolved with answer a^*, the repaired summary $\hat{y}$ recovers the primary motivation while preserving the secondary detail. The keep rule accepts $\hat{y}$ only when penalized entailment improves and the guardrails pass.

5 Baselines, Ablations, and Evaluation

5.1 Experimental Setup

All variants use the same base summarizer, `facebook/bart-large-cnn`, to keep comparisons controlled. We use a task-specific summarization model (BART) rather than a general instruction-tuned LLM for four reasons. First, nucleus sampling from BART at $N = 8$ is cheap and produces stable disagreement estimates across runs, whereas equivalent semantic diversity from a frontier LLM is more expensive per token and more variable across sampling temperatures. Second, BART-large-cnn is trained on CNN/DailyMail, so output length and register match the reference distribution; LLM outputs vary in length and style across samples, which would inflate apparent disagreement on the NLI graph and pollute U_{global} with stylistic rather than semantic divergence. Third, an open-weights BART fixes the generation distribution and avoids API versioning drift across the project timeline, which matters for reproducibility of calibration thresholds. Fourth, the contribution of this work is the gating and localization machinery rather than the generator itself, so holding the generator fixed and modest isolates where gains come from. The gating and localization stages are generator-agnostic, and substituting an instruction-tuned LLM in Stage 0 is a natural extension that we leave to future work.

As a single-decoder reference point, we additionally report a greedy BART baseline: a single greedy decode from `facebook/bart-large-cnn` with no sampling, no uncertainty gate, and no clar-

ification. This isolates the contribution of the full pipeline's multi-sample generation, output reranking, and clarification relative to a vanilla one-pass summarizer.

Hyperparameters were selected on the development split and fixed before test evaluation. The pipeline uses nucleus sampling with $N = 8$, temperature 1.0, and top-$p = 0.88$. We use $N = 8$ as a practical cost-quality tradeoff that provides stable disagreement estimates without excessive runtime. For uncertainty estimation and sentence-level factuality scoring, all NLI-based components use the same cross-encoder, `cross-encoder/nli-deberta-v3-large`, to preserve a controlled comparison across variants. All keep-rule thresholds were also fixed before test evaluation. These include the contradiction penalty weight λ, the minimum penalized-score gain δ_{pen}, the minimum semantic-preservation threshold η_{pres}, the hotspot-support cutoff for asking σ_{ask}, and the repaired-to-original length-ratio bounds $[\rho_{\min}, \rho_{\max}]$.

5.2 Evaluation Metrics

Our primary metric is document entailment. Using the same notation as in the method: $\text{sent}(\cdot)$ for sentences and $\text{ent}(d \to c)$ for NLI entailment probability, we define

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

the average per-sentence source support. This score is the main evaluation metric, while Stage 3 decisions use the stricter penalized-entailment keep rule described above. We additionally report ask rate, calibration outputs, and ablations.

5.3 Ablations

We compare against the following variants:

1. **Greedy BART baseline**: a single greedy decode with no sampling, gating, or clarification.
2. **Full**: graph gate + sentence-level localization + clarification/repair.
3. **Entropy gate**: replace U_{global} with an entropy-of-weights gate $H = -\sum_i \tilde{p}_i \log \tilde{p}_i$.
4. **No localization**: keep the gate but replace targeted (i^*, j^*) localization with a simpler one-question branch.

5.4 Calibration and Protocol

The decision threshold τ is calibrated on the 150-example development split. For each development item, the pipeline records the graph uncertainty score and estimates what would happen under the direct-output and clarification branches. The calibration sweep then chooses the threshold that maximizes expected document entailment among thresholds whose dev-split ask rate falls within a target band. The band is a calibration-time constraint on threshold selection, not a runtime cap on the test split: realized test-split ask rates may differ from the dev-split ask rate at the chosen τ. This matters because a trivial policy could ask too often, while another trivial policy could never ask; the project goal is selective interaction, so the threshold must trade off faithfulness gains against the cost of asking.

All reported test results are computed on the 350-example held-out split after calibration. The primary score is the NLI-based document-entailment proxy used throughout our pipeline. This metric is imperfect, but it gives a consistent automatic signal for comparing many variants without human annotation. We therefore interpret the absolute numbers cautiously and focus primarily on controlled comparisons between variants evaluated with the same scorer.

6 Results

The main evaluation uses the 350-example test split. Calibration on the 150-example development split selects $\tau = -0.27$.

On the 350-example test split (Table 1), the full pipeline asks on 146 examples and outputs directly on 204, for an ask rate of 0.4171. The overall mean final $\mathrm{EntScore}$ is 0.5670. Compared to the greedy BART baseline (0.4982), the full pipeline improves by $+0.0688$ (paired permutation $p < 0.001$, 95% CI $[0.0551, 0.0833]$ over 5000 bootstrap resamples). The entropy-gate ablation reaches 0.5640 with an ask rate of 0.1057, and the no-localization ablation reaches 0.5247 with the same 0.4171 ask rate as the full pipeline.

The full pipeline therefore improves by $+0.0030$ over the entropy-gate ablation and by $+0.0423$ over the no-localization ablation. The no-localization gap is especially informative because that variant asks at exactly the same rate as the full system, isolating the contribution of targeted hotspot localization. Decomposing the full-vs-baseline delta by gate decision yields a mean delta of $+0.0959$ on the 146 examples flagged for clarification (95% CI $[0.0714, 0.1225]$) versus $+0.0495$ on the 204 direct-output examples (95% CI $[0.0352, 0.0652]$), suggesting that gains concentrate where the gate identifies uncertainty even though direct-output reranking contributes meaningfully throughout.

Method	Mean Entailment	Ask Rate	Difference vs. Full
Greedy baseline	0.4982	0.0000	-0.0688
Full pipeline	**0.5670**	0.4171	–
Entropy-gate ablation	0.5640	0.1057	-0.0030
No-localization ablation	0.5247	0.4171	-0.0423

Table 1: Results on the 350-example test split. The full pipeline improves over the greedy BART baseline by $+0.0688$ EntScore (paired permutation $p < 0.001$, 95% CI $[0.0551, 0.0833]$). The entropy-gate ablation is essentially tied with the full pipeline despite a much lower ask rate, suggesting that the choice of gate signal is less important than whether interaction is used at all. The no-localization ablation, which asks at the same rate as the full pipeline but replaces targeted hotspot selection with a simpler one-question branch, trails by -0.0423, isolating the contribution of where the system asks.

6.1 Interpreting the Ablations

The two ablations isolate different effects. The entropy-gate variant retains the same candidate generation and output selection but replaces U_{global} with a simpler entropy over generation weights and asks at a much lower rate (0.1057 vs. 0.4171); the fact that it still reaches 0.5640 shows that uncertainty-aware sampling and reranking already produce most of the gain even when clarification fires sparingly. The no-localization variant asks at exactly the same rate as the full pipeline (0.4171) but scores 0.0423 lower, which isolates the contribution of targeted hotspot selection: *where* the system asks matters, not just *whether* it asks.

6.2 Error Analysis

Inspecting per-example outputs surfaces two recurring failure modes that account for a meaningful share of the residual EntScore gap.

Failure Mode 1: Confident consensus on a poorly supported summary. When BART's samples collapse onto near-paraphrases of the same hallucinated claim (an over-specific entity, date, or causal link not in the source), bidirectional NLI agreement is high, U_{global} falls well below τ, and the pipeline commits to direct output without ever entering the clarification branch. The resulting summary then scores poorly under document entailment despite an internally consistent candidate pool. This is a structural weakness of any agreement-based gate, and the fix is not more sampling but an orthogonal source-support floor that overrides the gate when no candidate clears a minimum entailment threshold.

Failure Mode 2: Repairs are rarely accepted by the keep rule. Even when the hotspot anchors a concrete claim, the conservative penalized-entailment keep rule rejects most candidate repairs. In the full evaluation, only 2.6% of test examples (6.2% of clarification attempts) result in an accepted repair, with the remaining clarifications falling back to the best direct-output candidate. The mean

penalized-entailment delta of repaired-vs-direct candidates is -0.0774, indicating that most repairs hurt the keep-rule score even when the underlying summary may be more informative. This explains the small full-vs-entropy-gate margin: when repairs are rarely kept, the clarification branch reduces to a slightly different direct-output reranking, similar to what the entropy-gate ablation already does.

Aggregate gains are also structurally bounded for reasons orthogonal to these failure modes: EntScore averages support over all summary sentences so a single repaired claim is diluted by already-supported ones; the multi-stage chain (gate $\rightarrow$ localize $\rightarrow$ ask $\rightarrow$ resolve $\rightarrow$ repair $\rightarrow$ accept) compounds losses at each step; and the keep rule favors precision over aggressiveness, rejecting partially helpful repairs. Two concrete next steps follow: the gate should combine semantic agreement with a source-support floor to catch confident hallucinations, and the keep rule should be tightened so that more useful repairs survive acceptance rather than being filtered as length or preservation violations.

7 LIMITATIONS

EntScore is an automatic NLI proxy, not a human factuality judgment, and the same NLI family is used internally for localization, selection, and final evaluation; reported gains should therefore be read as internal automatic-evaluation gains rather than definitive human factuality improvements, and a stronger final evaluation would add human annotation or an independent factuality evaluator. The clarification interface is restricted to a single binary question, which keeps evaluation clean but cannot resolve ambiguities involving multiple entities, uncertain temporal ordering, or missing context. The artifact is semi-automatic and tied to CNN/DailyMail news, so behavior on legal, medical, or long-form scientific text, where evidence is less linear and more dispersed, is unverified. Finally, the pipeline is computationally expensive (multi-sample generation, NLI graph construction, sentence-level localization, calibration, and multiple ablation passes), so deployment at scale would require caching, batched inference, or cheaper uncertainty approximations.

The ask-rate calibration band $[0.20, 0.35]$ was enforced when selecting thresholds on the development split, where the chosen graph threshold $\tau = -0.27$ produced a dev ask rate of 0.30. On the held-out test split, however, realized ask rates differed: the full pipeline and no-localization ablation (which share the graph threshold) reached 0.4171, and the entropy-gate ablation reached 0.1057. This dev-to-test ask-rate gap reflects imperfect threshold transfer under distribution shift on a relatively small dev split, and a larger development set would likely yield more stable calibration.

8 CONCLUSION

We implemented a clarification-aware summarization pipeline that combines multi-sample disagreement, NLI-based uncertainty estimation, sentence-level localization, and one binary clarification question under an offline artifact-based protocol. On the 350-example test split, the full system improves over a greedy BART baseline by $+0.0688$ EntScore (95% CI $[0.0551, 0.0833]$, $p < 0.001$), is essentially tied with the entropy-gate ablation, and improves by $+0.0423$ over the no-localization ablation. Decomposing by gate decision shows that gains concentrate where the gate triggers clarification (mean delta $+0.0959$) but direct-output reranking also contributes substantially ($+0.0495$); because the same NLI family drives both system decisions and evaluation, these gains should be read as automatic-evaluation gains rather than human factuality improvements. The analysis suggests three concrete next steps: combine semantic agreement with a source-support floor at the gate to catch confident hallucinations, tighten the keep rule so that more useful repairs survive acceptance, and use a larger development set so the chosen threshold transfers more reliably to test data. Promising further directions include a clustering-based hotspot search that groups paraphrastic candidate sentences before computing aggregate risk, and substituting an instruction-tuned LLM for BART in Stage 0 to test whether the observed gains transfer to stronger generators.

CONTRIBUTION STATEMENT

All three authors contributed jointly to scoping the problem, designing the pipeline, and writing the report. Anders Vestrum led the methodological core of the system: the NLI-graph uncertainty signal

U_{global} and the sentence-level hotspot localization. Noah Lund Syrdal led the multi-sample generation pipeline, the penalized-entailment repair acceptance rule with its guardrails, and the ablation and evaluation scripts. Mihail Dimitrov led the construction of the ambiguity-sensitive CNN/DailyMail artifact, the clarification question-generation and answer-resolution components and the calibration protocol that ties the gate threshold to the development split. The team did not work with any collaborators outside of the course.

GenAI Acknowledgement

During the project, two GenAI assistants were consulted: ChatGPT (`https://chat.openai.com`) and Claude (`https://claude.ai`). Their role was bounded to three concrete uses: looking up unfamiliar library behavior in PyTorch and Hugging Face Transformers, sanity-checking and tidying snippets of code we had already drafted, and proof-reading the LaTeX manuscript at the level of grammar and sentence flow. Every method choice in the pipeline (uncertainty gate, hotspot localization, repair acceptance rule, ablation design), every numerical result, and every interpretive paragraph in this report originates with the authors; we did not paste assignment text, dataset identifiers, or experiment prompts into either tool, nor did we ask either tool to draft analysis on our behalf. We have the ability to explain and replicate the work done in this document if asked by an instructor.

A Implementation Details

The project materials include the notebook pipeline, the generated ambiguity artifact, and CSV outputs for calibration, evaluation, and ablation runs. The default open-model path does not require an OpenAI API key; the clarification backend defaults to local open models. GPU execution is strongly recommended because both BART generation and DeBERTa NLI scoring are repeatedly invoked across hundreds of examples.

A.1 Models and Sampling Hyperparameters

The main evaluation uses `facebook/bart-large-cnn` for summarization, `cross-encoder/nli-deberta-v3-large` for NLI scoring, `all-MiniLM-L6-v2` for semantic deduplication, and `google/flan-t5-base` as the default open instruction model. The full pipeline uses $N = 8$ sampled summaries, temperature 1.0, top-$p = 0.88$, semantic deduplication threshold 0.92, and a maximum of 10 sampling attempts. The open one-question baseline uses temperature 0.9 and top-$p = 0.9$.

A.2 Thresholds and Calibration Constants

The exact threshold values used by the implementation are as follows:

- ask-rate calibration band: $[0.20, 0.35]$;
- minimum candidate diversity for asking: $N_{\min}^{\text{clarify}} = 2$;
- hotspot-support cutoff for asking: $\sigma_{\text{ask}} = 0.90$;
- contradiction penalty weight: $\lambda = 0.35$;
- minimum penalized-score gain: $\delta_{\text{pen}} = 0.01$;
- minimum semantic-preservation threshold: $\eta_{\text{pres}} = 0.78$;
- repaired-to-original length-ratio bounds: $\rho_{\min} = 0.65$, $\rho_{\max} = 1.35$.

A.3 Evaluation Constants

The full evaluation uses a 150-example development split and a 350-example test split. Calibration is carried out on the development split, selecting $\tau = -0.27$, and the threshold is then fixed for test evaluation. Statistical reporting uses 5000 bootstrap resamples for confidence intervals and 10000 paired permutations for p-values.

REFERENCES

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. Semantically diverse language generation for uncertainty estimation in language models. *arXiv preprint arXiv:2406.04306*, 2024. doi: 10.48550/arXiv.2406.04306. URL https://arxiv.org/abs/2406.04306.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, Feng Zheng, and Liang He. Enhancing uncertainty modeling with semantic graph for hallucination detection. *arXiv preprint arXiv:2501.02020*, 2025. doi: 10.48550/arXiv.2501.02020. URL https://arxiv.org/abs/2501.02020.

Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. STYLE: Improving domain transferability of asking clarification questions in large language model powered conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10633–10649, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.632. URL https://aclanthology.org/2024.findings-acl.632/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

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. URL https://papers.nips.cc/paper_files/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html.

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. URL https://openreview.net/forum?id=rygGQyrFvH.

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, 2020. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, 2023. doi: 10.18653/v1/2023.ijcnlp-main.20. URL https://aclanthology.org/2023.ijcnlp-main.20/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.17651.

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 Gülçehre, 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, 2016. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028/.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. *arXiv preprint arXiv:2506.00245*, 2025. doi: 10.48550/arXiv.2506.00245. URL https://arxiv.org/abs/2506.00245.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824, 2023. doi: 10.18653/v1/2023.findings-emnlp.248. URL https://aclanthology.org/2023.findings-emnlp.248/.

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/.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14148–14161, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.841. URL https://aclanthology.org/2024.findings-acl.841/.

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/.