Faithful Evaluation of Population Claims in Opinion Summaries

Anonymous ACL submission

Abstract

Opinion summaries routinely make claims like "*most guests loved the location*" and "*almost all reviewers recommend this product*", but existing faithfulness evaluators cannot catch when these claims are wrong.

We show why: population claims require *distributional ground truth* (does the reviewer fraction meet the threshold implied by "most" or "almost all"?), but Natural Language Inference (NLI) evaluators check only *documentary ground truth* (does any span support the claim?). We prove this gap is structural: a supporting span always exists when the claimed direction is correct, so NLI passes overclaiming summaries silently regardless of model quality.

We introduce *distributional verification*, which assigns claims to popularity classes with explicit population-fraction thresholds and issues verdicts of JUSTIFIED, OVERCLAIMED, or WRONGDIRECTION. The overclaiming gap rises from near-zero on highly positive-skewed review corpora to $\sim$8 pp as sentiment distributions become balanced, an increase confirmed across six datasets spanning products, restaurants, hotels, and film reviews.

We further introduce *Temporal Leave-Future-Out* (TLFO): a summary generated at time T should be verified against the *future* distribution, turning any timestamped corpus into a free evaluation benchmark. Across three corpora, claims justified at snapshot time show 11–21 pp higher overclaim rates after distributional change-points.

A human annotation study (κ=0.897) validates the verdict pipeline and finds a 22.0% overclaim gap among direction-correct claims. We release a temporally-indexed product review benchmark and a 424-claim human-annotated verdict dataset.

1 Introduction

Opinion summarization systems are routinely evaluated for faithfulness: does the summary say things that are grounded in the source reviews? Existing evaluators operationalize faithfulness as *documentary faithfulness*: for each claim in the summary, retrieve a relevant span from the source and run NLI to check entailment (Min et al., 2023; Tang et al., 2024; Kryściński et al., 2020).

This approach is appropriate for *atomic claims* such as "*the battery lasts 12 hours*" or "*the hotel is near the Vatican*," but ill-suited for population claims such as "*most reviewers loved the battery*" or "*guests overall found the location convenient*." A population claim is true or false relative to a statistic over the reviewer population, not relative to any individual review span. A supporting span for "*most reviewers loved the battery*" always exists in the corpus when at least one reviewer did love it, but that span provides no information about whether the majority loved it.

This incompleteness is *silent*: NLI-based evaluators score overclaiming summaries as faithful because they find a supporting span, never checking the fraction. The same failure affects LLM-based evaluators (e.g., RAGAS (Es et al., 2024)): without access to the reviewer fraction, no text-based verifier can distinguish a justified from an overclaimed population claim, a point confirmed empirically in §5. We call this *overclaiming*: the summary's consensus language is directionally correct but quantitatively overstated. Our contributions are:

1. A formal taxonomy distinguishing atomic claims (documentary ground truth) from population claims (distributional ground truth), and a proof that NLI-based verification is structurally incapable of detecting overclaiming in the latter (§3).

2. Distributional verification: a claim-level eval-

uation method that assigns claims to popularity classes with population-fraction thresholds and issues JUSTIFIED, OVERCLAIMED, or WRONGDIRECTION verdicts (§4).

3. Cross-dataset validation across six datasets confirming the predicted gap gradient and its broadly monotonic dependence on distributional diversity (§5).

4. Temporal Leave-Future-Out (TLFO): evaluation against the future distribution, demonstrated across three independent corpora (§6).

5. AmaSum-T: a temporally-indexed product benchmark, and a 424-claim human-annotated verdict dataset ($\kappa = 0.897$; inter-annotator agreement subset (IAA): $\kappa = 0.922$) confirming a 22.0% overclaim gap.

2 Background and Related Work

Faithfulness evaluation. FActScore (Min et al., 2023) decomposes generated text into atomic facts and verifies each against retrieved Wikipedia passages via NLI. MiniCheck (Tang et al., 2024) trains an efficient NLI model optimized for factual grounding. FENICE (Scirè et al., 2024) and SummaC (Laban et al., 2022) apply NLI to summarization faithfulness. All implement existential satisfiability: a claim is faithful if *some* retrieved span entails it. This is the correct criterion for atomic claims; we show it is structurally wrong for population claims (§3).

Opinion summarization. Extractive methods select representative review sentences (Angelidis et al., 2021). Abstractive methods (Bražinskas et al., 2020; Iso et al., 2021) generate novel language. Recent LLM-based approaches include large-scale incremental summarization (Muddu et al., 2024), aspect-aware decomposition (Li et al., 2025), and facet-guided generation (Wang et al., 2025). Evaluation has focused on ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and LLM-as-judge metrics; none of these natively evaluate whether consensus quantifiers are calibrated to rating distributions; we examine LLM judges as a baseline in §5.

Distributional bias in opinion summarization. Several recent papers identify systematic distributional distortions in opinion summaries. Huang et al. (2023) document that summaries over-represent majority sentiment and under-represent minority views. Huang et al. (2025) propose REFER, a frequency-framed prompting strategy that biases generation toward underrepresented opinion fractions. Lei et al. (2024) apply reinforcement learning with a polarity-distance reward to encourage balanced polarity in generated summaries. Most recently, Aghaebe et al. (2026) propose faithful belief aggregation for opinion summarization, explicitly preserving the distribution of expressed beliefs during synthesis. These papers address *generation*: they build summarizers that produce more calibrated output. Our work addresses *evaluation*: given any summary (including those produced by bias-mitigating systems such as REFER, polarity-balanced generation (Lei et al., 2024), or faithful aggregation), we ask whether its population claims are calibrated to the actual rating distribution. To our knowledge, no prior work defines population claims as a distinct claim type, proves that NLI cannot verify them, or connects popularity-class thresholds (§3) to distributional ground truth.

Temporal evaluation. Lazaridou et al. (2021) and Luu et al. (2022) show that NLP models degrade as the world changes after their training cutoff. TLFO applies the same insight to opinion summaries: a summary generated from pre-T reviews may overclaim once post-T reviews reveal a shifted distribution, and the future reviews serve as a free ground-truth oracle requiring no additional annotation.

3 Population Claims and the NLI Impossibility

Opinion summaries make two structurally distinct kinds of claims that existing faithfulness evaluators treat identically. We formalize the distinction because the choice of verification function follows directly from it.

3.1 Two Claim Types

Definition 3.1 (Atomic claim)**.** *An* atomic claim *a is a proposition that is true or false relative to one or more evidential spans in the source corpus R. Formally, a is* supported *if $\exists\, s \in R$ such that $s \models a$ under an NLI entailment relation $\models$. Atomic claims have* documentary ground truth*: the locus of truth is a set of retrievable text spans.*

Example. "*The battery charges in 90 minutes*" is atomic. Supported if any review contains a span entailing this duration.

Popularity class	Example quantifiers	τ_c
Universal	*unanimously, almost all, nearly everyone*	≥ 0.85
Majority	*most, the majority, generally, overall*	≥ 0.55
Plural minority	*many, a large number of*	≥ 0.35
Existential	*some, a few, several*	≥ 0.10

Table 1: Popularity classes with minimum population fractions (Def. 3.3).

Definition 3.2 (Population claim)**.** *A* population claim p *is a proposition about the fraction of a reviewer population* $\mathcal{R}$ *holding polarity* $\pi \in \{+, -\}$*:*

$$p \equiv \left[\text{pos_frac}(\pi, \mathcal{R}) \geq \tau_c\right] \quad (1)$$

where $\tau_c \in (0, 1]$ *is the consensus threshold encoded by the claim's consensus quantifier (hereafter* cue word c*) and*

$$\text{pos_frac}(\pi, \mathcal{R}) = \frac{|\{r \in \mathcal{R} : \textit{polarity}(r) = \pi\}|}{|\mathcal{R}|}.$$

When $\mathcal{R}$ *is fixed, we write* $\text{pos_frac}(\pi)$*; for* $\pi{=}+$, $\text{pos_frac}(\mathcal{R})$*. Population claims have* distributional ground truth*: the locus of truth is a statistic over the population, not any individual span. Here* $\textit{polarity}(r)$ *is topic-scoped – overall sentiment or aspect-specific (e.g., location); in this paper we operationalize it via overall star ratings, which are directly observable across all datasets (see Limitation 4, §7).*

Example. "*Most reviewers found the battery sufficient*" is a population claim with $c = \texttt{most}$, $\tau_c = 0.55$, $\pi = +$. It is true iff $\geq 55\%$ of reviewers rate the battery positively.

Definition 3.3 (Popularity classes)**.** *We define four* popularity classes*, each with a minimum required population fraction* τ_c *(Table 1). A population claim's consensus quantifier (e.g., most, some) places it in a class – a straightforward classification for any LLM or lexical matcher. These thresholds are calibration conventions grounded in quantifier semantics, treated as operationally defined and validated empirically in §5.*

3.2 The NLI Verification Function

Definition 3.4 (NLI-based faithfulness)**.** *A claim* q *is* NLI-faithful *with respect to source* R *if*

$$\exists\, s \in R \;\; s.t. \;\; P_{\text{NLI}}(\text{entail} \mid s,\, q) \geq \theta \quad (2)$$

where P_{NLI} *is an NLI model's entailment probability and* θ *is a decision threshold.*

FActScore retrieves relevant spans before the NLI step; MiniCheck uses a grounding-optimized model; SummaC operates sentence-by-sentence – all implement the same existential satisfiability check of Eq. (2). This is the correct verification function for atomic claims: the claim is grounded if some span supports it.

3.3 NLI Cannot Verify Population Claims

Theorem 3.1 (Structural non-contradiction)**.** *For any overclaimed positive-polarity population claim* p *with* $\text{pos_frac}(\mathcal{R}) = f > 0$*, there always exists a span* $s^* \in \mathcal{R}$ *such that: (i)* s^* *does not contradict* p *under any NLI model without access to the full population count; and (ii)* s^* *directly instantiates the property asserted by* p*. Consequently, the existential check in Eq. (2) cannot distinguish* $(s^*, p_{\textit{overclaimed}})$ *from* $(s^*, p_{\textit{justified}})$ *on structural grounds alone. The same argument holds for negative-polarity claims whenever* $f^- = \text{pos_frac}(\pi{=}-, \mathcal{R}) > 0$*.*

Proof sketch. Since $\text{pos_frac}(\mathcal{R}) = f > 0$, at least one review $r^* \in \mathcal{R}$ expresses positive sentiment; extract span $s^* \subset r^*$ instantiating the claimed property. Under any NLI model without population-count access, (s^*, p) receives at minimum NEUTRAL – regardless of how far f falls below τ_c. See Appendix F for the full argument. □

Corollary 3.1.1 (Silent failure)**.** *NLI models that assign* ENTAIL *to positive-instance spans will pass every overclaimed population claim whose direction is correct, regardless of the degree of overclaiming, given the existential design of Eq. (2).*

Remark. That NLI models assign ENTAIL to positive-instance spans is the common empirical case; it is confirmed in §5.3 where `roberta-large-mnli` passes 100% of both JUSTIFIED and OVERCLAIMED claims at near-identical scores.

In practice: a 3-star product whose summary says "almost all reviewers loved it" passes NLI identically to a 4.9-star product – a positive span exists in both cases, and the §5.3 experiment confirms NLI assigns entailment at 100% on human-labeled data where direction is unambiguous.

A formal counting-gap corollary and a threshold-boundary proposition (with proofs) are given in Appendix F. The key practical implication: overclaiming is not a model failure – it is a structural consequence of representing continuous distributions with discrete language.

Empirical calibration. A template baseline mechanically selecting the highest cue whose threshold $\leq \text{pos_frac}(\mathcal{R})$ produces 0% overclaiming by construction – the floor all real summarizers exceed (§5). JUSTIFIED and OVERCLAIMED claims both pass NLI; WRONGDIRECTION claims are only partially detectable, and score calibration (Min et al., 2023) cannot transform an existential verifier into a population counter.

4 Distributional Verification

The correct verification function replaces the existential span check with a fraction comparison. Given claim p with cue c, polarity π, and distribution (f^+, f^-):

- JUSTIFIED: $\text{pos_frac}(\pi) \geq \tau_c$ (fraction meets threshold).
- OVERCLAIMED: $\text{pos_frac}(\pi) > 0$ but $\text{pos_frac}(\pi) < \tau_c$ (some evidence supports the claimed polarity direction, but the fraction is too low to meet the threshold – the typical overclaiming case).
- WRONGDIRECTION: $\text{pos_frac}(\pi) \approx 0$ – the claimed polarity direction is essentially absent from the distribution (e.g., a positive claim on a predominantly-negative-reviewed product with near-zero positive fraction).

The human annotation study (424 claims, §5.3) confirms this distinction: negative-polarity existential claims on 92%-positive products ("*Some report that the pump broke*", $f^- = 6\%$) receive OVERCLAIMED not WRONGDIRECTION, because $f^- > 0$; the direction is partially present, only the quantified fraction is overstated. The automated pipeline uses a proxy (dominant polarity $\neq$ claimed polarity) that over-assigns WRONGDIRECTION; the human-validated benchmark adopts the $\text{pos_frac}(\pi) > 0$ / $\text{pos_frac}(\pi) \approx 0$ distinction, yielding WRONGDIRECTION 3.1% vs. the automated proxy's 22.7%. The only change from NLI: ground truth is $\text{pos_frac}(\pi, \mathcal{R}) \geq \tau_c$, not $\exists s \in \mathcal{R}$.

Popularity class assignment. Each candidate sentence is assigned to a popularity class (Def. 3.3) by its consensus quantifier; our pipeline uses a 22-entry quantifier lexicon matched over a ± 3-token window for reproducibility. Polarity is the sign of the net sentiment count (#positive $-$ #negative lexicon matches) over the same window. Claims whose pos_frac falls within ± 0.10 of τ_c are excluded as statistically unreliable. Thresholds follow scalar implicature research (Papafragou and Musolino, 2003) and are validated by the annotation study (κ=0.922); both lexicons are released with the benchmark.

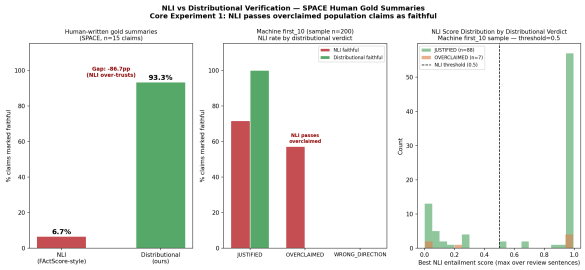

Figure 1: NLI vs. distributional verification on SPACE human gold summaries (left two panels) and machine-generated first-10 summaries (right panel). NLI passes 57.1% of distributional-OVERCLAIMED claims as faithful. Entailment scores are indistinguishable between JUSTIFIED and OVERCLAIMED claims (right), confirming the structural nature of Theorem 3.1.

LLM extractor ablation. A `gpt-4.1-mini` ablation on the IAA subset shows the LLM extractor achieves higher precision than the lexical baseline but misses 32% of genuine population claims. Because benchmark claims are human-validated independently of extractor choice, overclaim gap numbers (Table 2) are identical under both strategies.

Canonical claim index. We precompute per-entity fractions f^+, f^- for $O(1)$ verification; for TLFO the index is time-sliced at the PELT (Pruned Exact Linear Time; Killick et al. 2012) change-point t^* with largest $|\Delta\text{pos_frac}|$. A *TLFO overclaim* is a claim that is JUSTIFIED at snapshot time but OVERCLAIMED or WRONGDIRECTION when verified against the post-t^* distribution.

SPACE NLI comparison. We quantify the structural failure empirically in §5: on SPACE's 50-hotel set, 57.1% of distributional-OVERCLAIMED claims are scored FAITHFUL by `roberta-large-mnli` (Figure 1). The remainder fail because retrieved spans are weakly entailing in borderline cases – not because NLI detects the overclaiming.

5 Experiments

5.1 Datasets

We evaluate on six datasets spanning four domains (Table 5 in Appendix C): AmaSum (Bražinskas et al., 2021) (19K Amazon products,

pos_frac ≈ 0.85, human Verdict/Pros/Cons summaries), AmaSum-T (our temporally-indexed extension, 14K ASINs, pos_frac ≈ 0.83, §6), Flipkart[1] (pos_frac ≈ 0.71, 363K reviews), SPACE (Angelidis et al., 2021) (8.8K TripAdvisor hotels, pos_frac ≈ 0.71, 1,050 gold summaries), Yelp[2] (150K restaurants, pos_frac ≈ 0.65, 7M reviews), and Rotten Tomatoes[3] (RT) (pos_frac ≈ 0.55, 9,121 critics' consensus texts; pos_frac is the fraction of critics scoring "fresh"). The pos_frac gradient – from highly skewed (AmaSum) to balanced (Yelp, RT) – is the key experimental design choice, enabling us to test whether the overclaiming gap scales with distributional diversity as predicted by Theorem 3.1.

5.2 Population Claim Gap Across Datasets

Metric and baselines. The *naive direction-check pass rate* is the fraction of claims whose polarity direction is consistent with the dominant sentiment (i.e., not WRONGDIRECTION), ignoring whether the population fraction meets the popularity-class threshold. The *overclaim gap* is the fraction of direction-correct claims that are OVERCLAIMED (direction correct but threshold not met): positive values indicate the verified justified rate is strictly lower than the naive rate, meaning the naive metric overstates faithfulness. We also include a *template* baseline that mechanically selects the highest cue whose $\tau_c \leq \text{pos_frac}(\mathcal{R})$, producing 0% overclaiming by construction and serving as the calibration floor against which other systems are compared. We measure all baselines on three intentionally naive extractive summaries – First-10 (first 10 reviews), Recent-10 (most recent 10), and Extractive top-3 (TF-IDF centrality) – averaged to smooth baseline-specific effects. These serve as a controlled lower bound: because they directly copy review language, any overclaiming is attributable to distributional mismatch rather than summary-generation artifacts. Where available we also evaluate human-written and LLM-generated summaries. Table 2 shows results.

The overclaim gap is broadly monotonic in pos_frac: near-zero at the highly positive-skewed end (AmaSum, AmaSum-T), rising to ∼7–8 pp at SPACE, Yelp, and Rotten Tomatoes (Figure 2). The rate is stable across N=10 to N=200 reviews on Yelp, confirming the failure is structural: more reviews do not help. RT shows 8 pp despite being more balanced because editors use hedged language that reduces strong-threshold cue frequency. AmaSum-T shows near-zero overclaim (∼1 pp) but elevated wrong-direction (∼11%): Cons-section existential cues on 83%-positive products.

Human summaries. SPACE provides 1,050 human-written gold summaries; Rotten Tomatoes provides 9,121 editor-written critics' consensus texts. Human annotators on SPACE overclaim at 3.8–6.7% with near-zero wrong-direction errors – they correctly identify polarity direction but casually use stronger consensus language than the distribution warrants (Figure 5, Appendix B). RT editors show 44% overclaim at the majority tier ($\tau_c = 0.55$–0.84). Human-written summaries exhibit the same overclaiming pattern as machine baselines across both datasets, consistent with the failure being a property of natural language description rather than of any particular system.

LLM judge baseline. We ran `gpt-4.1-mini` on all 225 unambiguous-verdict benchmark claims (424 total minus NOTAPOPULATIONCLAIM and split-verdict; see §5.3) under *blind* (no fraction disclosed) and *informed* (pos_frac + threshold disclosed) conditions. The blind judge correctly detects only 20.8% of OVERCLAIMED claims, passing the remaining 79% as faithful – the same structural failure as NLI. The informed condition improves detection to 68.8% but marks 59.4% of JUSTIFIED claims wrong – a false-positive rate making it unusable. Injecting the exact fraction reduces the LLM to a noisy threshold comparison that distributional verification performs directly (Table 3; IAA-subset pilot in Appendix A).

5.3 Annotation Study and Human Validation

Three annotators independently labeled a 94-claim IAA subset randomly sampled from AmaSum candidate claims before verdict filters were applied, assigning JUSTIFIED, OVERCLAIMED, WRONGDIRECTION, or NOTAPOPULATIONCLAIM (NAPC); see Appendix G for full guidelines and annotator details. The task proved highly reliable (Fleiss' κ=0.922, "almost perfect"; pairwise $\kappa \geq 0.85$; per-class κ: 0.90 JUSTIFIED, 0.91 OVERCLAIMED), with 90.4% unanimous three-way agreement across all 94 claims (85/94); 66 received a distributional verdict; 63 are used in the NLI ex-

[1] https://www.kaggle.com/datasets/niraliivaghani/flipkart-dataset
[2] https://www.yelp.com/dataset
[3] https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

Dataset	Domain	f^+	Naive pass	Distrib. justified	Overclaim gap	Wrong-dir
AmaSum	Amazon products	∼0.85	∼90%	∼88%	∼2 pp	∼10%
AmaSum-T	Amazon products	∼0.83	∼89%	∼88%	∼1 pp	∼11%
Flipkart	Consumer products	∼0.71	∼99%	∼99%	∼0 pp	∼1%
SPACE	TripAdvisor hotels	∼0.71	∼89%	∼82%	∼7 pp	∼11%
Yelp	Restaurants	∼0.65	∼92%	∼84%	∼8 pp	∼8%
Rotten Tomatoes	Film criticism	∼0.55	∼79%	∼71%	∼8 pp	∼21%

Table 2: Distributional verification results across datasets (machine baselines only: averaged over First-10/Recent-10, All reviews, Extractive top-3; human-written summary results are in Figure 5 and §5.2). *Naive pass*: fraction of claims whose polarity direction is consistent with dominant sentiment (i.e., not WRONGDIRECTION). *Overclaim gap*: fraction of direction-correct claims that fail the distributional threshold (OVERCLAIMED); the gap is broadly monotonic in distributional diversity (see Figure 2). *Wrong-dir*: fraction with opposite-polarity direction. AmaSum-T and AmaSum show near-zero overclaim gap but elevated wrong-direction rates from Cons-section existential cues applied to overwhelmingly-positive products.

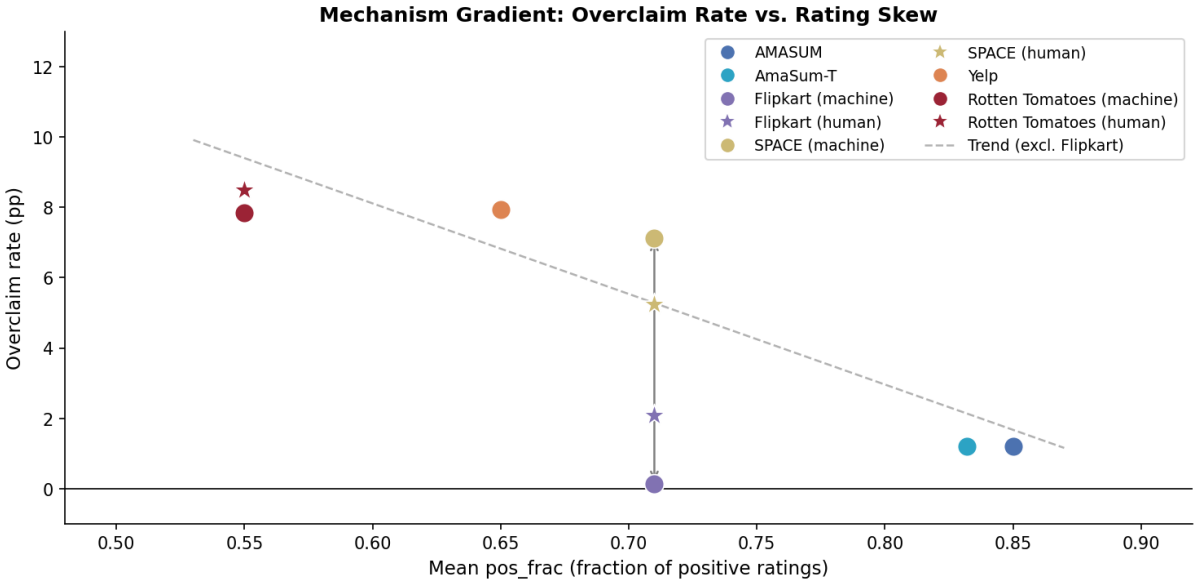

Figure 2: Overclaim rate (fraction of direction-correct claims that are OVERCLAIMED) vs. mean pos_frac across all datasets. Circles = machine baselines; stars = human-written summaries. The dashed line shows the linear trend for machine baselines. The overclaim gap is broadly monotonic in distributional diversity: highly positive-skewed datasets (AmaSum, AmaSum-T) show near-zero overclaim rates (∼1 pp), while more balanced datasets (SPACE, Yelp, Rotten Tomatoes) show overclaim rates of 7–8 pp. AmaSum-T additionally has ∼11% wrong-direction claims from Cons-section negative cues (see Table 2). The bracket at pos_frac≈0.71 marks the Flipkart/SPACE anomaly: despite identical pos_frac, Flipkart's overclaim rate is near zero because 99.9% of its claims use "some" (τ_c=0.15, trivially satisfied), while SPACE's long prose generates high-threshold cues ("overall," "everyone") that challenge the threshold boundary.

periment below (3 existential-tier borderline cases excluded) and all 66 in the LLM judge pilot (Appendix A).

Two annotators then independently labeled the remaining 330 claims (3 annotators rotating, 2 per claim), yielding 424 total annotated claims. Agreement remained high (Cohen's κ=0.897; 93.6% unanimous). Of the 225 claims with an unambiguous verdict (excluding 171 NOTAPOPULATIONCLAIM and 28 with no majority), human labels show 75.6% JUSTIFIED, 21.3% OVERCLAIMED, and 3.1% WRONGDIRECTION. Among direction-correct claims (JUSTIFIED+OVERCLAIMED), the overclaim gap is **22.0%** (48/218 claims), with a 21.3 pp difference between naive direction-only pass rate (96.9%) and distributional verification (75.6%). Overclaiming is concentrated in Cons-section claims (90.2% overclaim rate) versus Pros-section claims (3.6%), confirming the structural source: summaries asserting negative reviewer fractions for overwhelmingly positive products.

Verifier	Acc.	OC det.	FP%
NLI (roberta-lg)	50.0	0.0	0.0
LLM blind	60.0	20.8	26.5
LLM informed	46.2	68.8	59.4
Distrib. verif.	$\approx 92^{\dagger}$	$\approx 100^{\dagger}$	$\approx 0^{\dagger}$

Table 3: Verifier comparison on 225 benchmark claims (170 JUSTIFIED, 48 OVERCLAIMED, 7 WRONGDIRECTION). *Acc.*: % agreement with human majority. *OC det.*: % of OVERCLAIMED claims correctly flagged. *FP%*: % of JUSTIFIED claims incorrectly flagged. NLI passes all claims as faithful; 50.0% = balanced accuracy (JUSTIFIED recall: 100%, OVERCLAIMED recall: 0%). LLM blind misses 79% of overclaims; LLM informed has 59.4% false positives. †By construction (κ=0.897).

Extractor precision. 28 of 94 IAA claims (30%) received a majority NOTAPOPULATIONCLAIM label from two false-positive patterns: cue words modifying product attributes ("overall design") or use contexts ("works in many areas") rather than reviewer fractions. Targeted filters for both patterns yield 424 validated claims from 2,100+ candidates.

Natural Language Inference (NLI) experiment. We ran `roberta-large-mnli` on the 63 valid claims (48 human-JUSTIFIED, 15 human-OVERCLAIMED), scoring each claim against all review spans and taking the maximum entailment probability (following the FActScore retrieval-then-NLI protocol (Min et al., 2023)). NLI assigned ENTAIL to 100% of claims in both categories (48/48 JUSTIFIED, 15/15 OVERCLAIMED), at nearly identical mean entailment probabilities ($\bar{p}$=0.990 vs. $\bar{p}$=0.989). This confirms Theorem 3.1 on human-labeled data: NLI is structurally unable to distinguish the two verdict classes. Figure 1 visualizes the same pattern on a parallel SPACE experiment with automated labels.

WRONGDIRECTION. WRONGDIRECTION is rare in human labels (3.1%, 7/225): annotators classify negative existentials on positive products as OVERCLAIMED (direction partially present, fraction overstated), not WRONGDIRECTION (direction absent). The automated pipeline over-assigns WRONGDIRECTION via a dominant-polarity proxy (22.7%); the human-validated benchmark corrects this, yielding JUSTIFIED 75.6%, OVERCLAIMED 21.3%, WRONGDIRECTION 3.1%.

6 Temporal Leave-Future-Out (TLFO)

Motivation. Static distributional evaluation verifies a summary against the same snapshot used to generate it, guaranteeing agreement even when the underlying opinion distribution has shifted – a product firmware update, a restaurant ownership change, a hotel renovation can render previously-justified claims overclaimed. TLFO reframes the question: given a summary generated at time T using reviews in $[0, T]$, do its population claims remain calibrated against the *future* distribution in $(T, T+\Delta]$? The future distribution serves as a free oracle – no additional annotation is required – so any timestamped corpus is automatically a temporal evaluation benchmark (formal definition in §4). Across our timestamped datasets, 72–95% of entities show a detectable temporal signal (Table 5), confirming that distribution shift is a pervasive rather than edge-case concern.

Yelp TLFO demonstration. We define *stable* businesses as those for which PELT detects no significant change-point (used as the control group). We split each change-point business's reviews at its detected change-point, use the first-10 pre-T reviews as the snapshot text, and verify its population claims against the post-T distribution. Strong-consensus claims ($\tau_c \geq 0.85$) show 11 pp higher overclaim in change-point businesses than stable ones (two-sample t-test, $p = 0.024$, uncorrected; Figure 3). Weaker cues show no signal: they are near-trivially satisfied at Yelp's $\mathrm{pos_frac} \approx 0.65$, so drift does not push them past their thresholds. The TLFO signal concentrates at the tier whose threshold is closest to the dataset's mean $\mathrm{pos_frac}$ – the tier most easily falsified by a moderate shift.

AmaSum-T. We match AmaSum ASINs (Bražinskas et al., 2021) to Amazon Reviews 2023, recovering per-review timestamps: 14,015 ASINs (72.5% match), 836K reviews, all with a 5+ year post-2018 window. Of these, 9,676 have ≥10 pre-2018 and ≥5 post-2018 reviews (TLFO-eligible); using first-10 pre-2018 reviews as the snapshot, products with detectable drift ($|\Delta\mathrm{pos_frac}| \geq 0.10$ post- vs. pre-cutpoint) show 20.8 pp higher overclaim than stable products ($|\Delta\mathrm{pos_frac}| < 0.10$) at the Universal tier ($\tau_c \geq 0.85$; p<0.0001; Figure 4). The signal concentrates here because AmaSum-T's $\mathrm{pos_frac} \approx 0.83$ sits just below the Universal threshold – the same mechanism as Yelp and

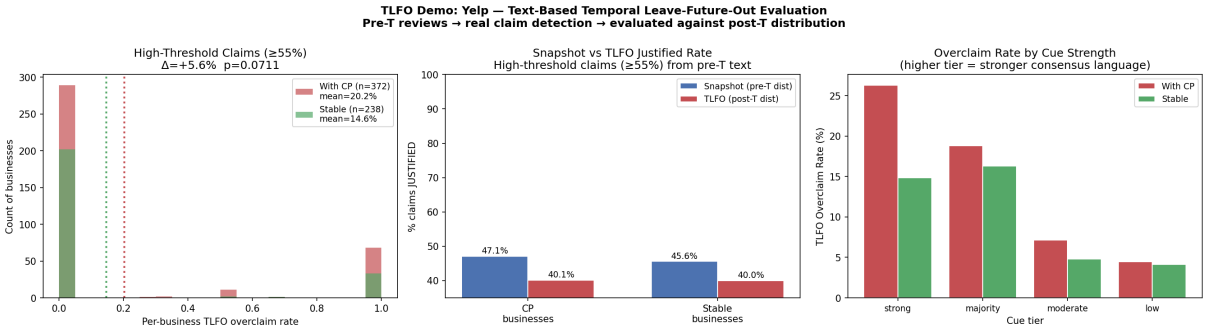

Figure 3: Yelp TLFO demonstration. **Left**: per-business TLFO overclaim rate distributions for change-point (CP) vs. stable businesses. **Center**: snapshot (pre-T) vs. TLFO (post-T) justified rate for high-threshold claims ($\tau_c \geq 0.55$); CP businesses degrade more. **Right**: TLFO overclaim rate by cue tier – Universal-tier cues ($\tau_c \geq 0.85$) show the clearest separation between CP and stable businesses.

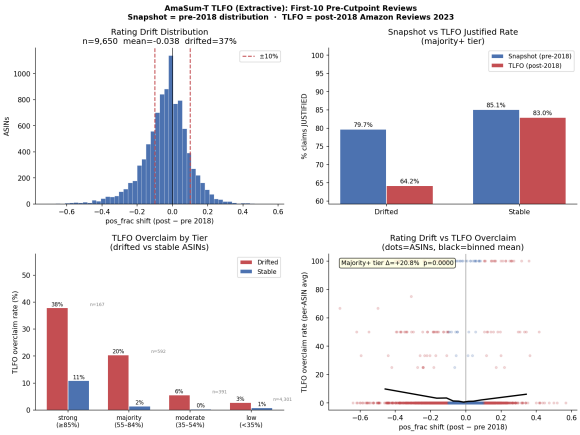

Figure 4: AmaSum-T extractive TLFO (n=9,676 ASINs, first-10 pre-2018 reviews as snapshot). Panel 1: pos_frac shift distribution; 37% of ASINs drift by ≥10 pp. Panel 2: snapshot vs. TLFO justified rate (Universal tier, $\tau_c \geq 0.85$). Panel 3: TLFO overclaim rate by tier, drifted vs. stable. Panel 4: Universal tier $\Delta = +20.8$ pp, p<0.0001.

SPACE. PrefLib TripAdvisor reviews for 1,713 SPACE-overlapping hotels replicate the tier-specificity prediction ($\Delta = +11.7$ pp, $p = 0.0023$; Appendix D). LLM-generated AmaSum summaries show no TLFO signal ($p = 0.36$): they hedge by omitting strong quantifiers (Appendix E).

7 Discussion

Generality. The population claim problem applies to any corpus-summarization domain (scientific synthesis, medical evidence review, news aggregation) where NLI-based faithfulness metrics have been adopted (Kryściński et al., 2020; Laban et al., 2022; Tang et al., 2024); opinion summarization is the clearest instantiation because star ratings provide a directly observable distributional ground truth.

8 Conclusion

We introduced population claims as a distinct claim type with distributional ground truth, proved that NLI-based faithfulness evaluation is structurally incapable of detecting overclaiming in such claims, and proposed distributional verification as the correct evaluation path. Across six datasets spanning four domains, the overclaim rate rises from near-zero on highly positive-skewed corpora to ∼8 pp as distributions become balanced, with the gap predicted by distributional diversity. TLFO extends distributional verification to the temporal dimension, turning every timestamped review corpus into a free evaluation benchmark. A human annotation study (IAA: 94 claims, κ=0.922; full: 424 claims, κ=0.897) validates the automated verdict pipeline and reveals a 22.0% overclaim gap among direction-correct claims, with overclaiming concentrated in Cons-section negative existentials on positive products. We release AmaSum-T, the 424-claim human-annotated benchmark, and the quantifier and sentiment lexicons.

Limitations

(1) Popularity class assignment is English-only; multilingual extension is left to future work. (2) Human annotation of 424 AmaSum claims (κ=0.897; IAA subset: κ=0.922) reveals a 22.0% overclaim gap; cross-domain annotation on SPACE and Yelp is left to future work. (3) TLFO requires timestamped reviews (Flipkart omitted; SPACE cross-checked via PrefLib, Appendix D). (4) Distributional ground truth is operationalized via overall star ratings. Aspect-specific verification (e.g.,

verifying location claims against location-aspect ratings) requires aspect-labelled corpora; SPACE provides per-aspect ratings and is a natural next step. (5) AmaSum and SPACE are used under academic research licenses; Yelp, Flipkart, and Rotten Tomatoes data are accessed under their respective terms of service for non-commercial research use. (6) The distributional verification pipeline requires no GPU (all verification is $O(1)$ lookup against a precomputed index). The NLI experiment uses `roberta-large-mnli` ($\sim$355M parameters) on 63 claims; the LLM judge uses the `gpt-4.1-mini` API on 225 claims. No model training was performed.

Risks. The primary risk is metric gaming: once popularity-class thresholds are published, systems optimized against distributional verification could learn to select cue words that hover just above a threshold rather than accurately representing the distribution. This is a general concern for any published evaluation benchmark. A secondary risk is rating bias: operationalizing sentiment via star ratings may encode demographic or cultural rating tendencies; summaries evaluated against a biased distribution will inherit that bias (Limitation 4 above). No sensitive personal data is collected or released; the method is not applicable to domains where misuse is a primary concern.

Acknowledgements

[Omitted for anonymous review.]

AI assistance. Claude (Anthropic) was used to assist with writing and editing portions of this paper. All content was reviewed and verified by the authors.

References

Sepideh Aghaebe, Kaveh Taghipour, and Hwee Tou Ng. 2026. Faithful belief aggregation and synthesis for opinion summarization. arXiv preprint. Preprint.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 400–415, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Few-shot learning for opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4119–4135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

James Y. Huang, Haytham M. Fayek, and Zhuohan Zhang. 2025. REFER: Mitigating bias in opinion summarisation via frequency framed prompting. In *Proceedings of the EMNLP Workshop on New Frontiers in Summarization (NewSum)*, Miami, Florida, USA. Association for Computational Linguistics.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex aggregation for opinion summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3885–3903, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Angeliki Lazaridou, Adhiguna Kuncoro, Edward Grefenstette, Dani Yogatama, Grzegorz Chrupała, Adina Williams, Oriol Vinyals, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc.

Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024. Polarity calibration for opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Decomposed opinion summarization with verified aspect-aware modules. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24805–24841, Vienna, Austria. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Shikhar Vashishth, Jill Noy, Christopher Re, and Percy Liang. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Nicholas Mattei and Toby Walsh. 2013. PrefLib: A library for preferences. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT)*, pages 259–270. Springer.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Tejpalsingh Siledar, Swaroop Nath, Pushpak Bhattacharyya, Swaprava Nath, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Sudhanshu Shekhar Singh, and Nikesh Garera. 2024. Distilling opinions at scale: Incremental opinion summarization using XL-OPSUMM. arXiv:2406.10886. Preprint.

Anna Papafragou and Julien Musolino. 2003. Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, 86(3):253–282.

Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7431–7448, Bangkok, Thailand. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Jian Wang, Yanjie Liang, Yuqing Sun, and Bin Gong. 2025. Inducing argument facets for faithful opinion summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A LLM Judge: IAA Pilot Results

The main text (Table 3) reports LLM judge results on the 225-claim full benchmark. Here we report the earlier 66-claim IAA pilot (49 JUSTIFIED, 17 OVERCLAIMED) for completeness. Results are directionally consistent with the main benchmark.

Condition	Acc.	OC det.	FP%
LLM blind	57.6	29.4 (5/17)	14.3
LLM informed	47.0	82.4 (14/17)	34.7

Table 4: IAA pilot (66 claims, 49 JUSTIFIED, 17 OVERCLAIMED). Informed condition detects more overclaims (82.4% vs 68.8% full benchmark) because the IAA subset excludes Cons-section claims. Both pilots confirm distributional verification dominates LLM judge approaches.

B Additional Figures

Figure 5 provides the full per-baseline breakdown of the overclaiming gap on SPACE, complementing the summary in Table 2 and the human-summary discussion in §5.2.

C Dataset Statistics

Table 5 provides full statistics for all six datasets. *Disagree* is the fraction of entities with predominantly negative opinion distributions ($\text{pos_frac} \leq 0.25$), which is the range where wrong-direction errors are most likely for positive-polarity claims. *Temp. sig.* is the fraction of entities with a detectable temporal signal (either a significant trend or a PELT change-point).

D PrefLib Cross-Check (SPACE Hotels)

The PrefLib TripAdvisor dataset (Mattei and Walsh, 2013) provides dated reviews (2001–2009) for 1,713 hotels whose IDs overlap with SPACE. Using first-10 pre-change-point PrefLib reviews as extractive snapshots (530 hotels with detectable cue words, 1,439 claims), the TLFO overclaim signal concentrates at the Plural minority tier ($\Delta = +11.7$ pp, $p = 0.0023$) – consistent with SPACE's $\text{pos_frac} \approx 0.71$, where Plural minority thresholds (35–54%) are the binding constraint and Majority-tier claims ($\tau_c \geq 0.55$) are trivially satisfied. This confirms the tier-specificity prediction on a third independent corpus.

E LLM Hedging and TLFO

AmaSum's GPT-4-generated summaries (132 ASINs, Verdict/Pros/Cons bullets) show no significant TLFO overclaim signal ($p = 0.36$): LLM bullets avoid explicit majority consensus quantifiers, leaving too few cue-word claims for the test to have power ($\sim$1.2 claims/ASIN vs. $\sim$4.7 for extractive). LLM summarizers implicitly hedge by omitting strong consensus language – sidestepping TLFO overclaiming at the cost of weaker consensus signals. This suggests that the TLFO test is most informative for extractive or template-based summarizers that do surface explicit quantifiers.

F Additional Proofs

Proof of Theorem 3.1. Let p be a population claim "Most reviewers found X satisfactory" with $\tau_c = 0.55$. Suppose the claim is overclaimed: $\text{pos_frac}(\mathcal{R}) = f < 0.55$.

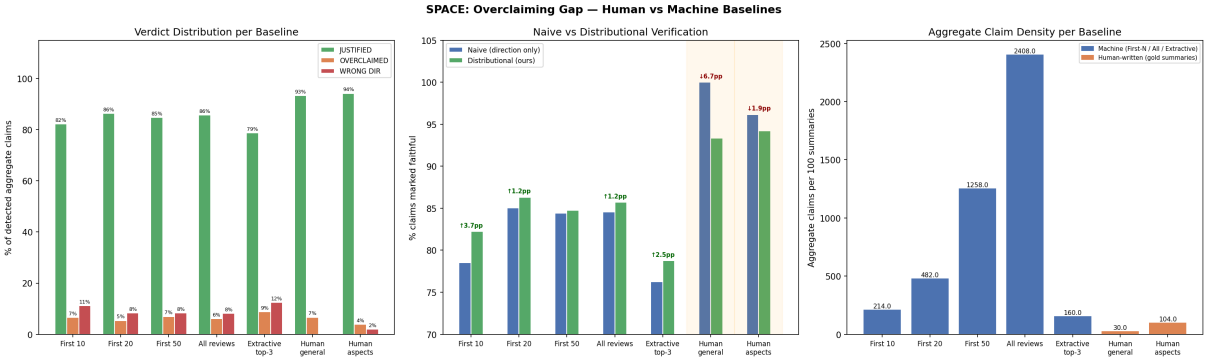

Figure 5: SPACE overclaiming gap: verdict distribution (left), naive vs. distributional pass rate (center), and claim density (right) for machine and human-written baselines on the 50-hotel SPACE eval set. Human gold summaries (orange-shaded columns, center panel) show the same overclaiming gap as machine baselines, confirming the failure is structural.

Dataset	Domain	Entities	Reviews	pos_frac	Disagree	Temp. sig.	Human summ.
AmaSum	Amazon products	19,319	3.5M	0.85	5.5%	71.8%	V/P/C[‡]
AmaSum-T	Amazon products	14,015	836K	0.83	–	75.3%	– (extr.)
Flipkart	Consumer products	1,173	363K	0.71	19.6%	N/A	No
SPACE	TripAdvisor hotels	8,789	1.2M	0.71	20.2%	PrefLib[†]	1,050 gold
Yelp	Restaurants	150K	7M	0.65	36.5%	94.7%	No
Rotten Tomatoes	Film criticism	13,664	1.4M	0.55	2.7%	N/A	9,121 consensus

Table 5: Dataset statistics. *Disagree*: fraction of entities with predominantly negative sentiment ($\text{pos_frac} \leq 0.25$). *Temp. sig.*: fraction with detectable temporal signal. [†]SPACE has no timestamps; via PrefLib 00040 (Mattei and Walsh, 2013), 1,713 overlapping hotels. [‡]Verdict/Pros/Cons.

Since $f > 0$, $\mathcal{R}$ contains a review $r^* \in \mathcal{R}$ expressing positive sentiment about X. Extract span $s^* \subset r^*$ of the form "I found X satisfactory."

The span s^* is: (i) not a contradiction of p – it does not assert the majority condition is false; (ii) a direct positive instance of the property claimed by p; (iii) logically consistent with p being true. Under standard NLI labels, (s^*, p) receives at minimum NEUTRAL from any model without access to the full population count. Since $f > 0$ holds whenever at least one positive review exists, such a span can always be found regardless of the degree of overclaiming. The argument is symmetric for negative-polarity claims by substituting $f^- > 0$ and a negative-sentiment span. □

Corollary F.0.1 (The counting gap)**.** *The ground truth of a population claim is a count* $|\{r \in \mathcal{R} : r \text{ expresses polarity } \pi\}|$. *NLI is a function of (claim, span) pairs. No function of (claim, span) pairs can compute a count over a population without accessing every member and performing summation. NLI-based faithfulness is therefore structurally incapable of verifying population claims, independent of model architecture, training data, or scale.*

This makes Theorem 3.1 precise: it is not that current NLI models are insufficiently powerful – it is that the verification function answers a different question than the one a population claim poses.

Proposition F.1 (Threshold-boundary overclaiming)**.** *Let $\mathcal{C}$ be a finite cue-word set with thresholds $\{\tau_c\}_{c \in \mathcal{C}}$. For any $\epsilon > 0$, there exists a population with* $\text{pos_frac} = \tau_c - \epsilon$ *for which cue c is the closest available descriptor: the next lower threshold cue c' satisfies $\tau_{c'} \ll \tau_c - \epsilon$ and thus understates the actual fraction. A summarizer minimizing description error will select c, overclaiming by ϵ.*

Proof sketch. The threshold set $\{\tau_c\}$ partitions $(0, 1]$ into discrete intervals. The continuous variable pos_frac falls in the interior of one interval with probability 1 under any non-atomic distribution over $[0, 1]$. Summarizers choosing from a finite vocabulary always face a non-zero overclaim for the best-fit cue; the overclaim vanishes only when pos_frac falls exactly on a threshold boundary, an event of measure zero. □

G Annotation Guidelines and Annotator Details

Task Overview

Annotators were shown sentences extracted from AmaSum human-written summaries (Pros, Cons, and Verdict sections) that contained consensus cue words (e.g., "most," "many," "some," "almost all"). For each claim, the annotation interface displayed:

- The full claim sentence and the summary section it came from (Pros / Cons / Verdict)
- The product's pos_frac value – the fraction of star ratings ≥ 4 out of all reviews for that product – shown as a decimal rounded to two places (e.g., 0.83)
- A reference card of the popularity-class thresholds (Table 1)

Annotators were instructed to assess whether the claim makes a quantitative assertion about the reviewer population and, if so, whether that assertion is justified by the product's pos_frac.

Label Definitions

NOTAPOPULATIONCLAIM – Not a Population Claim. The sentence does not assert a claim about a fraction of reviewers. Apply this label when: (a) the sentence describes a product property rather than a reviewer opinion ("the battery lasts 12 hours"); (b) a cue word modifies a product attribute rather than a reviewer fraction ("overall design is clean"); or (c) a cue word describes a use context rather than a population ("works well in many areas").

JUSTIFIED. The sentence is a population claim – it asserts that a fraction of reviewers hold a view – the claimed polarity direction is consistent with the dominant sentiment ($\text{pos_frac} \geq 0.50$ for positive claims; $\text{pos_frac} < 0.50$ for negative claims), *and* the population fraction meets the popularity-class threshold for the cue word used. Formally, for a positive-polarity claim with cue c: $\text{pos_frac} \geq \tau_c$.

OVERCLAIMED. The sentence is a population claim, the direction is consistent with the dominant sentiment, but the implied fraction does *not* meet the popularity-class threshold. The claim is directionally right but quantitatively overstated. Formally: $\text{pos_frac} < \tau_c$ for a positive-polarity claim with cue c.

WRONGDIRECTION. The sentence is a population claim but the claimed direction is inconsistent with the dominant sentiment (e.g., "most reviewers loved the product" when $\text{pos_frac} < 0.50$).

Decision Procedure

Annotators applied the following steps in order:

1. **Is this a population claim?**
   Does the sentence use consensus language asserting a fraction of reviewers hold a view? If *no* → NOTAPOPULATIONCLAIM.
2. **Is the claimed direction consistent with the dominant sentiment?**
   Positive-polarity claim: is $\text{pos_frac} \geq 0.50$? Negative-polarity claim: is $\text{pos_frac} < 0.50$? If *no* → WRONGDIRECTION.
3. **Does the population fraction meet the cue-word threshold?**
   Look up the cue word's popularity class from Table 1. For positive-polarity: is $\text{pos_frac} \geq \tau_c$? If *yes* → JUSTIFIED; if *no* → OVERCLAIMED. For negative-polarity: is $(1-\text{pos_frac}) \geq \tau_c$? Same rule.

Worked Examples Shown to Annotators

Claim	pos_frac	Label
"Most reviewers praise the build quality."	0.83	JUST. (0.83≥0.55)
"Most reviewers praise the build quality."	0.45	OC (0.45<0.55)
"Almost all reviewers recommend this."	0.71	OC (0.71<0.85)
"Almost no reviewers recommend this."	0.83	WD
"Many reviewers appreciate fast charging."	0.62	JUST. (0.62≥0.35)
"The battery life is impressive overall."	0.79	NAPC (product)
"Works well in many areas."	0.75	NAPC (context)

JUST. = JUSTIFIED; OC = OVERCLAIMED; WD = WRONGDIRECTION; NAPC = NOTAPOPULATIONCLAIM.

Edge Cases Clarified During IAA Pilot

Negative existentials on positive products. Cons-section claims such as "some reviewers found the build quality lacking" on a product with $\text{pos_frac} = 0.85$ should be labeled OVERCLAIMED, not WRONGDIRECTION. The direction is partially correct (some negative-sentiment reviewers do exist), but the Cons-section framing implies a higher fraction than the existential threshold warrants. Annotators were instructed to use OVERCLAIMED whenever direction is at least partially present but the consensus quantifier overstates the fraction.

Cue-word scope. The test for NOTAPOPULATIONCLAIM is: *does this sentence assert something about the fraction of reviewers, or about the product itself?* "Overall design is clean" → NOTAPOPULATIONCLAIM ("overall" modifies "de-

sign," not a reviewer fraction). "Overall, reviewers were satisfied" → population claim ("overall" modifies the reviewer consensus).

Annotator Details

Recruitment and background. Annotation was conducted through a managed third-party labeling vendor using Labelbox as the annotation platform. All annotators are fluent English speakers employed full-time by the vendor. Annotators had no prior exposure to the distributional verification methodology or to any of the specific claims before the IAA phase.

Compensation. Annotators are full-time employees of the labeling vendor with healthcare benefits and guaranteed living wages in their city and country of residence, in accordance with our organization's ethical standards for third-party labeling vendors. The vendor also provides career advancement and educational opportunities, and annotators are not required to moderate graphically violent content.

Ethics and consent. All claims annotated were drawn from AmaSum (Bražinskas et al., 2021), a publicly available dataset of Amazon product review summaries; annotation consisted solely of labeling existing publicly available text and no new data was collected from human subjects. No personally identifiable information was present in the annotation interface. This study uses pre-existing publicly available data and is exempt from IRB review.