When More Extractive Summaries Receive Higher Automated Factuality Scores: Evidence from Distilled Scientific Summarization

Joseph Moretto
Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA, USA
jam5cq@virginia.edu

Peter A. Beling
School of Data Science
University of Virginia
Charlottesville, VA, USA
beling@virginia.edu

Tyler Cody
School of Data Science
University of Virginia
Charlottesville, VA, USA
tyler@tcody.net

Donald E. Brown
Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA, USA
deb@virginia.edu

Abstract

Scientific papers are long and technically dense, and frontier large language models can summarize them effectively, but their cost and API dependencies limit large-scale use. We distill scientific summarization from Claude Opus 4.6 into LLaMA 3.1 8B with LoRA and LLaMA 3.3 70B with QLoRA using 1,500 arXiv papers. On a 150-paper test set, the fine-tuned students improve BERTScore F1 against paper abstracts over zero-shot base models ($p < 0.001$; Cohen's d up to 1.40 on the aggregate test set) and closely match the teacher's abstractiveness profile, including novel trigram rate ($\sim$77% vs. $\sim$64% for base) and extractive density ($\sim$2.5 vs. $\sim$4.6). These gains are not consistently reflected in BLEU, ROUGE-2, or LongDocFACTScore, which often favor the more extractive base outputs. Across 750 summaries, LongDocFACTScore is positively associated with extractive density (Pearson $r = 0.52$), and the association remains positive when analyzed separately by system and separately by domain. An earlier 8B decoding run shows the same directional ordering with SummaC. Because we lack human factuality annotations, we do not determine which system is most faithful. Instead, we show that automated factuality rankings can be sensitive to extractiveness and should be reported alongside semantic-similarity and explicit extractiveness measures.

1 Introduction

Scientific documents are long and technical, requiring faithfulness to specific methodological details that makes them considerably harder to summarize than news articles. Frontier large language models (such as Claude Opus) can produce high-quality scientific summaries, but their computational cost and API dependencies make them impractical for literature review tools or research database indexing at scale.

Knowledge distillation, using a capable teacher to generate training data for a smaller student, has become a recognized approach to this problem (Das and Rudrapal, 2025; Zhang et al., 2026). Yet prior distillation work for summarization has relied on weak teachers (GPT-2 in Jung et al., 2024; mBART and mT5 in Lee et al., 2025), full fine-tuning rather than parameter-efficient methods (Liu et al., 2024b; Agarwal et al., 2024), and news or medical domains rather than scientific text. LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) enable efficient adaptation of large models, and KD-LoRA (Azimi et al., 2024) combines the two, but only for classification with encoder-only models. Bao et al. (2024) target scientific papers with structure-aware fine-tuning but without frontier LLM distillation, and Chain of Density (Adams et al., 2023) improves summary density via prompting without training student models.

Evaluating a summarization system requires a working definition of summary quality. SummEval (Fabbri et al., 2021) decomposes summary quality into four dimensions: coherence, consistency, fluency, and relevance. A good scientific summary should be fluent and coherent prose, factually faithful to the source paper, and should select for the main contribution, methodology, key results, and significance rather than peripheral details. Hu-

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1092–1104
October 17-21, 2026 ©2026 Association for Computational Linguistics

man reference summaries are not a clean ground truth. Zhang et al. (2024) find that instruction-tuned LLMs surpass the heuristically derived reference summaries of CNN/DailyMail and XSum on faithfulness, coherence, and relevance, while rating on par with summaries newly commissioned from freelance writers, and Wan et al. (2024) confirm this through atomic content unit decomposition, showing that reference summaries on standard benchmarks contain unsupported claims at rates comparable to or higher than instruction-tuned model outputs. Paper abstracts, which we use as one of our references, face a related problem in that they are written to attract readers and signal novelty rather than to inform a reader who wants to know what a paper does. Even direct human annotation of factual consistency is unreliable. Pagnoni et al. (2021) report only moderate agreement on fine-grained factual error categories (Fleiss $\kappa = 0.39$, against $\kappa = 0.58$ for the binary factuality judgment), and Mujahid et al. (2026) note that large-scale human annotations for long-document summaries remain scarce, leaving automated metrics largely unvalidated in our setting.

Given these constraints, we take the position that a good scientific summary should do the following. First, cover the four canonical dimensions of contribution, methodology, results, and significance. Second, preserve technical claims accurately. Third, be written in fluent abstractive prose rather than verbatim copying. Extractive summaries preserve source wording exactly and avoid one class of factual error, but at the cost of the synthesis and reframing that distinguish summarization from quotation. We use Claude Opus 4.6 and its abstractiveness profile as a target for distillation, not as ground-truth quality. Fine-tuning transfers this target style to the students, and the resulting outputs receive different rankings under different metrics. Because we lack human factuality annotations, we do not claim that teacher-like outputs are more faithful than the base-model outputs.

To our knowledge, few studies combine frontier-teacher distillation, parameter-efficient adaptation, and multi-domain scientific-document summarization. We make three contributions. First, we show that distillation transfers an abstractive target style to open-weight students and increases their semantic similarity to both held-out teacher summaries and paper abstracts, bringing abstract-referenced student BERTScore to within $\sim$0.01 of the teacher on a 150-paper test set. Second, we document a systematic ranking disagreement: against paper abstracts, fine-tuning improves BERTScore and several reference-based metrics but not BLEU or ROUGE-2, and it lowers LongDocFACTScore. Third, we quantify the association between LongDocFACTScore and extractive density, which the metric's authors flag qualitatively (Bishop et al., 2024), and show that it holds across systems, within systems, and across domains. Combined with controlled perturbation evidence from Mujahid et al. (2026), this indicates that automated factuality scores can be sensitive to source-form overlap, complicating comparisons between systems with different abstractiveness profiles. We therefore position the paper primarily as an evaluation analysis, with distillation serving as the setting that exposes the disagreement.

The rest of this paper is organized as follows. Section 2 describes the data collection pipeline, teacher summary generation with Claude Opus 4.6, and the LoRA/QLoRA fine-tuning setup. Section 3 presents our metrics, significance testing, and evaluation design. Section 4 reports results against both paper abstracts and held-out teacher summaries, the abstractiveness analysis, and the association between extractiveness and factuality scores. Section 5 interprets these findings and their implications for evaluating abstractive summarization systems, and Section 6 concludes. Appendices provide hyperparameter justifications, the full prompts, qualitative summary examples, and complete per-domain results.

2 Methodology

2.1 Data Collection

We collect 1,500 papers from arXiv (500 each from CS, physics, and mathematics) by querying the arXiv API with the general category wildcard for each domain (`cs.*`, `physics.*`, `math.*`), sorting results by submission date in descending order and walking backward from the most recent submissions until 500 papers per domain pass the extraction criterion below. Cross-listed papers are deduplicated across categories with a global seen-identifier set. Beyond requiring downloadable LaTeX source and successful section extraction (below), no content-based inclusion filters are applied; the source-availability requirement skews the corpus toward papers whose authors uploaded LaTeX. Collection was performed at the beginning of May 2026; the most recent submissions returned by the

API at crawl time dated to late April, and all 1,500 retained papers were first submitted in April 2026. A custom extraction pipeline extracts four canonical sections (introduction, methods, results, conclusion) using flexible heading matchers for discipline-specific conventions (e.g., "Theoretical Framework" for physics, "Main Theorem" for mathematics), removes display equations and figure/table environments while retaining captions and inline math, and performs brace-depth-aware LaTeX command cleaning. The paper abstract is stored separately as a reference summary and is never part of the model input, for either the teacher or the students. Papers are retained if at least two canonical sections are extracted. This corpus is small relative to the largest prior distillation efforts; Jung et al. (2024), for instance, distil a 1.5M-pair sentence-summarization corpus. Frontier-teacher generation over full paper texts is far more expensive per example than the sentence-level setting that method targets, which bounds the corpus size we can afford. Small-scale LLM-reference training is nonetheless established practice: Liu et al. (2024b) fine-tune with as few as 100–1,000 references from a reference LLM, on top of a ∼10K-summary warm-up.

Training data contamination. Public arXiv papers may in principle appear in the pretraining data of the teacher or the students. Our collection window largely rules this out. arXiv identifiers encode the month of first submission, and every identifier in our corpus carries the `2604` prefix: the first version of every paper was submitted in April 2026, so no earlier arXiv version of any paper exists. This postdates the December 2023 pretraining cutoff of both LLaMA students (Grattafiori et al., 2024) and the August 2025 training-data cutoff of the teacher,[1] so neither the papers nor their abstracts can have appeared in any model's pretraining corpus. Any residual risk also runs in a conservative direction: contamination would inflate the base models' zero-shot scores, shrinking rather than exaggerating the reported fine-tuning gains, and the extractiveness analysis in Section 4.3 concerns the behavior of the scoring metrics rather than of the models, so it is unaffected.

[1]Anthropic documents Claude Opus 4.6 with a training-data cutoff of August 2025 and a reliable-knowledge cutoff of May 2025; the Claude Opus 4.6 System Card (Anthropic, 2026) states only that the model was trained on "publicly available information from the internet up to May 2025," alongside undated non-public and internally generated data.

2.2 Teacher Summary Generation

We use Claude Opus 4.6 (Anthropic) as the teacher, prompted to generate 150–250 word summaries covering the main contribution, methodology, key results, and significance. The teacher and student prompts convey the same constraints (150–250 words, the four target aspects, third-person present-tense prose, no bullet points): the teacher receives a detailed version and the students a condensed one, both reproduced verbatim in Appendix B. Teacher inputs contain the full extracted text up to ∼25k tokens, with simple tail truncation beyond that; the students' tighter 8,192-token context handling is described in Section 2.3. Summaries are filtered by length (150–300 words, allowing modest overshoot of the requested 150–250 range) and API success status; all 1,500 yield valid outputs. The total generation cost was $30 via the Anthropic batch API. Data is split 80/10/10 into training (∼1,200), validation (∼150), and test (150 papers) with domain stratification, where domain labels group related arXiv archives with their parent domain (e.g., `quant-ph` with physics, `stat` with math).

2.3 Fine-Tuning

We train **LLaMA 3.1 8B Instruct** with LoRA (Hu et al., 2022) in bf16, and **LLaMA 3.3 70B Instruct** with QLoRA (Dettmers et al., 2023) (4-bit NormalFloat). LoRA adapters (rank 16, alpha 32, dropout 0.1) target all projection layers. Training uses 3 epochs, learning rate 1e-4, effective batch size 16, cosine schedule with warmup ratio 0.05, weight decay 0.01, gradient clipping 0.3, and maximum sequence length 8,192 tokens via SFTTrainer (TRL) with gradient checkpointing and completion-only loss masking. The 8B trains in ∼3.7 hours on one NVIDIA A6000 (48GB); the 70B in ∼24 hours on two A6000s. Full justification for each choice is in Appendix A.

The 8B achieves train loss 0.369 / eval loss 0.529; the 70B achieves 0.570 / 0.634. The 8B has a larger train-evaluation gap (0.160 vs. 0.064), consistent with greater overfitting. We do not use the absolute losses to compare model quality across the two parameterizations. Papers exceeding the 8,192-token context window are truncated using a head-tail strategy: 70% of the token budget from the beginning of the extracted input (introduction, methods) and 30% from the end (results, conclusion), with cuts snapped to section boundaries where possible. The instruction header is preserved outside

the truncation budget, and a [...TRUNCATED...] marker is inserted at the cut point between the two segments.

3 Evaluation

Metrics ROUGE-1/2/L F1 (Lin, 2004), BERTScore F1 with DeBERTa-xlarge-MNLI (Zhang et al., 2020), corpus-level BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), LongDocFACTScore (Bishop et al., 2024) for factual consistency, and novel n-gram percentages (See et al., 2017) with extractive fragment coverage/density (Grusky et al., 2018), and vocabulary diversity (type–token ratio). We exclude LLM-as-judge metrics such as G-Eval (Liu et al., 2023). G-Eval correlates well with human judgments on news summarization, but its authors document a bias toward LLM-generated text: G-Eval-4 scores GPT-3.5 summaries above human-written ones even when human judges prefer the latter. Exact metric implementations and settings are documented in the released evaluation code.

Significance We use a paired two-sided sign-flip permutation test with 10,000 iterations and a Wilcoxon signed-rank test, with Cohen's d as an effect-size measure. We report unadjusted p-values. For each model-scale/reference setting, the six metrics form one comparison family; every result reported at $p < 0.001$ remains significant under Bonferroni correction ($\alpha = 0.0083$).

Design All metrics are computed against paper abstracts and held-out teacher summaries on a 150-paper test set (55 CS, 48 Physics, 47 Math). Because abstracts are withheld from all model inputs (Section 2.1), abstract-based scores are not affected by direct reference exposure. The abstracts nevertheless remain imperfect references drawn from the same papers. We compare fine-tuned vs. zero-shot base models at each scale. All four student systems share an identical prompt (Appendix B), so fine-tuned-vs-base comparisons hold prompting constant; the teacher's prompt differs, and teacher scores serve as reference points rather than controlled comparisons. Inference uses the training prompt template with fixed seed (42), temperature 0.7, top-p 0.9, repetition penalty 1.1, max 512 new tokens. The 8B processes one paper in ∼20 seconds; the 70B in ∼2 minutes.

Metric	FT 70B	B 70B	FT 8B	B 8B	Teach
ROUGE-1	.469	.459	.466	.428	.489
ROUGE-2	.167	.171	.164	.161	.186
ROUGE-L	.240	.236	.237	.222	.258
BERT F1	.646	.623	.644	.605	.654
BLEU	.069	.081	.066	.063	.079
METEOR	.342	.320	.336	.334	.357

Table 1: Results vs. abstracts ($n = 150$). BLEU = corpus. B = base (zero-shot). FT = fine-tuned.

Metric	FT 70B	B 70B	FT 8B	B 8B
ROUGE-1	.663	.559	.636	.541
ROUGE-2	.347	.241	.311	.229
ROUGE-L	.432	.311	.390	.300
BERT F1	.760	.683	.742	.668
BLEU	.214	.124	.182	.108
METEOR	.500	.364	.465	.395

Table 2: Results vs. teacher summaries ($n = 150$). BLEU = corpus.

4 Results

4.1 Reference-Based Metrics

Table 1 presents results against abstracts. Both fine-tuned models numerically outperform their base counterparts on ROUGE-1, ROUGE-L, BERTScore, and METEOR (significance in Table 5). The fine-tuned 70B achieves the highest student scores (ROUGE-L 0.240, BERTScore F1 0.646), closely approaching the teacher (0.258, 0.654). The 8B performs comparably (0.237, 0.644). On BLEU the pattern splits by scale: the fine-tuned 8B slightly exceeds its base (0.066 vs. 0.063), while the fine-tuned 70B falls significantly below its base ($p < 0.01$, $d = -0.26$). ROUGE-2 shows the same scale split at smaller, non-significant margins.

Table 2 shows distillation fidelity against teacher summaries. The 70B achieves ROUGE-L 0.432 vs. 0.311 for base ($\Delta = 0.121$); the 8B achieves 0.390 vs. 0.300 ($\Delta = 0.090$).

4.2 Extractiveness and Abstractiveness

Table 3 shows the abstractiveness comparison. Both fine-tuned models closely match the teacher's higher-order n-gram novelty and extractive density, while base models are markedly more extractive. This holds whether the base model produces longer output (8B base, 289 words) or shorter (70B base, 198 words).

The extractive-density contrast is the clearest signal: fine-tuned models average 2.46–2.62, near

Metric	FT 70B	B 70B	FT 8B	B 8B	Teach
Novel 3g	78.2%	66.3%	76.5%	62.8%	75.6%
Novel 4g	88.0%	76.6%	86.4%	72.9%	86.1%
Ext. density	2.46	4.06	2.62	5.16	2.74
Vocab div	0.68	0.60	0.68	0.52	0.66

Table 3: Abstractiveness analysis (full table in Appendix D). Novel ng = fraction not in source. Ext. density = sum of squared extractive-fragment lengths divided by summary length (Grusky et al., 2018). Vocab div = type–token ratio, reported descriptively because it is sensitive to summary length. All FT vs. Base differences $p < 0.001$.

Scope	FT 70B	B 70B	FT 8B	B 8B	Teach
Overall	−4.39	−3.76	−4.33	−4.07	−4.18
CS	−4.41	−3.76	−4.34	−4.07	−4.21
Physics	−4.30	−3.66	−4.29	−3.97	−4.13
Math	−4.46	−3.86	−4.35	−4.18	−4.22

Table 4: LongDocFACTScore (BARTScore log-prob; higher = greater consistency according to the metric).

the teacher's 2.74, whereas base models average 4.06–5.16. Density is defined as the sum of squared extractive-fragment lengths divided by summary length, so it gives disproportionate weight to long source-overlapping fragments; it is not itself an average span length. The higher base-model values therefore indicate substantially longer source-overlapping spans. Summary length does not explain the pattern: the 70B base is shorter than the fine-tuned 70B (198 vs. 217 words) but remains markedly more extractive.

This helps explain the BLEU pattern. Fine-tuned models generate more novel phrasings, reducing exact n-gram matches, and BLEU, which relies entirely on precision of n-gram overlap, penalizes this. ROUGE-2 shows the same pattern. BERTScore, which operates in embedding space, is the only metric that favors the fine-tuned models at $p < 0.001$ at both scales (Table 5); ROUGE-1 also improves at both scales, but only marginally at 70B. (Tables 1 and 2 report corpus-level BLEU, while the paired tests in Tables 5 and 6 aggregate per-summary BLEU, so the BLEU columns differ across each table pair; against abstracts, the significant 70B effect has the same sign under both and the 8B differences are negligible in either direction.)

4.3 Factuality Scores and Extractiveness

Table 4 shows LongDocFACTScore results. For both model sizes, the ordering is base $>$ teacher $>$ fine-tuned, the opposite of what reference-based quality metrics suggest.

Extractiveness offers a plausible explanation for this ordering. LongDocFACTScore uses BARTScore to compute the conditional log-probability of each summary sentence given retrieved source passages. Because source-like wording can receive high conditional probability, the metric may favor extractive realizations over abstractive paraphrases. Consistent with this mechanism, the more extractive base outputs score higher than the fine-tuned outputs. Bishop et al. (2024) anticipate this qualitatively, noting in their limitations that LongDocFACTScore and other state-of-the-art factuality metrics "favour extractive summaries" and recommending that the metric be used in conjunction with other metrics. Our contribution is to quantify that association and to show that it holds across systems, within systems, and across domains. Our data do not determine whether this gap reflects surface-form sensitivity, genuine faithfulness differences, or both.

An earlier decoding run of the two 8B systems provides a limited robustness check with SummaC (Laban et al., 2022). In that run, the fine-tuned model scored lower than the base model (-0.31 vs. -0.15) in all three domains. Because these were different generations and the comparison did not include the 70B systems or the teacher, we treat the SummaC result as directional supporting evidence rather than as a second full replication. Across all 750 system–paper outputs, extractive density and LongDocFACTScore have a pooled descriptive correlation of Pearson $r = +0.52$ and Spearman $\rho = +0.60$ (Figure 1). Because each paper contributes five summaries, we do not interpret the pooled p-value as an independent-observation significance test. The pooled partial correlation controlling for summary length is $r = +0.56$. More importantly, within each system — where each paper contributes one observation — the association remains positive, with Pearson r ranging from $+0.48$ to $+0.63$ (all $p < 10^{-9}$). The pooled descriptive association is also positive in each scientific domain: CS $r = +0.58$, physics $r = +0.48$, and mathematics $r = +0.46$. These analyses show that the pooled trend is not solely explained by differences in system means, but they do not rule out paper difficulty, retrieval quality, or genuine faithfulness differences as shared causes. Mujahid et al. (2026) provide complementary evidence from the perturbation side: rather than varying a summa-

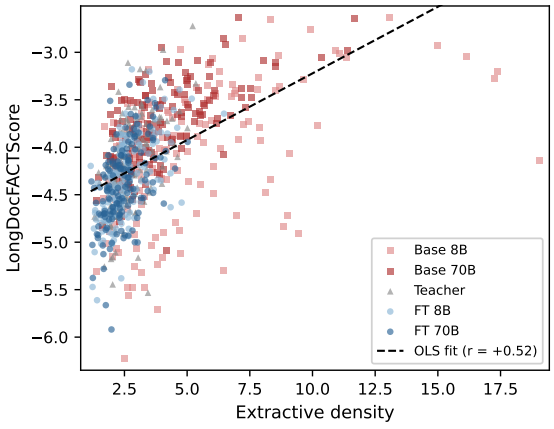


Figure 1: Per-summary extractive density vs. LongDocFACTScore across all five systems and 150 test papers ($n = 750$). Each point is one summary scored against its source paper. The dashed line is a descriptive pooled OLS fit. Because each paper contributes one point per system, inferential results are reported within system in Section 4.3; the association remains positive in every system.

rization system, they apply meaning-preserving edits (paraphrasing, simplification, synonym replacement) to human-written reference summaries and find substantial score shifts in BARTScore, AlignScore, and SummaC variants on legal text, though the same edits leave those metrics nearly flat on multi-document scientific text. Together, our observational result and their controlled perturbations indicate that these metrics are sensitive to surface form in addition to factual content. The rankings these metrics produce should therefore not be interpreted as pure measures of faithfulness when systems differ substantially in extractiveness. Our study does not establish which system is factually superior.

4.4 Statistical Significance

Table 5 confirms the results against abstracts. BERTScore is significant at $p < 0.001$ for both model sizes with large effect sizes ($d = 1.40$ for 8B, $d = 0.85$ for 70B). The 70B improvements are smaller because the 70B base is already stronger, narrowing the margin. Against teacher summaries (Table 6), all 12 comparisons are significant at $p < 0.001$ with $d > 1.0$.

4.5 Per-Domain and Length Analysis

BERTScore F1 is significant at $p < 0.001$ in all six domain-model combinations. For the 8B, mathematics shows the strongest gains ($d = 1.72$); for the 70B, CS is strongest ($d = 0.96$). Per-domain significance and score tables are in Appendix E.

Both fine-tuned models match the teacher's length profile: the 8B averages 212.6 words (std 15.7), the 70B 217.4 (std 15.7), and the teacher 221.9 (std 16.8).[2] Of 150 summaries, 149 (8B) and 146 (70B) fall within the 150–250 word target, compared to 35/150 for the 8B base and 147/150 for the 70B base. One caveat: 14/150 (9%) of 8B-base generations end without sentence-final punctuation, consistent with reaching the 512-token generation cap; no other system shows this signature. Truncation likely depresses the 8B base's reference-based scores, so a small part of the 8B fine-tuning gap on those metrics may reflect the cap, but the factuality ordering and density gap are unchanged when these 14 papers are excluded (LongDocFACTScore -4.00 vs. -4.33 for the fine-tuned model; density 5.24). The 70B base follows length instructions without fine-tuning but does not adopt the teacher's abstractive style (Section 4.2), showing that format compliance and summarization style are separable.

	8B			70B		
Metric	Δ	d		Δ	d	
ROUGE-1	+.038	0.69	***	+.010	0.18	*
ROUGE-2	+.003	0.07		−.004	−0.09	
ROUGE-L	+.016	0.35	***	+.004	0.09	
BERT F1	+.038	1.40	***	+.023	0.85	***
BLEU	−.001	−0.03		−.009	−0.26	**
METEOR	+.002	0.04		+.022	0.40	***

Table 5: FT vs. Base against **abstracts** ($n = 150$). Paired sign-flip permutation test. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Δ and d are computed from unrounded scores; BLEU here is per-summary (corpus-level in Table 1).

	8B			70B		
Metric	Δ	d		Δ	d	
ROUGE-1	+.095	1.69	***	+.104	1.95	***
ROUGE-2	+.082	1.40	***	+.106	1.69	***
ROUGE-L	+.090	1.45	***	+.121	1.85	***
BERT F1	+.074	2.37	***	+.077	2.55	***
BLEU	+.067	1.20	***	+.090	1.45	***
METEOR	+.070	1.20	***	+.137	2.02	***

Table 6: FT vs. Base against **teacher summaries** ($n = 150$). All $p < 0.001$. Δ and d are computed from unrounded per-summary scores.

[2] Whitespace-delimited word counts of the generation texts, matching the summary statistics in our released evaluation results. The `summary_length` field in the released abstractiveness data uses the evaluation pipeline's tokenizer and runs ∼5–7% higher; all relative comparisons are unaffected.

4.6 Model Size: 70B vs. 8B

Against teacher summaries, the 70B shows clear advantages: ROUGE-L 0.432 vs. 0.390 for the 8B. Against abstracts, the gap narrows: ROUGE-L 0.240 vs. 0.237, BERTScore 0.646 vs. 0.644. The 8B achieves 98.8% of the 70B's ROUGE-L and 99.7% of its BERTScore against abstracts, but only 90.3% of the 70B's ROUGE-L against teacher summaries. The 8B requires one GPU and 3.7 hours of training vs. two GPUs and 24 hours, making it the preferred choice when near-teacher fidelity is not required.

5 Discussion

5.1 What distillation transfers

We expected fine-tuning to improve metric scores, but it also changed the model's writing behavior more thoroughly than anticipated. The 70B base already follows length instructions well, producing summaries in the target word range without any fine-tuning, yet its extractive density remains substantially higher than the teacher's. Fine-tuning closes that gap: novel n-gram rates and extractive density shift to the teacher's profile, slightly overshooting it (density 2.46–2.62 vs. the teacher's 2.74). Type–token ratio moves in the same direction but is interpreted descriptively because it is sensitive to summary length. The qualitative examples in Appendix C make this visible. Fine-tuned models produce flowing prose, whereas base models produce templates with headers and formatting that the prompt explicitly instructs against ("flowing prose without bullet points").

5.2 Metric Sensitivity to Extractiveness

Metric choice reverses the apparent ranking of the systems. BERTScore favors fine-tuning at both model scales and in all domains, whereas LongDocFACTScore favors the more extractive base outputs at both scales; BLEU and ROUGE-2 show mixed or null effects against abstracts. The earlier 8B SummaC run shows the same direction as LongDocFACTScore but is only a directional robustness check. These results demonstrate that semantic reference similarity, lexical overlap, extractiveness, and source-based factuality scores capture different properties. They do not establish that BERTScore identifies the most faithful system. We therefore recommend reporting these dimensions separately, with extractiveness accompanying any factuality-based system ranking.

Not all copying is problematic: scientific summaries must preserve technical terms, and unnecessary paraphrasing can introduce errors. Coverage is similar across the four student systems (80–85%; Appendix D), while density is markedly higher for the base models. Because density gives quadratic weight to long fragments, this combination indicates that the difference is driven by longer clause-level source overlaps rather than only by isolated technical terms. Appendix C illustrates this contrast. Separating term-level from clause-level copying remains useful future work.

5.3 What we cannot conclude

Two gaps in our setup keep this evidence about metric attribution rather than ground-truth factuality: we have no human factuality annotations, and we do not compare against a student fine-tuned on human abstracts rather than teacher summaries. Base models copying verbatim spans receive high LongDocFACTScore because verbatim text has high conditional probability under the scoring model, but verbatim copying does not guarantee correctness either, since a copied claim can still be misleadingly truncated, attributed to the wrong subject, or stripped of qualifying context. We cannot say whether the more abstractive fine-tuned summaries contain more, fewer, or roughly the same number of factual errors as the base outputs. Closing these gaps requires either careful human annotation, an abstract-supervised baseline, or a controlled experiment that perturbs extractiveness while holding factual content constant.

5.4 Efficiency

For $30 in batch API costs, we generated 1,500 teacher summaries. The fine-tuned 8B then processes a paper in ∼20 seconds on a single A6000 with no API key needed; with 4-bit quantization its memory footprint fits a 16GB consumer GPU, though we did not benchmark consumer hardware directly. Summarizing 100,000 papers via the Claude API would run roughly $2,000 at current batch rates for Claude Opus 4.6 ($5/$25 per million input/output tokens standard, halved under the 50% batch discount), given our average token counts. These figures exclude one-time GPU access costs for training (∼3.7 hours on a single A6000 for the 8B), and API pricing is subject to change. Whether that tradeoff is worth it depends on the application. For literature indexing pipelines, where the cost difference compounds across hundreds of thousands

of papers, the distilled model is a competitive option. A locally deployable model is necessary in settings where external API calls are not permitted, including offline or air-gapped environments and applications with security or privacy constraints.

6 Conclusion

Our experiments show that LongDocFACTScore rankings are strongly associated with extractiveness. Base outputs with longer source-overlapping spans outscore both fine-tuned students and the teacher, and extractive density is positively associated with score in the pooled data, when analyzed separately by system, and when analyzed separately by domain. An earlier 8B decoding run shows the same directional ordering with SummaC. Without human factuality labels, we cannot determine whether these differences arise from surface-form sensitivity, genuine faithfulness differences, or both. We therefore recommend reporting extractiveness alongside factuality scores whenever compared systems differ substantially in copying behavior.

As a secondary contribution, parameter-efficient fine-tuning transfers the teacher's abstractive target style and increases semantic similarity to both teacher summaries and paper abstracts, bringing abstract-referenced student BERTScore to within $\sim$0.01 of the teacher. This demonstrates efficient style transfer but does not by itself establish superior factuality.

Future work should compare teacher-supervised and abstract-supervised students and evaluate factuality metrics under content-preserving changes in surface form. Until then, rankings based on a single automated factuality metric should be interpreted cautiously.

Limitations

The most important missing baseline is fine-tuning on human abstracts rather than teacher summaries, which would tell us how much of the improvement comes from the teacher's output specifically versus fine-tuning on any reasonable target. We leave this for future work.[3]

[3]Few-shot prompting is impractical here: each paper consumes 3,000–8,000 tokens, so two in-context examples plus the query paper would total roughly 9,000–25,000 tokens, exceeding the 8B's 8,192-token context. Even for larger context windows, Liu et al. (2024a) and Levy et al. (2024) show performance degrades well below the technical maximum, making few-shot results hard to interpret.

We have no human evaluation, so we restrict our claims to metric behavior and do not infer which system contains fewer factual errors. A future study should assess faithfulness and informativeness with explicit guidelines, multiple annotators, and agreement reporting.

LongDocFACTScore is our only fully evaluated factuality metric. An earlier decoding run of the 8B systems showed the same directional ordering with SummaC (Section 4.3), but we could not run QuestEval or AlignScore due to dependency conflicts with our evaluation environment. The claim would be stronger with more metrics. Similarly, our exclusion of LLM-as-judge evaluation rests on concerns about self-preference bias; a judge from a different model family would mitigate this and could assess dimensions such as coverage and coherence, and remains a reasonable extension.

Smaller issues: our conclusions rest on a 150-paper test set spanning three STEM domains, which supports the paired comparisons we report but not broad generalization to other domains or document types. Our preprocessing removes display equations and figure/table environments (retaining captions), an untested choice that may discard content central to mathematics and physics papers. Both students are LLaMA-family models and we distill from a single frontier teacher; base models from other families may have different default copying behavior and other teachers may induce different styles, so the distillation findings may not transfer, although the metric-level association analysis does not depend on either choice. Stochastic decoding introduces run-to-run variation (mitigated by fixed seeds), and we apply no multiple-comparison correction, though all primary results ($p < 0.001$) survive Bonferroni. Training data contamination is addressed in Section 2.1: the April 2026 collection window postdates all models' training cutoffs.

Ethics Statement

This work uses publicly available arXiv papers, retrieved and used for research purposes as permitted by arXiv's Terms of Use. Teacher summaries are generated by a commercial LLM (Claude Opus 4.6); the distilled models may inherit biases present in the teacher's training data. The fine-tuned models are intended for research use in scientific literature tools and should not substitute expert reading of primary sources.

Supplementary Materials Availability Statement: The source documents are publicly available arXiv papers. We release their arXiv IDs, which arXiv makes available as descriptive metadata under a CC0 1.0 Public Domain Dedication; we redistribute the extracted body text for the 692 of 1,500 papers whose authors chose a license permitting it (653 CC BY 4.0, 26 CC BY-SA 4.0, 13 CC0 1.0), each under its own license, and release metadata only for the remaining 808, since arXiv's default non-exclusive license does not grant third parties a redistribution right. The teacher-generated training, validation, and test summaries, the held-out test split (arXiv IDs) and the split configuration, the training and evaluation pipeline code, and the per-summary metric scores underlying all tables and figures are available at `https://github.com/Joseph-Moretto/Summary_Fine_Tuning`; the code is released under an MIT license and the teacher-generated summaries under CC BY 4.0. The section-extraction script is not included, but Section 2.1 describes the procedure it implements. The LoRA/QLoRA adapter configurations and tokenizers are included there; the adapter weight files themselves exceed the repository host's file-size limit and are available from the authors on request. The base LLaMA models are available from Meta under the LLaMA community license, and the teacher model (Claude Opus 4.6) is a commercial API model that we cannot redistribute; the released prompts (Appendix B) and batch API settings allow regeneration of comparable teacher outputs.

Acknowledgments

Funding provided in part by the National Security Data & Policy Institute, Contracting Activity #2024-24070100001. Compute was provided through a sponsor-funded allocation on UVA's Rivanna high-performance computing cluster. We used Claude Code to organize the released code repository and to draft its README documentation. We thank the anonymous INLG 2026 reviewers and the area chair for their comments.

References

Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Anthropic. 2026. Claude Opus 4.6 system card. `https://www-cdn.anthropic.com/c788cbc0a3da9135112f97cdf6dcd06f2c16cee2.pdf`. Training data cutoff: August 2025; reliable knowledge cutoff: May 2025.

Rambod Azimi, Rishav Rishav, Marek Teichmann, and Samira Ebrahimi Kahou. 2024. KD-LoRA: A hybrid approach to efficient fine-tuning with LoRA and knowledge distillation. In *Proceedings of the 4th NeurIPS Efficient Natural Language and Speech Processing Workshop*, volume 262 of *Proceedings of Machine Learning Research*, pages 73–80.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tong Bao, Heng Zhang, and Chengzhi Zhang. 2024. Enhancing abstractive summarization of scientific papers using structure information. *Expert Systems with Applications*, 258:125529.

Jennifer A. Bishop, Sophia Ananiadou, and Qianqian Xie. 2024. LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10777–10789, Torino, Italia. ELRA and ICCL.

Shyamashri Das and Dwijen Rudrapal. 2025. Advancing text summarization with knowledge distillation: Techniques, datasets, and challenges. In *Proceedings of the IEEE International Conference on Electrical, Electronics, and Computer Engineering (IC2E3 2025)*. IEEE.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Aaron Grattafiori et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR 2022)*.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2024. Impossible distillation for paraphrasing and summarization: How to make high-quality lemonade out of small, low-quality model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4439–4454, Mexico City, Mexico. 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.

Chanseo Lee, Sonu Kumar, Kimon A. Vogt, Muhammad Munshi, Panindhra Tallapudi, Antonia Vogt, Hamzeh Awad, and Wasim Khan. 2025. Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. *Scientific Reports*, 15(1):27619.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024b. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8647–8664, Mexico City, Mexico. Association for Computational Linguistics.

Zain Muhammad Mujahid, Dustin Wright, and Isabelle Augenstein. 2026. Stress testing factual consistency metrics for long-document summarization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31914–31933, San Diego, California, United States. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. ACUEval: Fine-grained hallucination evaluation and correction for abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10036–10056, Bangkok, Thailand. Association for Computational Linguistics.

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

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2026. A comprehensive survey on automatic text summarization with exploration of LLM-based methods. *Neurocomputing*, 663:131928.

A Hyperparameter Justification

This appendix provides the rationale for each hyperparameter choice in Section 2.3. Choices that are widely standard are noted as such; choices that deviate from defaults are justified explicitly.

LoRA rank 16, alpha 32. These are the defaults in Hugging Face TRL's `ModelConfig`, the argument dataclass we use with TRL's trainer. Setting alpha to twice the rank is a widely used convention rather than a PEFT library default, which is $r = 8$, $\alpha = 8$.

Target modules: all projection layers. The QLoRA paper (Dettmers et al., 2023) showed that targeting all linear layers matters more than increasing rank for matching full fine-tuning performance. We adapt all q/k/v/o projections and the gate/up/down projections in the MLP blocks. Limiting adapters to attention only would miss the feed-forward layers, where stylistic behavior is partly encoded.

LoRA dropout 0.1. Added as regularization given the small training set (1,200 examples). The PEFT default of 0.0 would leave the adapters without dropout regularization.

Epochs 3. Conventional for LoRA on small datasets. The 8B's train-eval gap (0.160) at three epochs is consistent with mild overfitting but not catastrophic; preliminary runs at higher epoch counts widened this gap without improving eval loss.

Learning rate 1e-4, cosine schedule, warmup ratio 0.05. 1e-4 is the QLoRA paper's recommendation for larger models. We use the same rate for both the 8B and 70B to avoid introducing a scale-vs-LR confound. Cosine annealing decays the rate smoothly rather than dropping at arbitrary steps, and 5% warmup stabilizes the first few updates.

Weight decay 0.01. Standard AdamW regularization, providing a mild L2 penalty on the trainable adapter parameters; the frozen base-model weights do not update.

Gradient clipping 0.3. More aggressive than the typical 1.0 default. We use the QLoRA paper's recommended value to stabilize the higher-precision adapter gradients while training over a 4-bit-quantized frozen base model.

Effective batch size 16. Small enough for frequent gradient updates on 1,200 examples (∼225 optimizer steps total across 3 epochs), large enough that gradient estimates are not dominated by noise. The 8B uses per-device batch size 4 with gradient accumulation 4. For the 70B, the quantized model is sharded across two GPUs; microbatch size 1 with gradient accumulation 16 yields an effective batch size of 16.

Maximum sequence length 8,192. The practical ceiling for the 8B on a single A6000 with LoRA and gradient checkpointing. Papers exceeding this are truncated using a head-tail strategy: 70% of the token budget from the beginning of the extracted input (introduction, methods) and 30% from the end (results, conclusion), with cuts snapped to section boundaries where possible. This preserves the contribution and methods context that summaries draw on most heavily, along with the results that typically come at the end of a paper.

Temperature 0.7, top-p 0.9, repetition penalty 1.1. Identical decoding settings for all four student systems (both fine-tuned, both base). The teacher was sampled through the Anthropic batch API with default sampling parameters and a 1,024-token cap. Temperature and top-p balance diversity against faithfulness. The mild repetition penalty prevents the looping behavior LLaMA models occasionally exhibit without distorting the output distribution.

Max 512 new tokens. Target summaries are 150–250 words (∼200–350 tokens). 512 provides headroom for the fine-tuned models and the 70B base; the exception is the 8B base, whose longer markdown-heavy outputs reach the cap in 14/150 cases (Section 4).

bf16 for 8B, 4-bit NormalFloat for 70B. The 8B fits on one A6000 in bf16 with LoRA and gradient checkpointing. The 70B in bf16 would require approximately 140GB, so we use QLoRA's NF4 quantization with double quantization to fit on two A6000s (96GB total).

Seed 42. Single seed for reproducibility. We report results from one training and one inference run

per model; multi-seed variance is not characterized.

B Prompt

Two prompts are used. The **teacher prompt** is used for teacher summary generation (Claude Opus 4.6). The **student prompt** is used for fine-tuning supervision (the user message is paired with the teacher summary as the training target for both LLaMA students) and for zero-shot and fine-tuned student inference; it condenses the teacher's instructions into a shorter system message and a single instruction sentence conveying the same constraints. `{title}` is substituted with the paper title; `{content}` and `{input_text}` are substituted with the concatenated section text. Both prompts are reproduced verbatim below.

Teacher prompt

System message.

```
You are an expert scientific paper summarizer. Your task is to generate concise, accurate summaries of scientific papers.

Your summaries must:
1. Be 150-250 words in length
2. Cover four key aspects:
- Main contribution or novelty of the work
- Methodology or approach used
- Key results or findings
- Significance or implications
3. Be FAITHFUL to the source - never include claims not present in the paper
4. Use clear, accessible language while maintaining technical accuracy
5. Be written in third person, present tense

Do NOT:
- Add information not present in the source
- Make speculative claims about future work
- Include citations or references
- Use bullet points or lists - write in flowing prose
```

User message.

```
Please summarize the following scientific paper.

Title: {title}

Paper Content:

{content}

Generate a 150-250 word summary covering the main contribution, methodology, key results, and significance.
```

Student prompt

System message.

```
You are an expert scientific paper summarizer. Generate concise, accurate summaries of scientific papers. Write in third person, present tense, using flowing prose without bullet points.
```

User message.

```
Summarize the following scientific paper in 150-250 words, covering the main contribution, methodology, key results, and significance.

Title: {title}

Paper Content:

{input_text}
```

C Qualitative Examples

We present two examples contrasting fine-tuned and base summaries; the first also includes the teacher summary.

C.1 Physics: "Surface Plasmons in the Continuum"

Teacher. "This study addresses the challenge of accurately calculating surface-plasmon resonances (SPRs) in aluminum clusters at energies above the ionization potential, where excited states are embedded in the ionization continuum. The authors demonstrate that standard real-time TDDFT calculations produce unphysical spectral artifacts because electrons excited into continuum states reflect off finite simulation domain boundaries [...]"

Fine-tuned 8B. "This study presents an ab initio framework for calculating surface-plasmon resonances (SPRs) in metal clusters at energies above the ionization potential, addressing a longstanding challenge in time-dependent density functional theory (TDDFT). The central contribution is demonstrating that TDDFT can reliably describe plasmonic excitations in the ionization continuum [...]"

Base 8B. "**Surface Plasmons in the Continuum / Main Contribution:** The authors develop a grid-based real-time time-dependent density functional theory (RT-TDDFT) approach to accurately describe the absorption spectra of metal clusters above the ionization potential [...]"

Observations. The fine-tuned model produces flowing prose matching the teacher's style. The base model uses bold headers and labeled sections despite the system prompt's explicit instruction to write flowing prose without bullet points, a template-following behavior that fine-tuning eliminates.

C.2 Physics (econophysics): "Sandpile Economics"

Fine-tuned 8B. "This paper develops the 'Sandpile Economy' framework, arguing that capitalist economies self-organize into permanent states of structural fragility through endogenous network evolution. The authors identify three mechanisms: competitive selection favoring specialization [...]"

Base 8B. "**Summary of the Scientific Paper / Main Contribution/Finding:** The authors propose a novel framework, dubbed 'Sandpile Economics,' which posits that the global production network organizes itself into a permanently critical geometric configuration [...]"

Observations. The fine-tuned model identifies three specific mechanisms, a content-level contribution absent from the base output. The stylistic contrast between prose and template is consistent across domains. The fine-tuned output also conflates two terms the source distinguishes, calling the framework "Sandpile Economy" — the source's name for the dynamical system of its Definition 3.6 — where the source names the framework "Sandpile Economics"; the base output uses the framework name correctly. This is a small illustration of the abstractive-error risk we cannot quantify without human annotation (Section 5).

D Full Abstractiveness Table

Metric	FT 70B	B 70B	FT 8B	B 8B	Teach
Novel 1g	19.6%	15.3%	19.0%	15.2%	12.8%
Novel 2g	57.8%	47.4%	56.8%	45.1%	53.0%
Novel 3g	78.2%	66.3%	76.5%	62.8%	75.6%
Novel 4g	88.0%	76.6%	86.4%	72.9%	86.1%
Coverage	80.4%	84.7%	81.0%	84.8%	87.2%
Density	2.46	4.06	2.62	5.16	2.74
Vocab div	0.68	0.60	0.68	0.52	0.66

Table 7: Full abstractiveness analysis. All FT vs. Base differences $p < 0.001$ (Wilcoxon).

E Per-Domain Results

Domain	8B d	70B d
CS	1.27	0.96
Physics	1.32	0.81
Math	1.72	0.75

Table 8: Per-domain FT vs. Base effect sizes (Cohen's d) for BERTScore F1 against abstracts. All six comparisons are significant at $p < 0.001$ (paired sign-flip permutation test).

Dom.	Model	R-1	R-2	R-L	BERT	BLEU	MET.	n
CS	FT 8B	.490	.172	.248	.656	.069	.346	55
	B 8B	.458	.168	.231	.618	.068	.340	55
	Teach	.515	.195	.269	.665	.088	.366	55
Phys	FT 8B	.464	.151	.225	.644	.063	.323	48
	B 8B	.438	.158	.219	.612	.058	.341	48
	Teach	.490	.175	.245	.656	.074	.352	48
Math	FT 8B	.440	.167	.238	.629	.067	.337	47
	B 8B	.383	.155	.214	.583	.063	.320	47
	Teach	.457	.188	.260	.638	.076	.353	47

Table 9: Per-domain results vs. abstracts (8B model).

Dom.	Model	R-1	R-2	R-L	BERT	BLEU	MET.	n
CS	FT 70B	.498	.175	.254	.659	.074	.350	55
	B 70B	.478	.176	.241	.631	.086	.320	55
	Teach	.515	.195	.269	.665	.088	.366	55
Phys	FT 70B	.470	.158	.231	.648	.066	.338	48
	B 70B	.473	.172	.237	.630	.075	.324	48
	Teach	.490	.175	.245	.656	.074	.352	48
Math	FT 70B	.435	.166	.232	.627	.066	.335	47
	B 70B	.423	.165	.228	.606	.079	.314	47
	Teach	.457	.188	.260	.638	.076	.353	47

Table 10: Per-domain results vs. abstracts (70B model).