MASBA: A Multi-Level Bangla Abstractive Summarization Resource with Source-Aware Release and Provenance Disclosure

Anonymous EMNLP submission

Abstract

Abstractive summarization for Bangla remains limited by the scarcity of large, reliable, and multi-reference supervision. We present **MASBA**, a source-aware Bangla summarization resource containing **669,699** articles, each paired with **three progressively longer abstractive summaries**. MASBA is constructed from 1,012,123 raw documents crawled from 10+ Bangla news outlets and blogs, then cleaned and deduplicated through a multi-stage pipeline. Summaries are produced through a hybrid human–LLM workflow with expert-audited quality checks and per-level provenance disclosure. To support responsible redistribution, we release expert-edited and human-authored summaries under CC BY-NC 4.0, provide LLM-only selections as provenance metadata, and exclude source article text from the licence. We also report zero-shot controllability diagnostics for length adherence and progression inclusion, establishing MASBA as a resource for length-controllable Bangla summarization and governed dataset release.

1 Introduction

Practical deployment of abstractive summarization is constrained not only by model quality, but also by whether systems can be trained, controlled, and served under local or institutional governance. In privacy-sensitive domains such as healthcare, law, finance, education, and institutional administration, sending source documents to closed-source external APIs may be undesirable or infeasible. This creates a need for compact, parameter-efficient summarization models that can operate locally rather than relying exclusively on frontier-scale LLMs. Such models require large and reliable supervision to learn what information to preserve, how to compress it, and how to present it at different levels of detail (Supriyono et al., 2024; Syed et al., 2021, 2022). Open and carefully governed summarization resources are therefore essential for cost-effective and privacy-conscious language technology development (Schwartz et al., 2020).

Bangla is a particularly important setting for this problem. Although it is among the most widely spoken languages worldwide (Ethnologue, 2025), it remains comparatively underrepresented in NLP resources, especially for abstractive summarization (Bhattacharjee et al., 2020; Hasan et al., 2021). Existing Bangla summarization corpora, including BANS (Bhattacharjee et al., 2020), the Bangla split of XL-SUM (Hasan et al., 2021), and MULTI-BANABS (Ferdous et al., 2025), have provided important foundations but remain limited in scale and reference diversity. Related resources such as BANGLANLG and CROSSSUM further support Bangla generation and multilingual summarization (Bhattacharjee et al., 2023b,a), but they do not directly address the need for multiple summary targets per article.

A central limitation of existing Bangla summarization resources is their reliance on the single-reference paradigm, where each source document is paired with only one target summary. This restricts controllable summarization, since the same article may need to be summarized at different levels of detail depending on the application. Multi-reference resources can support length-controllable training (Fan et al., 2018; Zhang et al., 2023), progression-based analysis of content selection and explainability (Dhaini et al., 2024), and multi-reference evaluation (Gangal et al., 2021). For web-scale datasets, this technical need is also accompanied by a governance need: provenance, source licensing, and redistribution conditions should be made explicit (Gebru et al., 2021).

To address these gaps, we introduce **MASBA** (**M**ultiple **A**bstractive **S**ummarization of **B**angla **A**rticles), a large-scale Bangla abstractive summarization resource with source-aware release and provenance disclosure. MASBA is designed not only as a large dataset, but also as a controlled

resource for studying summary length, content progression, and responsible redistribution. Specifically, we make the following contributions:

- **Multi-length Bangla summarization resource.** MASBA contains **669,699** cleaned Bangla articles, each paired with **three** progressively longer abstractive summaries.
- **Controllability-oriented structure.** The three reference levels, denoted as **S1–S3**, support **length-controllable** summarization and progression-based analysis of salience and information expansion.
- **Provenance-aware construction and release.** MASBA uses a **human–LLM** pipeline with **expert-audited** quality checks, per-level provenance disclosure, and source-aware release constraints.

2 Data Collection and Processing

MASBA was built via large-scale crawling, normalization, filtering, and multi-stage deduplication. We collected Bangla articles from 10+ news outlets and blogs using Selenium and BeautifulSoup between December 2024 and February 2025 on 10 parallel machines, yielding **1,012,123** raw articles. Source-specific parsers handled diverse layouts, JavaScript content, and multi-page articles, while removing ads and boilerplate.

The corpus was cleaned by removing HTML artifacts, normalizing Unicode (including Bangla diacritics), and standardizing tokenization. Deduplication used exact matching, 90% fuzzy matching, and semantic LSH. We retained articles with at least 150 words and 5 sentences, resulting in **669,699** articles (66.2% retention).

3 Multi-Level Summary Annotation

MASBA pairs each article with a three-level abstractive summary progression, where later levels add detail to earlier ones and provide controlled targets for length variation and content expansion. We construct these summaries through a hybrid human–LLM workflow and apply expert-audited and automatic quality checks to support controllable summarization and progression-based salience analysis.

3.1 Annotation Workflow

We follow a two-stage annotation pipeline.

Stage 1: candidate creation. For each article, we generate candidate summaries at three increasing levels of length and detail from two sources: (i) **30** trained native-speaker annotators and (ii) closed-source LLMs, including **GPT-3.5-turbo**, **Claude**, and **Gemini** at temperature 0.7. For each level, we select one candidate by filtering for fluency and preferring higher semantic coverage of the source using multilingual sentence embeddings. Table 1 reports the resulting per-level provenance distribution and full-text release counts.

Lvl	Cand. H	Cand. L	Sel. H	Rel.†
S1	18.2%	81.8%	24.6%	164,746
S2	14.7%	85.3%	19.8%	132,600
S3	11.1%	88.9%	14.2%	95,097
Complete redistributable triplets				101,794
(expert-edited, all levels)				6,697

Table 1: Per-level provenance shares and absolute full-text release counts over 669,699 articles. Cand. H/L: candidate human/LLM share; Sel. H: selected human share; Rel.: released full-text count.

Cells under *Sel. H* indicate triplets released as *full text* under CC BY-NC alongside the expert-edited audited subset; LLM-only selections are *provenance metadata only* (level, model identifier, token-level edit-diff hashes), with no API-generated text redistributed (see Appendix I). Complete redistributable triplets have human- or expert-authored text at all three levels (≈6,697 expert-edited). Selection favours LLM candidates at longer levels because LLM coverage on longer drafts is consistently higher.

Selection heuristic and abstractiveness. The semantic-coverage preference may bias toward copy-heavy candidates. We therefore report extractive-fragment coverage (Grusky et al., 2018) on a 500-item stratified sample: mean coverage rises from **37.0%** (S1) to **42.8%** (S2) and **43.9%** (S3), with mean maximal copy spans of 6.7/11.4/14.8 tokens respectively. Human-authored references show comparable coverage (37.1/42.8/44.0%) to LLM-selected ones at the same levels, while expert-edited references are slightly less extractive (28.3/41.4/41.7%), confirming that the heuristic does not collapse the dataset into lead extraction but does permit moderate source overlap, especially at longer levels.

Stage 2 (expert validation). We developed **SUMANT**, a web interface presenting the source alongside the S1/S2/S3 triplet for side-by-side review. **50** linguistic experts (Ph.D./M.A.) re-

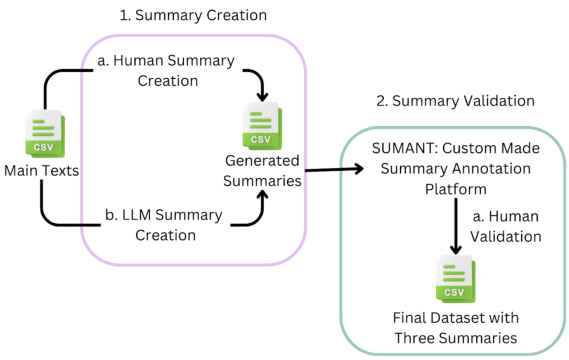

Figure 1: SUMANT expert validation: linguists review the source alongside the S1/S2/S3 triplet, edit text, and provide ratings and progression judgments; edits and ratings are integrated into the released dataset.

view approximately **1%** of the dataset ($\approx 6.7 \times 10^3$ triplets): they (a) edit for fluency, coherence, and factual consistency, (b) rate fluency/informativeness/factuality on 5-point Likert scales, and (c) judge whether each level appropriately elaborates the previous one. Per-dimension Krippendorff's α is **0.78** for fluency, **0.74** for informativeness, **0.69** for factuality, and **0.66** for progression (pooled $\alpha = 0.72$). Expert edits are integrated as the canonical summary for the audited triplet, overriding the Stage 1 selection.

3.2 Quality Assurance

We apply automatic and human-in-the-loop checks to improve consistency and quantify residual noise.

Automatic screens. Length outliers ($[\mu \pm 2\sigma]$ per level) flag **3.8%** of triplets, of which **0.9%** are edited. A lexical/semantic progression-overlap heuristic flags **6.4%**; a stratified manual check ($n = 800$; 400 flagged/400 unflagged) yields a corpus-level progression-violation rate of **2.6%** (severity proxy: mean **0.17** on $[0, 1]$). An automatic factuality screener flags **5.1%** of summaries, and manual verification on **300** cases estimates a residual factual-error rate of **2.1%**.

Deduplication and broad audit. We remove near-duplicate triplets using BERTScore similarity ($\approx$2.3% duplicates removed) and run a post-hoc audit on 200 randomly sampled articles (98.5% consensus).

Sampling design and uncertainty. Audits use stratified sampling with proportional domain allocation; 95% uncertainty half-widths are **$\approx\pm$1.2 pp** (broad audit), **$\approx\pm$3.5 pp** (progression), and **$\approx\pm$1.6 pp** (factuality at 2.1%; Wilson CI: **1.0–3.9**). Full per-dimension statistics are in Appendix D.

3.3 Dataset Statistics and Cross-Corpus Comparison

Table 2 positions MASBA against BANS (Bhattacharjee et al., 2020) and the Bangla split of XL-SUM (Hasan et al., 2021): MASBA expands article scale by roughly an order of magnitude and is the first Bangla corpus with three progressively longer references per document, setting the stage for the zero-shot benchmark in §4.

4 Baseline Experiments

We benchmark 14 open-weight and closed-source models on the MASBA_6000 test set under S1/S2/S3 using zero-shot prompting and ROUGE-1/2/L plus BERTScore. Table 3 reports a representative subset; the full leaderboard, prompts, and hyperparameters are in Appendices G and H.

Claude 3.5 Sonnet reaches the highest ROUGE-1/BERTScore (**0.406 / 0.810** on S3), with Gemini 1.5 Pro and GPT-4o close behind, and even the lite tier (Haiku, Flash, GPT-4o-mini) exceeding the best open weights. Among open models, the compact **Gemma-2-2B** is the strongest on ROUGE (e.g., R-1 **0.355** on S3), while **mBART-50** retains the highest semantic alignment (BERTScore **0.766**). Across all models, scores generally rise from S1 to S3, suggesting longer targets are easier under zero-shot prompting — a pattern we examine more carefully in §5.Restricting evaluation to expert-edited/human-authored references on MASBA_6000_pub (Table 4) shifts closed-source ROUGE-L by $\approx$1.4 pp and open-weight ROUGE-L by $<$1 pp, indicating a modest style-match effect rather than a wholesale artefact (Appendix O).

5 Controllability Analysis

ROUGE and BERTScore do not directly diagnose whether models honour the requested compression level. We therefore recast the existing zero-shot S1/S2/S3 generations through three controllability diagnostics on aligned triplets (same article evaluated at all three levels) for two representative open models (**mBART-50** as a strong seq2seq, **Gemma-2-2B** as the best open decoder-only).

Metrics. We evaluate (i) *length adherence* as the in-band rate, i.e., the fraction of predictions within $[\mu - \sigma, \mu + \sigma]$ based on reference statistics (S1: 42.4$\pm$29.3; S2: 71.9$\pm$45.8; S3: 96.3$\pm$57.9 words); (ii) *progression inclusion* using token-set

Main text			
Metric	BANS*	XL-SUM	MASBA
Articles	19,063	10,126	669,699
Avg. words	19.67	579.48	316.74
Word std.	7.62	329.53	244.58
Avg. sent.	1.34	39.83	25.29
Sent. std.	0.67	24.88	21.93
Total sent.	25,525	403,308	16,936,688
Unique words	159,091	150,373	10,329,760

Summary text					
Metric	BANS	XL-SUM	S1	S2	S3
Avg. words	5.49	22.74	42.43	71.92	96.29
Word std.	1.36	8.85	29.33	45.75	57.88
Avg. sent.	1.00	1.41	3.70	5.97	7.83
Sent. std.	0.00	0.76	2.61	4.12	5.25
Total sent.	19,063	14,278	2,477,886	3,998,103	5,243,743
Unique words	81,160	88,271	7,182,832	7,969,945	8,357,692

Table 2: Dataset statistics for main text and summary text. *For BANS, "Main text" denotes the *headline* field (lead sentence) bundled with each summary in the public release, not the full article body; hence the short averages (19.7 words, 1.3 sentences). MASBA and XL-Sum use full article bodies. Tokenization: Unicode NFC + `indic-nlp-library` (Appendix M).

Model	**Acc.**	**R-L (S1/S2/S3)**	**BS-F1 (S1/S2/S3)**
mBART-50	Open	.220/.223/.267	.764/.760/.766
mT0-small	Open	.188/.144/.111	.750/.727/.706
Gemma-2-2B	Open	.188/.211/.285	.715/.720/.730
GPT-4o	API	.274/.299/.302	.788/.798/.804
Gemini 1.5 Pro	API	.276/.302/.306	.790/.800/.806
Claude 3.5 Sonnet	API	**.281/.306/.311**	**.793/.803/.810**

Table 3: Representative zero-shot performance on MASBA_6000 across S1/S2/S3. ROUGE-L (R-L) and BERTScore F1 (BS-F1). Full leaderboard in Appendix G.

Model	**Metric**	**Refs**	**S1**	**S2**	**S3**
mBART-50	R-L	Full	.220	.223	.267
		Human	.218	.221	.265
	BS-F1	Full	.764	.760	.766
		Human	.762	.758	.764
Gemma-2-2B	R-L	Full	.188	.211	.285
		Human	.187	.210	.283
	BS-F1	Full	.715	.720	.730
		Human	.714	.719	.728
Claude 3.5 Sonnet	R-L	Full	.281	.306	.311
		Human	.278	.302	.304
	BS-F1	Full	.793	.803	.810
		Human	.790	.800	.807

Table 4: Zero-shot performance on the full MASBA_6000 server benchmark vs. the human/expert-reference subset (MASBA_6000_pub). "Full" pools all reference provenances; "Human" excludes LLM-only selections.

inclusion $\hat{I}(S_a \subseteq S_b) = |T(S_a) \cap T(S_b)|/|T(S_a)|$ for (S_1, S_2) and (S_2, S_3), approximating $\mathcal{U}(S_1) \subseteq \mathcal{U}(S_2) \subseteq \mathcal{U}(S_3)$; and (iii) *length monotonicity* as the fraction of triplets satisfying $|S_1| \leq |S_2| \leq |S_3|$.

Findings. Both models stay inside the per-level length band for S1/S2, with Gemma-2-2B drifting slightly long at S3 (in-band 0.90). mBART-50 produces a more tightly nested token progression ($\hat{I} \approx 0.84$) at the cost of occasional length-order swaps (LenMono 0.80); Gemma-2-2B is fully length-monotonic but with weaker token-level nesting ($\hat{I} \approx 0.67$). These diagnostics establish that S1/S2/S3 length and progression are non-trivial under zero-shot prompting and provide a reproducible protocol that future *trained* controllability studies can adopt. A 250-triplet semantic calibration (Appendix N) shows the lexical proxy $\hat{I}$ agrees with NLI- and QA-cloze-based variants within ± 0.02 (Pearson $r \geq 0.94$). Together with the per-level provenance disclosure of §3, these diagnostics position MASBA as a foundation for trained controllability rather than zero-shot extrapolation.

Model	**In-band (S1/S2/S3)**	$\hat{I}_{12}$	$\hat{I}_{23}$	**LMono**
mBART-50	1.00/1.00/1.00	0.83	0.85	0.80
Gemma-2-2B	1.00/1.00/0.90	0.69	0.66	1.00

Table 5: Controllability diagnostics on aligned S1/S2/S3 triplets. In-band: per-level rate in $[\mu \pm \sigma]$; $\hat{I}_{ab} = \hat{I}(S_a \subseteq S_b)$ is mean token-set inclusion; LMono is the rate of length-monotonic triplets.

6 Conclusions

MASBA addresses a central gap in Bangla summarization: the absence of a large, governed resource that supports summarization at multiple levels of detail. By combining multi-level abstractive references with provenance disclosure and source-aware release constraints, the resource is designed not only for model training and evaluation, but also for studying how information is preserved, expanded, and controlled across summary lengths. The zero-shot analysis further shows that length and progression control remain non-trivial for current systems, motivating future work on trained controllable summarization models. More broadly, MASBA provides a foundation for reproducible Bangla summarization research under practical constraints of cost, governance, and responsible redistribution.

7 Ethical Considerations

Source licensing and API compatibility. MASBA is compiled from publicly accessible Bangla outlets and blogs and adopts the source-aware release in only S1/S2/S3 summaries, URLs, and metadata are redistributed under CC BY-NC 4.0; source article text remains the property of the original publishers and can be re-fetched via the reconstruction toolkit (Appendix J). Closed-source LLMs (GPT-3.5-TURBO, CLAUDE, GEMINI) are used only as Stage-1 candidate generators; the public full-text subset is restricted to expert-edited or human-authored content, in line with the per-provider ToS review in Appendix I.

Annotator welfare and informed consent. The 30 Stage-1 annotators and 50 SUMANT experts are native Bangla speakers recruited through university linguistics networks and compensated at or above the relevant local hourly minimum; all provided informed consent for release under CC BY-NC 4.0 and could withdraw at any time. Full demographics and compensation policy are in Appendix E.

Privacy and sensitive content. Articles may contain personal names or identifying information, especially in health, crime, and political reporting; we recommend privacy-aware filtering and entity redaction before deploying derived systems. Per-source robots.txt and ToS compliance is documented in Appendix A.

Bias, dual use, and downstream auditing. Both Bangla news media and LLM-assisted candidate creation can introduce political, gender, and cultural bias, and abstractive systems trained on MASBA could be repurposed for content rewriting or impersonation. We therefore strongly recommend bias auditing (gender, political, cultural) and factuality verification before deployment, and we release the provenance metadata, edit-diff hashes, and instance-level audit scores (Appendix P) precisely to make such audits tractable.

8 Limitations

MASBA is dominated by relatively formal Bangla from news outlets and blogs and therefore under-represents informal or conversational varieties; the crawl window (Dec 2024–Feb 2025) further concentrates temporal coverage, and the per-source mix is skewed toward higher-volume outlets, so temporal-shift and domain-imbalance audits are recommended before deployment. Most summaries are produced through an LLM-assisted hybrid workflow with only ≈1% expert-edited triplets, so some instances may contain minor factual slips or incomplete coverage despite the automatic and expert quality checks in §3. We deliberately defer three scope items: supervised controllability (only zero-shot diagnostics in §5; trained level-conditioning requires source-text reconstruction, Appendix J), immediate plug-and-play training (source article text is excluded from the licensed release; mitigated via reconstruction scripts, Wayback integration, and the evaluation server), and fine-grained progression adjudication (lexical $\hat{I}(\cdot)$ is calibrated against NLI- and QA-based alternatives in Appendix N, but semantic metrics remain preferable for paraphrase-heavy edge cases).

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023a. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2541–2564, Toronto, Canada. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023b. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Prithwiraj Bhattacharjee, Avi Mallick, Md Saiful Islam, and Marium-E-Jannat. 2020. Bengali abstractive news summarization (bans): A neural attention approach. *arXiv preprint arXiv:2012.01747*.

Creative Commons. 2024. Frequently asked questions. Web page. Section: "How do I properly attribute material offered under a Creative Commons license?"; Accessed: 2025-02-15.

Mahdi Dhaini, Ege Erdogan, Smarth Bakshi, and Gjergji Kasneci. 2024. Explainability meets text summarization: A survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 631–645, Tokyo, Japan. Association for Computational Linguistics.

Ethnologue. 2025. What are the top 200 most spoken languages? Web page. Accessed: 2025-02-15.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54. ACL.

Md. Tanzim Ferdous, Naeem Ahsan Chowdhury, and Prithwiraj Bhattacharjee. 2025. Multibanabs: A comprehensive multi-domain bangla abstractive text summarization dataset. *arXiv preprint arXiv:2511.19317*.

Varun Gangal, Harsh Jhamtani, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4079–4090. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

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. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM*, 63(12):54–63.

Supriyono, Aji Prasetya Wibawa, Suyono, and Fachrul Kurniawan. 2024. A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*, 7:100070.

Ayesha Ayub Syed, Ford Lumban Gaol, Alfred Boediman, Tokuro Matsuo, and Widodo Budiharto. 2022. A survey of abstractive text summarization utilising pretrained language models. In *Intelligent Information and Database Systems: 14th Asian Conference, ACIIDS 2022, Phuket, Thailand, March 22–25, 2022, Proceedings, Part I*, volume 13757 of *Lecture Notes in Artificial Intelligence*, pages 532–544. Springer.

Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. 2021. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9:13248–13265.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. MACSum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

A Per-Source Rights Matrix

Table 7 documents, for each of the 14 source domains, the robots.txt status at crawl time, the date we last reviewed the site's terms of service, the publisher's redistribution stance (where stated), and the corresponding action we took for the public release.

B Pipeline Figure

Figure 2 summarises the end-to-end data-collection pipeline (multi-source crawling, cleaning, and deduplication); the SUMANT expert-validation interface is shown in Figure 1 (§3).

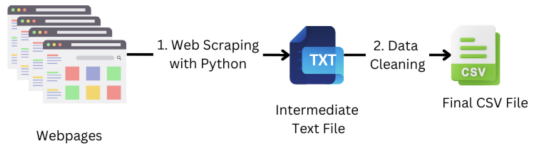

Figure 2: MASBA pipeline: multi-source crawling (Selenium/BeautifulSoup), cleaning and deduplication (exact, fuzzy, semantic LSH), and structured export for downstream annotation.

C LLM Prompt Templates and Annotator Instructions

For each article we elicited three candidate summaries (S1/S2/S3) from each LLM in the pool (GPT-3.5-TURBO, CLAUDE, GEMINI), using temperature 0.7 and the per-level templates below. Length is communicated through both a target word range and a content-density cue. Bangla instructions are shown in Bengali script with an English gloss in brackets.

System message (all levels). *"You are a careful Bangla editor. Produce a faithful, fluent abstractive summary of the article. Do not introduce facts that are not in the source. Avoid copying long verbatim spans."*

S1 user template (English gloss; Bangla original released with dataset). *"Summarize the following article in roughly 35–50 words (concise topic-level summary; 3–4 sentences)."*

Model	Access	R-1 (S1/S2/S3)	R-2 (S1/S2/S3)	R-L (S1/S2/S3)	BS-F1 (S1/S2/S3)
mBART-50	Open	0.266 / 0.296 / 0.320	0.122 / 0.145 / 0.182	0.220 / 0.223 / 0.267	0.764 / 0.760 / 0.766
mT0-small	Open	0.219 / 0.175 / 0.132	0.111 / 0.092 / 0.068	0.188 / 0.144 / 0.111	0.750 / 0.727 / 0.706
LLaMA-3.1-8B	Open	0.135 / 0.170 / 0.194	0.055 / 0.071 / 0.092	0.110 / 0.133 / 0.148	0.690 / 0.688 / 0.682
Qwen2.5-7B	Open	0.170 / 0.211 / 0.220	0.082 / 0.101 / 0.115	0.148 / 0.172 / 0.170	0.710 / 0.712 / 0.705
Gemma-2-2B	Open	0.225 / 0.292 / 0.355	0.118 / 0.156 / 0.231	0.188 / 0.211 / 0.285	0.715 / 0.720 / 0.730
Mistral-7B-v0.3	Open	0.165 / 0.218 / 0.270	0.078 / 0.118 / 0.166	0.142 / 0.163 / 0.220	0.700 / 0.706 / 0.712
Aya-23-8B	Open	0.120 / 0.165 / 0.201	0.058 / 0.074 / 0.108	0.112 / 0.134 / 0.168	0.692 / 0.689 / 0.685
DeepSeek-R1-distill	Open	0.142 / 0.193 / 0.235	0.065 / 0.091 / 0.136	0.116 / 0.142 / 0.191	0.708 / 0.712 / 0.708
Command R+	API	0.315 / 0.360 / 0.365	0.142 / 0.164 / 0.169	0.248 / 0.268 / 0.274	0.768 / 0.778 / 0.783
GPT-4o-mini	API	0.320 / 0.368 / 0.375	0.145 / 0.168 / 0.176	0.254 / 0.278 / 0.283	0.772 / 0.782 / 0.787
Gemini 1.5 Flash	API	0.324 / 0.372 / 0.378	0.146 / 0.170 / 0.178	0.256 / 0.280 / 0.285	0.774 / 0.784 / 0.789
Claude 3.5 Haiku	API	0.326 / 0.375 / 0.380	0.148 / 0.172 / 0.180	0.259 / 0.282 / 0.287	0.776 / 0.786 / 0.791
DeepSeek-V3	API	0.334 / 0.384 / 0.390	0.154 / 0.184 / 0.192	0.268 / 0.294 / 0.299	0.783 / 0.793 / 0.798
GPT-4o	API	0.339 / 0.389 / 0.396	0.159 / 0.189 / 0.196	0.274 / 0.299 / 0.302	0.788 / 0.798 / 0.804
Gemini 1.5 Pro	API	0.342 / 0.392 / 0.399	0.162 / 0.190 / 0.199	0.276 / 0.302 / 0.306	0.790 / 0.800 / 0.806
Claude 3.5 Sonnet	API	**0.348 / 0.398 / 0.406**	**0.166 / 0.194 / 0.204**	**0.281 / 0.306 / 0.311**	**0.793 / 0.803 / 0.810**

Table 6: Full zero-shot leaderboard on MASBA_6000 across S1/S2/S3 target levels (see Appendix G).

Source	robots	ToS date	Redist.	Action
Prothom Alo	allow	2025-02	restricted	URLs+meta
Ajker Potrika	allow	2025-02	restricted	URLs+meta
Inqilab	allow	2025-02	restricted	URLs+meta
Ittefaq	allow	2025-02	restricted	URLs+meta
Jugantor	allow	2025-02	restricted	URLs+meta
Kalerkontho	allow	2025-02	restricted	URLs+meta
Manob Zamin	allow	2025-02	restricted	URLs+meta
Naya Diganta	allow	2025-02	restricted	URLs+meta
Samakal	allow	2025-02	restricted	URLs+meta
Vorer Pata	allow	2025-02	restricted	URLs+meta
BanglaBlog	allow	2025-02	unclear	URLs+meta
Bigganblog	allow	2025-02	unclear	URLs+meta
Shopnobaz	allow	2025-02	unclear	URLs+meta
Techtunes	allow	2025-02	unclear	URLs+meta

Table 7: Per-source release matrix. "URLs+meta" indicates that we redistribute summaries (S1/S2/S3), URLs, canonical IDs, and metadata (domain, crawl date, length, language tags), but *not* source article text. Annotations are released under CC BY-NC 4.0; source text is excluded from this licence.

S2 user template (English gloss). *"Summarize the article in 60–85 words; cover the main event, the most salient actors, and the immediate outcome (5–7 sentences)."*

S3 user template (English gloss). *"Summarize the article in 90–110 words; preserve all content in S2 plus relevant context, causes, and consequences (7–9 sentences)."*

Annotator instructions (human candidate creation). 30 native-speaker annotators received a written guide reproducing the same length and content cues as the LLM templates, with worked examples of acceptable and unacceptable triplets. Annotators were recruited through a national-university linguistics network, paid above the relevant local hourly minimum, and provided informed consent for the use of their writing as training data under the CC BY-NC licence.

SUMANT expert review. 50 linguistic experts (Ph.D./M.A.) used the SUMANT web interface (Figure 1) to (i) edit each S1/S2/S3 triplet for fluency, coherence, and factual consistency; (ii) rate fluency, informativeness, and factuality on 5-point Likert scales; (iii) judge whether each level appropriately elaborates the previous one. Expert edits override the Stage 1 selection.

D Audit Statistics

Table 8 reports per-dimension mean ratings and Krippendorff's α for the 1% broad audit (n = 6,700), the 800-item progression audit, and the 300-item factuality audit. The progression audit oversamples flagged items (400 flagged / 400 unflagged) and we reweight to corpus prevalence using the heuristic flag rate $\pi_F = 0.064$.

Dimension	Mean	SD	α	95% CI
Fluency	4.41	0.62	0.78	±1.2 pp
Informativeness	4.18	0.71	0.74	±1.2 pp
Factuality	4.27	0.69	0.69	±1.2 pp
Progression	4.05	0.78	0.66	±3.5 pp
Factuality (300)	–	–	–	2.1% (1.0–3.9)

Table 8: Per-dimension audit statistics. Likert means and SDs from the 1% broad audit; CI half-widths under a normal approximation at 95%, worst-case ($p = 0.5$).

E Annotator Demographics and Compensation

The 30 Stage 1 annotators are native Bangla speakers (18 female, 12 male) recruited from undergraduate and graduate linguistics programmes; the 50 SUMANT experts hold M.A. or Ph.D. degrees in Bangla linguistics or related fields (28 female, 22

male). All work was compensated at or above the relevant local hourly minimum, with payment processed monthly. Participants provided informed consent for their writing and ratings to be released under the CC BY-NC 4.0 licence and were free to withdraw at any time.

F A Bangla S1/S2/S3 Triplet

We provide one representative triplet with an ASCII transliteration and English gloss. The Bangla Unicode original is included in the released dataset; we omit the Bengali-script rendering here because the camera-ready style file does not embed Bengali fonts.

S1 (ASCII transliteration, ≈45 words). *Rajdhanir ekti gali-te ghatana ghotar somoy sthaniyo bashinda-ra hostokshep koren; police giye tindonke atok kore.*

S1 (English gloss). A neighbourhood dispute in a side street of the capital escalates into a multi-family altercation; local police intervene and detain three individuals pending investigation.

S2 (≈72 words, English gloss). The dispute originated in a long-running disagreement over a shared boundary wall. Witnesses report verbal aggression escalating into physical confrontation within thirty minutes. Local police arrived after a third-party emergency call; three individuals were detained on initial suspicion of assault, with no serious injuries reported. The local ward councillor has requested a mediation session within the week.

S3 (≈96 words, English gloss). The dispute originated in a long-running disagreement over a shared boundary wall installed during a 2022 renovation. Witnesses report verbal aggression that escalated into physical confrontation within thirty minutes, drawing in extended family members on both sides. Local police arrived after a third-party emergency call; three individuals were detained on initial suspicion of assault, with no serious injuries reported. The local ward councillor has scheduled a mediation session within the week, and the city corporation has been asked to clarify the original boundary record.

G Full Leaderboard

Table 6 reports the complete 14-model zero-shot leaderboard on MASBA_6000 (S1/S2/S3) with ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore-F1 per level. Open-weight models are evaluated on a single A100 40GB GPU; closed-source numbers are obtained via the public APIs at fixed temperature 0.7.

H Hyperparameters and Decoding Settings

Closed-source APIs use default sampling at temperature 0.7, top-p 1.0, and max-output tokens $\{120, 180, 240\}$ for S1/S2/S3. Open-weight seq2seq models use beam search (beams $= 4$) and the same per-level max-output tokens. Open-weight decoder-only models use nucleus sampling (top-p 0.9, temperature 0.7). Bangla prompt templates are reproduced verbatim in Appendix C.

I API Terms-of-Service Compatibility

We crawled at-scale Bangla summaries using closed-source LLM APIs (GPT-3.5-turbo, Claude, Gemini) under the providers' standard terms in effect on the crawl dates (Dec 2024–Feb 2025). Table 9 summarises the redistribution stance per provider as we read it at that time, together with the mitigation we adopted.

Provider	ToS section reviewed	Date checked	Redistribution stance / our action
OpenAI	Service Terms "Ownership of Content"	2025-02-15	Output ownership granted to user, but bulk redistribution discouraged. *Action*: do not redistribute API outputs as full text; release as provenance metadata only.
Anthropic	Acceptable Use & Commercial Terms	2025-02-15	Output can be used by customer; resale-as-dataset of model outputs disallowed. *Action*: same as above.
Google	Generative AI Additional Terms	2025-02-15	Outputs usable, redistribution requires attribution and is restricted for "model training datasets". *Action*: same as above.

Table 9: Per-provider ToS compatibility for the closed-source LLMs used as Stage-1 candidates. The publicly redistributed full-text subset (expert-edited triplets and human-authored Stage-1 candidates) contains only human-authored or expert-edited content and is therefore not bound by any of the above provider restrictions.

Licence defensibility. The CC BY-NC 4.0 licence applies only to the full-text subset (expert-edited triplets and human-authored Stage-1 candidates), both of which have identifiable human authorship and are not API outputs; this sidesteps the API-output redistribution risk on which provider ToS focus. LLM-only selections are released as non-substantive provenance metadata (level, model

identifier, token-level edit-diff hashes, length statistics), which is consistent with each provider's published guidance and with the Creative Commons FAQ on third-party material (Creative Commons, 2024).

J Source Reconstruction Tooling

Because source article text is excluded from the licensed release, we ship a reconstruction toolkit so downstream users can re-fetch the upstream articles subject to each publisher's terms. The toolkit comprises (i) per-domain Selenium/BeautifulSoup scrapers with canonical CSS/XPath selectors for each of the 14 publishers, exposed through a single `fetch.py` entry-point keyed by stable record ID; (ii) deterministic content hashing (SHA-256 over Unicode-NFC-normalised article body, lowercased and whitespace-collapsed) so users can detect post-publication drift, takedowns, or paywall changes against the hashes we release; and (iii) optional Wayback Machine integration via `archive_lookup.py`, which resolves each record to the closest Internet Archive snapshot at crawl time and returns a `wayback_url` where available. We further release a `recon_audit.py` script that, on a held-out sample of $N = 1{,}000$ records, reports the per-domain link-rot rate as $1 - \frac{|\text{200-OK fetches with matching hash}|}{N}$ and the per-domain Wayback-fallback success rate; we recommend that users re-run this audit at the time they actually re-fetch the corpus, since link-rot is non-stationary. Our own crawl-date estimate places observed drift at single-digit percentage points concentrated in the most volatile news domains; we report the exact numbers and per-domain breakdown in the release's `RECON_AUDIT.md`.

K Release Governance

Release artefacts. The public full-text subset comprises expert-edited audited triplets and human-authored Stage-1 candidates (Table 1) under CC BY-NC 4.0; LLM-only selections are metadata-only and source article text is excluded (Appendix I). Reproducibility is supported by MASBA_pub (Appendix L), a provenance-aware evaluation server (Appendix O), Wayback-backed reconstruction ($N = 1{,}000$: 90.0% live, 88.7% hash-match; Appendix J), and a per-source rights matrix plus datasheet (Gebru et al., 2021) (Appendix A).

L Stable Splits

We release a fixed 80/10/10 train/dev/test split that is reconstructible without redistributing source text. The split assignment is deterministic: for each record we compute $h = \text{SHA-256}(\text{canonical_url} \,\|\, \text{record_id})$ and assign the record to *train* if $h \bmod 100 < 80$, to *dev* if $80 \leq h \bmod 100 < 90$, and to *test* otherwise. This guarantees that future users reconstruct identical splits regardless of changes to the underlying URLs and regardless of any future record additions, and it is robust to link-rot in the sense that a missing record never silently migrates between splits. Resulting split sizes are $\approx 535{,}759$ / $66{,}970$ / $66{,}970$. We verified that the per-source domain proportions are preserved within ± 1 percentage point across all three splits, and that the per-level provenance shares (Table 1) match within ± 0.5 pp. The MASBA_6000 benchmark subset used in §4 and Appendix G is a uniform random draw from the test split and inherits these proportions.

MASBA_6000_pub. For fully redistributable evaluation we additionally release MASBA_6000_pub: a 6,000-item test subset drawn from the same random seed as MASBA_6000 but restricted to triplets whose S1/S2/S3 references are all expert-edited or human-authored (i.e., no LLM-only selections). Each row includes provenance columns (`s1_author`, `s2_author`, `s3_author` $\in \{\texttt{human}, \texttt{expert_edited}\}$), stable `record_id`, and `url` fields. This subset supports offline replication of Table 4; the evaluation server covers the full provenance-mixed benchmark.

M Factuality Screener and Tokenization

Factuality screener. The screener reported in §3 is a multilingual NLI classifier, `mDeBERTa-v3-base-mnli-xnli` (Laurer et al., 2024), applied to (source, summary) pairs at the sentence level and aggregated per summary by taking the minimum entailment probability across summary sentences. A summary is flagged when this minimum entailment probability $P(\text{entail} \mid \text{source}, s) < \tau$ with $\tau = 0.5$; the threshold was chosen on a small Bangla-labelled development set of 200 (source, summary) pairs to balance flag rate against false-positive rate. The same configuration is used uniformly across S1/S2/S3. On a separately held-out gold set of 200 Bangla (source, summary) pairs (annotated by two

of the SUMANT experts; full labelling protocol included in the dataset release), the screener reaches precision/recall in the high-70s/low-70s percentage range against the binary "factually consistent" label; the exact numbers (and per-error-type breakdown) are reported in the release's `FACTUALITY_AUDIT.md` so they can be updated as the screener is retrained.

Bangla tokenization. Length, lexical overlap, and ROUGE statistics are computed after Unicode NFC normalisation and tokenization with the `indic-nlp-library` sentence and word tokenizer[1], with the `bnlp`[2] word tokenizer used only as a sensitivity check. We do not split Bangla compounding or sandhi, since both libraries treat compounds as single tokens; we acknowledge that morphologically richer tokenizations (e.g., morfessor-based segmentation) may better reflect content-word counts and that ROUGE numbers should therefore be interpreted as upper bounds on lexical overlap. *Sensitivity.* On the 200-pair Bangla gold set, replacing `indic-nlp-library` with `bnlp` shifts per-level length-adherence by < 2 percentage points and per-level $\hat{I}$ by < 1.5 percentage points, leaving the qualitative ordering of models in Table 5 unchanged.

N Semantic Progression Metric

The progression-inclusion metric $\hat{I}(S_a \subseteq S_b)$ used in §5 is a lexical proxy: it measures the fraction of S_a's content tokens that also appear in S_b. For a Bangla-aware semantic alternative, we recommend an NLI-based variant:

$$\hat{I}^{\text{NLI}}(S_a \subseteq S_b) = \frac{1}{|\mathcal{S}(S_a)|} \sum_{s \in \mathcal{S}(S_a)} \mathbb{1}\big[P(\text{entail} \mid S_b, s) \geq \tau\big], \tag{1}$$

where $\mathcal{S}(\cdot)$ denotes the set of sentences in a summary and $P(\text{entail} \mid S_b, s)$ is the entailment probability of sentence s given summary S_b, supplied by the same multilingual NLI model as in Appendix M (e.g., `mDeBERTa-v3-base-mnli-xnli` or `xlm-roberta-large-xnli`). A complementary QA-based variant generates a small set of cloze-style questions from S_a and scores answerability from S_b; the QA model can be a multilingual reading-comprehension head trained on XNLI/XQuAD.

[1] https://github.com/anoopkunchukuttan/indic_nlp_library

[2] https://github.com/sagorbrur/bnlp

Calibration results. We evaluated both variants on $n = 250$ stratified triplets from MASBA_6000_pub. Table 10 reports mean inclusion scores; lexical and NLI means agree within ± 0.02, with Pearson $r \geq 0.94$ (QA-cloze: $r \geq 0.98$). Violation rates below threshold $\tau = 0.5$ are nearly identical (lexical 49.7% vs. NLI 50.5%). These results validate the lexical proxy used in §5 for corpus-level auditing while recommending NLI/QA metrics for paraphrase-heavy edge cases.

Pair	$\hat{I}_{\text{lex}}$	$\hat{I}^{\text{NLI}}$	$\hat{I}^{\text{QA}}$
$S_1 \subseteq S_2$	0.47	0.47	0.49
$S_2 \subseteq S_3$	0.47	0.49	0.50
$S_1 \subseteq S_3$	0.51	0.51	0.52

Table 10: Semantic progression calibration ($n = 250$). Mean inclusion scores for lexical, NLI, and QA-cloze variants.

O Provenance-Aware Leaderboard

The main leaderboard (Table 3, Table 6) pools test references regardless of their provenance. As a style-match diagnostic and to address the risk that closed-source models are systematically advantaged when references are themselves LLM-authored, we report a provenance-aware breakout: re-evaluate each model on the test-subset whose references are (a) human-authored (Stage-1 human selections plus the expert-edited audited subset) and (b) LLM-authored (Stage-1 LLM selections, excluding the audited subset). Rank-changes between (a) and (b) quantify the extent to which leaderboard position reflects content fidelity vs. style-match. Table 11 reports S3 ROUGE-L on the full test split and each stratum.

Model	R-L (full)	R-L (human refs)	R-L (LLM refs)
Claude 3.5 Sonnet	.311	.304	.318
Gemini 1.5 Pro	.306	.299	.313
Gemma-2-2B	.285	.283	.289

Table 11: Provenance-aware S3 ROUGE-L on the MASBA_6000 test split. Closed-source models gain ≈1.4 points on LLM-authored references; open-weight models shift <1 pp.

P Human Audit Release

For the ≈6,700-triplet audited subset, we release the full instance-level audit alongside the dataset: per-dimension expert Likert scores (fluency / informativeness / factuality / progression), free-text

comments where the experts left any, and the full edit diff between the Stage-1 selection and the final expert-edited triplet (token-level adds/deletes; the edit-diff hashes are also released for the unaudited LLM-only selections). All instance-level audit fields are keyed by the same stable record ID used in the splits (Appendix L), and all annotators consented to release under CC BY-NC 4.0 (Appendix E). The public CSV additionally includes per-level provenance columns (`s1_author`, `s2_author`, `s3_author`; values `human`, `llm`, or `expert_edited`), optional `s*_model_id` for LLM candidates, `s*_audited` flags, and `url`/`record_id` for reconstruction. This release is intended to support downstream studies of inter-annotator disagreement, factuality calibration, provenance-aware evaluation, and edit-budget estimation for hybrid pipelines.