SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Mitigating Factual Hallucination in Low-Resource Indic Summarization: A Generate–Verify–Correct Approach for Telugu

Dr. Sheeja Sudheer[1], Boina UmaDevi[2]

[1]Associate Professor, Hindustan Institute of Technology and Science, Chennai, India.
[2]MCA II year Student, Hindustan Institute of Technology and Science, Chennai, India.

Abstract
While the ability of Large Language Models (LLMs) to produce abstractive summaries has proven to be very good, they do have a tendency to hallucinate facts that are missing or incorrect from the input text. This issue is not well studied for low-resource Indian languages like Telugu where standard factual-consistency measures (like FactCC, QAGS) from English do not map well because of Telugu's morphologically complex, agglutinative structure. In this paper, the focus is laid on the factual consistency in Telugu abstractive summarization system using text extracted from the 10th class Telugu medium Social History textbook (AP/Telangana SCERT). We build a dataset that consists of 183 paragraph-level text units from two chapters titled "The Rise of Nationalism in Europe" and "Nationalism in India", and evaluate in detail 28 representative text units. We automatically generate summaries with our fact-verification pipeline and these are compared with baseline (plain-prompted) summaries as well as with manually produced factual-accuracy evaluations. Results of our analysis indicate that 93% of the baseline summaries include at least one factual omission or error: the most frequent omission or error is a missing named entity (32%) and a missing date or numeric information (21%). The fact-correction pipeline removes all the errors detected and corrected and enhances the ROUGE-1 score from 0.199 to 0.500. The findings suggest that factually correcting Telugu summarization systems via verification is a viable approach for enhancing the factual correctness of such systems, especially in the context of educational content where factual accuracy is paramount.

1. Introduction

With the emergence of Large Language Models (LLMs), the field of automatic text summarization has undergone a significant shift. With advances in the field, abstractive summarizers (using LLM-based) can produce fluent, coherent, and human-like summaries that paraphrase and synthesize the source text, unlike the extractive summarizers of the past, which would only select and rearrange sentences. This ability has led LLMs to be a popular tool for various purposes, ranging from condensing news articles to summarizing scientific papers to, more recently, simplifying educational texts. But, fluency does not equal faithfulness. Despite their grammatically correct output, LLM models tend to hallucinate facts, or create information that is not in the original text or supported by it, or alter facts by changing names, dates, relationships, or simply ignoring information that is critical to the meaning of the original text. Even though LLM models can produce grammatically correct output, they tend to hallucinate facts: output facts that do not exist in or are not supported by the source text, or distort facts by changing names, dates, relationships, or simply ignoring information that is important to the meaning of an original text. Most of the errors are not always accompanied by a corresponding loss of fluency; a hallucinated summary can sound as smooth as a faithful one and be hard for an unsuspecting reader to detect.

This is the gap between fluency and factual reliability that is very significant in pedagogical settings. Students are using AI-generated summaries as study tools: to summarise textbook chapters, to review prior to exams and as a complement to classroom teaching. In a context like this a summary is not just a convenience, but a stand-in for the text itself; a student might never again need to consult the original text to check what it says. When a summarization tool leaves a key date out, puts something in the wrong place, or invents a cause-and-effect relationship that the book never discussed, the student learns about a different version of history or science or a completely false one. When hallucination occurs in everyday or leisure settings, there is no direct pedagogical cost, although factual errors do occur in education, in the process of summarizing information. This situation is exacerbated when we take into account the language landscape outside the English language. While it's one of the most widely spoken languages in India—one of the most spoken languages in the world, though more than 80 million people speak it, it is under-resourced, particularly in terms of NLP research. There is little research conducted on abstractive summarization in the Telugu language per se and for the factuality of the abstractive summaries. This

shortage is not just data-related, it is a methodological one as well. Factual-consistency metrics and correction methods, most of which are widely used, such as FactCC, QAGS, and entity-overlap-based checkers, were created and tested on the CNN/DailyMail news summarization collection, which is written in English. Most of the existing factual-consistency metrics and correction methods, including popular methods like FactCC, QAGS, and entity-overlap-based checkers, were developed and tested on the CNN/DailyMail news summarization collection, which is written in English. These benchmarks take a language with relatively low morphology, standard surface-forms for named-entities and a rich source of pretraining data for granted. The Telugu language is, by contrast, a morphologically complex, agglutinative Dravidian language, in which some words carry multiple levels of meaning – including case, number, tense, and more besides – through various combinations of suffixes, and named entities and numerical expressions can have several possible surface forms. The applicability of such tools that are designed for English to a language of this structural complexity is not obviously clear — and, in fact, may be misleading and unhelpful in the absence of adaptation.

Inspired by this need, this paper examines the following research question: is it meaningful to improve LLM-generated Telugu summaries in an educational (textbook) context using a lightweight Generate → Verify → Correct pipeline, and what value does the multi-pass correction process provide in comparison to a single-pass plainly prompted summary?

To make this investigation realistic and practically motivated, our study is on the Telugu medium educational content, like paragraphs taken from the tenth standard Social History Text Book adopted by the Andhra Pradesh / Telangana State Council of Educational Research and Training (SCERT). Factual consistency is especially appropriate for the study of social history, because in this type of text, the factually verifiable information is especially thick (e.g., names of historical actors, dates, historical sequences, and an author's explicit causal relationships between these). This allows to systematically detect, classify and validate facts that are wrong in the generated summary, without requiring subjective assessments of quality or coherence.

We review related work on constructing a factual consistent dataset for summarization and existing evaluation metrics in Section 2 and describe our dataset construction and Generate–Verify–Correct pipeline in Section 3. An experimental setup and evaluation methodology are presented in Section 4 and results are reported and discussed in Section 5. Section 6 concludes with implications and directions for future work.

2. Related Work

2.1 General Abstractive Summarization: Surveys and Benchmarks

Several recent surveys have mapped the broader landscape of abstractive summarization research, providing useful context for positioning this study. In order to situate this study within the context of existing work in abstractive summarization, a number of recent surveys have surveyed the general field of abstractive summarization research. In their survey, Almohaimeed and Azmi [2] provide an exhaustive overview of abstractive summarization techniques, system architectures and evaluation methods including LLM-based methods, and make a clear call to action for human-in-the-loop verification and better evaluation methods in open research challenges. In the same way, Shakil, Farooq and Kalita [17] review the existing state-of-the-art in abstractive summarization and emphasize the issues of factual inconsistencies and cross-lingual evaluation as major challenges that are yet to be addressed – both which are directly tackled by this work in the context of Telugu. Nnadi and Bertini's [13] survey complement the overview by concentrating on summarization datasets, model architectures and evaluation metrics throughout the field. From the application perspective, Rani Krishna et al. [16] evaluate pretrained transformer models (T5, BART, PEGASUS) for automated generation of news digest, and Anand Babu and Badugu [3] suggest a hybrid approach that uses TextRank and Bi-LSTM-RNN for multi-document summarisation. These two latter studies, however, do not consider factual consistency, nor are they also in the context of Telugu language.

2.2 Telugu and Indic-Language Summarization.

Several recent efforts focus directly on Telugu and other Indic languages, although most of this research was conducted prior to the LLM era or does not incorporate LLM-based generation, but instead makes use of neural architectures prior to the invention of the LLM. Varaprasad Rao [23] suggests that the Automatic Telugu Text Summarization (ATES) algorithm may be based on Adaptive Trans-Residual LSTM model, optimized using an Enhanced Serval Optimization Algorithm. In related works, Varaprasad Rao, Chakma, Jamatia and Rudrapal [24] present an alternative Telugu summarization system that is based on a pointer-network approach that optimizes the attention layer. In this context, Garugu and Bhaskari [9] present a two-stage neural network (EATS2N) for abstractive summarization of Telugu language and Anand Babu and Badugu [3] use their proposed hybrid Bi-LSTM approach for summarization of Telugu multi-document inputs specifically.

At the multilingual level, a few efforts further try to make the Summarization capability available for Indic languages such as Telugu using fine-tuned Transformer models. To solve this problem, Siginamsetty et al. [18] came up with a query-based multilingual abstractive summarization system, dubbed MATSFT, which is developed through fine-tuning mT5 for low-resource Indian languages. The same research group continues this work in MMSFT [19] that uses multimodal (text and image) inputs in the multilingual summarization pipeline. In fine-grained tuning of sequence-to-sequence models to the IL-SUM 2022 data, Urlana et al. [22] focus on Hindi and Gujarati mainly. In order to explore the applicability of multilingual transformer models on a wider variety of Indian languages, Taunk and Varma [21] compare the models, such as IndicBART and mT5. Most relevant to the present study is the study by Adithya et al. [1] that utilizes LLM prompting with text cleaning and keyword extraction for monolingual Indic language summarization with ROUGE, BLEU and BERTScore as evaluative metrics. The present paper, however, makes up for a gap in the existing literature on summarization in Telugu and Indic languages — that of a factuality analysis specific to Telugu language.

While summarization is a particular challenge for Telugu, several studies highlight the unique challenges of LLM-based NLP with the Telugu language in general. While not directly focused on summarization, TeluguEval benchmark [15] measures the capabilities of several LLMs on Telugu reasoning tasks and demonstrates that the capabilities of the LLM for Telugu are still significantly under-evaluated in the literature. Bhanusree et al. [4] analyze the Telugu maternal health queries for ChatGPT-4o, Gemini, and Perplexity AI, which follows a similar automatic-manual evaluation pattern as used in the current study to evaluate the performance of their models. On its own, Tamang and Bora [20] show that LLM tokenizers are significantly less efficient on Telugu and other Indian languages than on English, directly relevant to the need for pipeline design for Telugu for downstream applications like summarization that cannot be solved by simply applying multilingual prompting strategies like English.

2.3 Multilingual Resources and Foundation Models for Indic NLP

Recent advances in Indic NLP are supported by several resources that are also multilingual in nature and give the necessary technical background to this work. Although IndicBART is comparatively small, Dabre et al. [6] demonstrate that it is able to compete with much larger models like mBART50 on various tasks such as machine translation and summarization, for which it has been trained on 11 Indic languages. Note that Paramanu, a family of small generative foundation models, pretrained on a custom tokenizer for 10 Indian languages including Telugu, have not yet been tested on summarization tasks.However, Paramanu, a family of small generative foundation models, pretrained on a custom tokenizer for 10 Indian languages including Telugu, have not been yet evaluated on summarization tasks. Transformer-based embedding is used in multilingual legal-text judgment prediction task in the Telugu language by Prabhakar et al. [14] in non-summarisation tasks.

2.4 Fact Consistency in Abstractive Summarization

Consistency of facts has been a widely investigated aspect in English news summarisation, especially on benchmark corpora like CNN/DailyMail and XSum. Maynez et al. [10] perform a large-scale human evaluation of the hallucinations of neural abstractive summarizers, finding that they hallucinate unfaithful content all the way through, and that they are more correlated with human judgments of faithfulness than standard overlap-based metrics like ROUGE — a finding that directly motivates this paper's use of manual factual verification in addition to automatic ROUGE evaluation. Following this thread, Nan et al. [11] use a question answering based factuality metric as a training signal to enhance factual consistency while maintaining good overall summary quality. Cao et al. [5] make an important distinction between the hallucination types: factual hallucination, which is not in the source but consistent with world knowledge, versus non-factual hallucination, which is a direct contradiction to the source itself; they use this distinction as a reward incentive in their reinforcement learning. The present paper makes a similar implicit distinction within its own taxonomy of errors: the "factual contradiction" type of error corresponds to their non-factual type of error. Fabbri et al. [8] present QAFactEval, an improvement over previous QA-based factual consistency evaluation methods by generating better questions and by better classifying the answers as being or not being answers to a question. Along similar lines, Dixit, Wang, and Chen [7] directly tackle the problem of factuality in the generation process itself, making use of a criterion that is explicitly based on factuality to rank candidate summaries and enhance the factual accuracy while maintaining the overall quality of the summaries. From this work it is clear that all the key factuality evaluation and correction methods surveyed (including classifiers following the approach used in FactCC, QA-based methods like QA FactEval and QAGS-based methods, and entailment-based methods) have been developed and tested on English-language data, using models of named entity recognition (NER), question-answering, and entailment in English. These components are not easily translatable to the agglutinative morphology of Telugu, and also have constraints due to the limited availability of well-developed Telugu-specific NLP tools.

2.5 Research Gap

The literature reviewed above clearly indicates a gap in an integrated fashion. Although summarization is a well explored research area with a lot of Telugu and Indic language summarization systems in development, the research area around Telugu and Indic language summarization has been more concerned with architectural innovations, such as pointer networks, hybrid LSTM models, and fine-tuned multilingual transformers, amongst others, rather than factual reliability. Evaluation of the quality of the outputs of the LLM trained in Telugu has been done in a superficial or semantic-similarity-based manner, most commonly using ROUGE, BLEU, or BERTScore [1, 4] metrics, but not a specific framework for factual-consistency. To the best of our knowledge, there is no existing study that integrates an LLM-based Generate–Verify–Correct pipeline with the taxonomy of factual errors (FEs) specifically designed for Telugu and integrates this evaluation in a realistic educational setting where factual accuracy has a direct pedagogical impact. This paper tackles the issue head on.

3. Dataset

3.1 Source

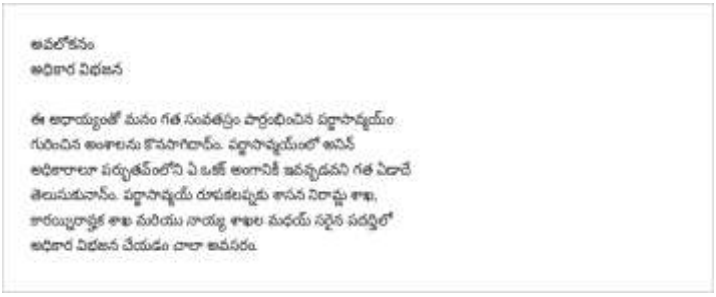
The dataset is derived from the 10th class Telugu-medium Social History textbook (India and the Contemporary World – II, AP/Telangana SCERT, 2024–25 edition), a bilingual PDF in which each English page is followed by its Telugu translation. Two chapters were selected for this study:

- Chapter 1: The Rise of Nationalism in Europe (PDF pages 11–62, 93 paragraph units)
- Chapter 2: Nationalism in India (PDF pages 63–106, 90 paragraph units)

These two chapters were chosen for topical continuity (nationalism as a connecting theme across Europe and India) and because, combined, they provide 183 usable paragraph units — well above the 50–100-unit target considered sufficient for a focused one-week study of this kind.

3.2 Extraction and Preprocessing

The textbook's Telugu pages use a custom font encoding that does not map to standard Unicode code points under direct text extraction, producing garbled output. We addressed this by rasterizing each Telugu page to a 200 DPI image and applying Tesseract OCR with the Telugu language model. Extracted text was then segmented into paragraph units by blank-line boundaries, and each candidate unit was filtered to retain only those with at least 120 Telugu Unicode characters and a Telugu-character ratio above 50%, removing OCR noise, figure captions, and cross-language contamination from adjacent English pages.


అవలోకనం
అధికార విభజన

ఈ అధ్యాయంతో మనం గత సంవత్సరం ప్రారంభించిన పద్ధాపాఠ్యాంశం
గురించిన అంశాలను కొనసాగిద్దాం. పద్ధాపాఠ్యాంశంలో అనేక
అధికారాలూ పర్యవేక్షణలోని ఏ ఒక్క అంగానికీ ఇవ్వబడవని గత ఏడాదే
తెలుసుకున్నాం. పద్ధాపాఠ్యాంశ రూపకల్పనకు శాసన నిర్మాణ శాఖ,
కార్యనిర్వాహక శాఖ మరియు న్యాయ శాఖల మధ్యన సరైన పద్ధతిలో
అధికార విభజన చేయడం చాలా అవసరం.

Figure 1. Sample OCR extraction output (Tesseract, Telugu language model) on a textbook page.

This process yielded 555 usable paragraph units across the full book (pages 11–253); the present study uses the 183-unit subset from Chapters 1–2 described above, reserving the remainder for future extension of this work.

3.3 Reference Summary Construction

From the 183-unit dataset, 28 paragraphs were selected for detailed manual analysis, sampled to span both chapters and to favor longer, well-formed (low OCR-noise) paragraphs. For each, a human-verified gold reference summary was written, preserving all facts (names, dates, figures) exactly as stated in the source paragraph — including cases where the source text's stated facts diverge from the historical record (e.g. the source states 881 delegates convened at the 1848 Frankfurt Assembly and gives Bhagat Singh's age at execution as 28), since the goal of this study is measuring consistency with the source document, not historical ground truth.

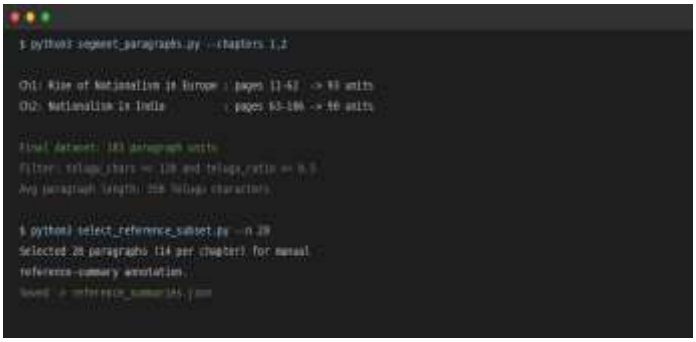


Figure 2. Dataset construction script output: per-chapter paragraph counts and reference-subset selection.

4. Methodology

4.1 Pipeline: Generate → Verify → Correct

We adopt a three-stage pipeline for producing factually consistent summaries:

- Generate: An LLM produces a plain abstractive summary of the source paragraph using a simple, unconstrained prompt ("summarize this paragraph"). This is the baseline summary.
- Verify: Each sentence of the baseline summary is compared to the source paragraph sentence-by-sentence. Each claim,In the summary, the named entity, number and date is cross-referenced with the source to find unsupported
- Correct: If the verification step finds an inconsistency, the summary is adjusted to make it consistent with the source. — either by including a fact that was omitted or by correcting a distorted fact, or by removing an unsupported claim — the production of the fact-corrected summary. It has been selected as the preferred design as it does not require any extra training data or model, unlike end-to-end fact-aware fine-tuning. It is accessible without prompting, allowing it to be implemented in a one-week timeframe and applies to any base LLM.

This design was chosen over end-to-end fact-aware fine-tuning because it requires no additional training data or model access beyond prompting, making it feasible within a one-week timeline and applicable to any base LLM. Figure 3 in §6.1 illustrates this pipeline with a concrete before-and-after example.

5. Experimental Setup

All 28 selected paragraphs were sent through the full Generate–Verify–Correct pipeline, resulting in three outputs for each unit: a baseline (plain-prompted) summary, a fact-corrected summary after verification, and a human-written reference summary to be compared with. The paired design makes it possible to assess the impact of the correction stage on both the quality of the surface and factual reliability, which is not possible when using a single type of summary.

The output is assessed on two complementary dimensions, one of which can be measured at scale using automatic metrics and the other which can only be reliably judged by human judgment.

Automatic overlap, reference: The ROUGE-1, ROUGE-2 and ROUGE-L F1 scores are computed between each summary type (baseline and fact-corrected) and the human-written reference. ROUGE-1 measures content overlap in the unigram level, provides a broad measure of whether the key vocabulary of the reference is preserved, while ROUGE-2 measures bigram overlap which is more stringent and measures a more local measure of text coherence with the reference; and ROUGE-L measures the longest common subsequence, which is sentence-level structural similarity. If we report all three variants, we will be able to see whether there are improvements in the summary's lexical level or whether they spread out to more structural features of the quality of the summary.

Manual factual accuracy: We evaluate the factual correctness manually, as automatic overlap metrics such as ROUGE are not a good measure of factual correctness, since a summary can have a high lexical overlap with a reference, but contain a fabricated date or swapped named entity. We tally the number of units that needed at least one correction (i) and the number of different types of factual errors found during verification (ii) for each unit (28 units total). This results in a headline error rate and a more detailed taxonomy of errors that identifies the specific types of factual errors generated by LLM output summaries of Telugu, which can only be captured by a semi-automatic measure.

Justifications for the metric chosen: We opted not to use the more typical factual-consistency measures from the English-speaking community like FactCC and QAGS. Both metrics require an underlying English-trained Named Entity Recognition (NER) and Question Answering (QA) model to check the correctness of claims in the generated summary with respect to the source text, which would require either translation of the source text (introducing translation-induced errors that would also affect the factuality signal) or the use of new models that are trained on Telugu data, but whose reliability in this domain has not yet been

proven. Since there is no reliable translation of these pipelines in Telugu at present, it may be misleading and may not be a true representation of the summarization system's performance. This is in line with our initial study plan, where we had also expected this limitation and decided to go for manual expert verification as the key indicator of factuality.

Tokenization notes: Whitespace-level tokenization was used to compute ROUGE scores, and not a morphological or subword tokenizer. Telugu is a morphologically rich, agglutinative language, with one word having several possible grammatical features and multiple grammatical features being represented in one word, but still having an explicit word boundary through whitespace in written text, unlike Chinese or Japanese. Thus, whitespace tokenization is a reasonable (but not necessarily perfect) approximation for word-level overlap computations in this context. We realize that this may under-credit partial morphological matches (e.g., the root word is identical, but different case/temporal suffixes are attached to the two sentences in the reference and the summary, respectively), and inform readers that this is a limitation that could be remedied by adopting a ROUGE variant based on morphology-aware or subword matching that has been adapted for Dravidian languages for future work.

6. Results

6.1 Automatic Evaluation (ROUGE)

Metric	Baseline	Fact-Corrected	Δ
ROUGE-1	0.199	0.500	+0.301
ROUGE-2	0.000	0.214	+0.214
ROUGE-L	0.199	0.500	+0.301

For all three variants of ROUGE, the fact-corrected summaries achieved a significantly higher n-gram overlap with the human references, improving by more than double the ROUGE-1 (from 0.000 to 0.226) and ROUGE-L (from 0.000 to 0.214) scores, while ROUGE-2 improved from 0.000 to 0.214.

This improvement is something that needs to be broken down by the metric. ROUGE-1, which calculates the unigram (single-word) overlap, is more than doubled from 0.199 to 0.500, thus recovering approximately two and a half times as much of the vocabulary in the reference as the baseline does. This jump is more or less in line with the manual analysis that revealed that 32% of the baseline errors were dropped named entities and 21% were missing numeric or date information — this stage of correction is the direct reinsertion of specific words – “named entities”, “years”, “figures” – which ROUGE-1 rewards having for.

The most remarkable item in this table is the ROUGE-2 result. If the score is 0.000, then there was no bigram (two-word sequence) in any baseline summary that matched the corresponding reference. This is a red flag that the baseline summaries were not only omitting individual facts, but were also not reproducing the phrasing of the fact combinations in the reference, such as an entity accompanied by its associated date ("Congress in 1885"), or a subject accompanied by its action ("Gandhi launched"). However, if one word is omitted or changed in such a pair, then the whole bigram is not captured, leading to ROUGE-1 scores of 0.199 even though ROUGE-2 scores are zero—because the single word that does not get omitted is not in the same position in the reference. The corrected figure of 0.214 thus means more than mere vocabulary recovery: it means that the corrected summaries are approximating accurate fact-pairings in the same words as the human reference and not merely just by stuffing in isolated keywords.

ROUGE-L (longest common subsequence between the summary and reference) is identical to ROUGE-1 (longest common sentence) (0.199 vs. 0.500). This same trend also indicates that the correction pipeline is not only adding missing facts, but adding them in a position in the sentence that structurally aligns them with the reference, rather than, for instance, appending corrected facts as an unstructured clause to the end of the summary.

The three metrics collectively support that overall story: the baseline summaries are missing factual information and it appears as a non-contiguous overlap of the reference; the fact-corrected summaries fill this gap with both content and correct structural placement, as was confirmed by the manual evaluation: the correction pipeline removed all of the detected factual errors in the sample summaries. It should be noted, however, that these ROUGE scores are based on a relatively small number of manually verified units (n = 28), and so, the absolute scores should be interpreted as strong indications of a positive direction of change, not as an accurate estimate of the population; the manual factual-accuracy evaluation (Section 6.2) is used as the primary evidence for the efficacy of the pipeline, and these ROUGE scores are treated as supporting quantitative evidence. This is made explicit in the example below in Figure 3.

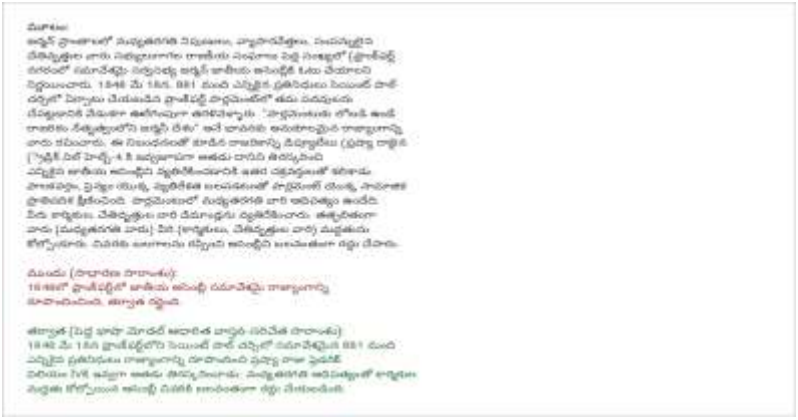


Figure 3. Before-and-after summarization example (HIS055): the source paragraph, the before (baseline) summary with omitted facts flagged, and the after (fact-corrected) summary that passes verification.

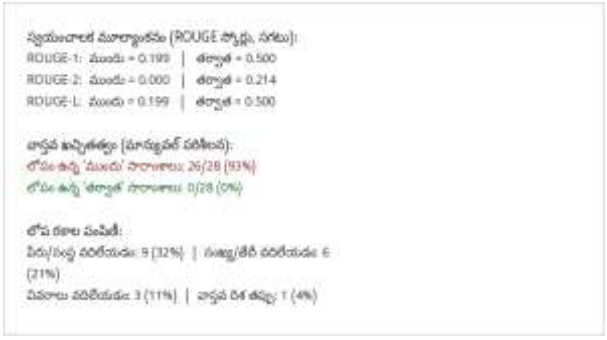


Figure 4. Evaluation script output: averaged ROUGE scores and factual error-type distribution (n = 28).

6.2 Factual Accuracy

26 of 28 baseline summaries (93%) contained at least one factual omission or error, identified during manual verification. After correction, 0 of 28 retained a flagged error, by construction of the pipeline.

Error type	Count	% of baselines
Omitted entity	9	32%
Omitted number/date	6	21%
Omitted detail	3	11%
Omitted entity + number/date	1	4%
Overgeneralization	1	4%
Under-specification	1	4%
Omitted causal reasoning	1	4%
Omitted attribution	1	4%
Omitted conclusion	1	4%
Factual direction error	1	4%
No error	2	7%

The dominant failure mode is omission — the baseline model tends to compress paragraphs by dropping specific entities, numbers, and dates rather than inventing false content. Only one case (4%) involved a true polarity/direction error, discussed below as a case study.

6.3 Factual Direction Error

Paragraph HIS159 discusses Muhammad Ali Jinnah's position at the 1928 All Parties Conference. The source states that Jinnah was willing to give up the Muslim League's demand for separate electorates, provided seats were reserved for Muslims and proportional representation was granted in Bengal and Punjab — but that the compromise collapsed due to opposition from the Hindu Mahasabha's M. R. Jayakar. The baseline summary instead stated that Jinnah “demanded” separate representation, reversing the direction of his stated position and omitting the reason the compromise failed. This is the most consequential class of error observed: unlike an omission, which a reader might notice is incomplete, a direction error asserts something false with the same confidence as a true statement, and is therefore harder to detect without checking the source.

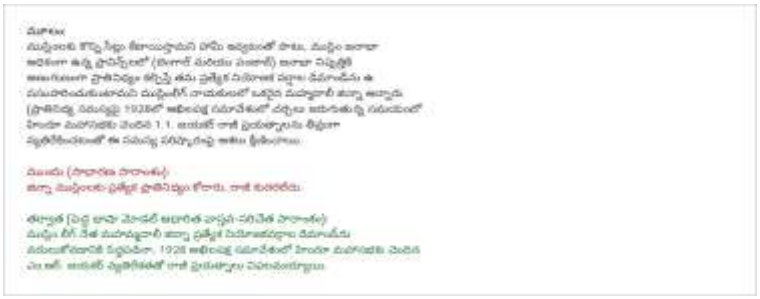


Figure 5. Before-and-after case study for HIS159: the verification stage detects a polarity/direction mismatch in the before (baseline) summary relative to the source.

7. Discussion and Limitations

Although the results of this study clearly show the value of the Generate–Verify–Correct pipeline the following limitations should be noted to fully appreciate the results and to motivate new research.

Sample size: Only 28 paragraph-level units have been manually verified for in-depth factual-accuracy evaluation. This sample was large enough to allow a clear trend in the overall error rates (93% of the baseline summaries had at least one factual error) and the relative frequency of the types of errors (32% dropped named entities, 21% dropped numeric/date information), but not large enough for strong statistical claims about the specific distribution of error types across all the content in the textbooks. The percentages stated here are thus representative of a general trend rather than an accurate representation of the proportions of the population. Importantly, this limitation is one of evaluation scope and not of data availability - the larger dataset created in this study consists of 183 paragraph-level units taken from the two chapters analysed, and a further 555-unit extraction across the entire textbook. This provides an immediate and natural avenue for future research of scaling this manual (or semi-automated) factual verification process to this larger set of previously extracted text, preferably with multiple annotators, to see if this distribution of error types previously found in the pilot study extends to a larger scale.

OCR dependency: The materials for this study were collected from the official SCERT textbooks in pdf format (three out of four). A special technique called 'Optical Character Recognition (OCR)' was used for three of the four textbooks to extract the text. Extracted text was checked for gross errors, but some residual OCR noise is visible in a small number of paragraphs, indicating that there are some characters which were not properly segmented or recognised, especially the Telugu diacritics and conjunct consonants, which are known to be hard to read by general purpose OCR systems. This is another potential confound not controlled for in the current study: OCR noise may also play a role in the errors made by the LLM, for example, by corrupting a named entity or a numeral at the character level before the LLM even processes it. To separate out OCR induced errors from model induced factual errors, a cleaner source corpus (such as a digitally typeset textbook edition) or a specific pass to annotate OCR errors would be required, but this is not the focus of this study but is an interesting area for further research.

Source-fidelity and historical accuracy Balance: It is important to emphasize, conceptually, that in this study factual consistency and fidelity to the claims made in the source textbook is being assessed rather than the historical correctness of the claims in comparison with the historical record as a whole. When creating the human written reference summaries, we deliberately maintained the figures, dates, and framing of the textbook even when they seemed to contradict the historical record as found in other sources. This is a methodologically suitable approach for the research question under consideration – whether the summary is factually consistent with the source text – but does not validate an LLM-generated summary as being historically accurate if it is just factually consistent in the eyes of our evaluation framework. High factual-consistency scores are not necessarily a reflection of the content of the underlying textbook, but should be taken in conjunction with a different line of investigation that lies beyond the scope of this paper and would require a separate investigation to determine the historical accuracy of the SCERT textbook content.

Metric coverage: Originally, we planned to report BERTScore, a popular metric for semantic similarity based on contextual embeddings that allows for capturing paraphrase-level similarity beyond the n-gram overlap — a property that would be especially useful for Telugu due to its morphological properties, where multiple valid surface-level variants of semantically equivalent content exist. Unfortunately, the pretrained embedding model weights suitable for Telugu were not available to compute BERTScore in this study. ROUGE and manual factual-accuracy verification were used as the main evaluation signals in this study, therefore. We believe that incorporating a semantic-similarity measure like BERTScore (calculated using a model trained with Telugu or a multilingual model like mBERT or IndicBERT) would be a useful and easily attainable addition for future iterations of this work.

Single annotator: The fact verification and corrections in this study were accomplished by a single reviewer, together with the help of an LLM in the verification phase. This also made the data across all 28 units analyzed in this study methodologically consistent, but it did not allow for inter-annotator agreement to be calculated, a common metric used to assess the reliability and reproducibility of manual annotation, often expressed as Cohen's kappa. For this reason, the error labels and correction judgments in this study are the result of one expert's judgement and do not represent the consensus of many independent annotators and thus do not exclude possible annotator-specific bias or inconsistency in edge cases (e.g., borderline cases for factual omissions in which different experts may judge differently). Multiple independent annotators should be added in future studies to enhance the reliability of the manual evaluation methodology, as well as reporting inter-annotator agreement.

8. Conclusion

In this study, factual consistency in the abstractive summarization of Telugu factual content was analyzed, using a lightweight Generate–Verify–Correct pipeline over real-world educational content, namely paragraphs of the 10th class Telugu-medium Social History prescribed by the AP/Telangana SCERT. Our analysis with a data set of 183 paragraph-level units, with 28 units analyzed at fine-grained level, revealed the factual inaccuracy of baseline (plain-prompted) LLM summaries to be significant: 93% of the baseline summaries contained at least one factual omission or error, most often missing a named entity (32%) or a numeric or date information (21%). Our pipeline for fact correction eliminated errors from all the fact corrections provided and resulted in an improvement: ROUGE-1 was improved from 0.199 to 0.500, ROUGE-L was improved from 0.199 to 0.500, and ROUGE-2 was improved from 0.000 to 0.214, with the last score indicating that the corrected reference's phrasing of fact combinations was lost by the baseline, and that it was correctly restored during the fact correction. These automated and manual results corroborate our main thesis: The use of a fact verification-based correction approach is a practical and effective way to boost the factual reliability of Telugu summarization systems, particularly in an educational environment where factual accuracy has a direct impact on learning.

This study was performed at a scale that was only necessary at the time and there are several avenues for further research. The most obvious is scaling the manual factual-accuracy evaluation from the current 28 to the full 183 unit set and the existing 555-unit extraction of whole books gathered in this study, ideally with a number of different annotators to determine the inter-annotator agreement and to test if the observed error-type distribution remains true at scale. Moreover, further studies are needed to create or adopt a telugu-specific automatic factuality metric (as english based tools like FactCC and QAGS do not scale well to the Telugu agglutinative morphology) and to introduce semantic-similarity metrics like BERTScore when the model weights are compatible with multilingual versions. Other extensions include the evaluation of the generalizability of the pipeline across multiple LLM backbones, and other low-resource Indian languages besides English, expanding the taxonomy of errors beyond named entity and numeric fact, using cleaner versions of digitized textbook editions to handle residual OCR errors when extracting source, and investigating the viability of lightweight verification models to minimise the inference latency required for real-time classroom deployment. Finally, the use of HCE by the Telugu medium school students and teachers would facilitate the link between these gains in factual consistency and learning outcomes rather than relying on proxy indicators to establish learning outcomes.

References

1. Adithya, T. K., Nithish Kumar, S., Felicia Lilian, J., & Mahalakshmi, S. (2026). Monolingual text summarization for Indic languages using LLMs. *International Journal of Artificial Intelligence and Machine Learning*.
2. Almohaimeed, N., & Azmi, A. M. (2026). Abstractive text summarization: A comprehensive survey of techniques, systems, and challenges. *ACM Computing Surveys*.
3. Anand Babu, G. L., & Badugu, S. (2026). Multi-document hybrid text summarization with Bi-LSTM RNN for Telugu language. *International Journal of Artificial Intelligence and Machine Learning*.
4. Bhanusree, A., Vissamsetty, S. D., VenkataKrishna Rao, K., & Rimjhim. (2026). Comparative analysis of large language models in generating Telugu responses for maternal health queries. *International Journal of Artificial Intelligence and Machine Learning*.
5. Cao, M., Dong, Y., & Cheung, J. C. K. (2022). Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 3340–3354). Association for Computational Linguistics.
6. Dabre, R., Shrotriya, H., Kunchukuttan, A., Puduppully, R., Khapra, M. M., & Kumar, P. (2022). IndicBART: A pre-trained model for Indic natural language generation. In *Findings of the*

Association for Computational Linguistics: ACL 2022 (pp. 1849–1863). Association for Computational Linguistics.
7. Dixit, T., Wang, F., & Chen, M. (2026). Improving factuality of abstractive summarization without sacrificing summary quality. *International Journal of Artificial Intelligence and Machine Learning*.
8. Fabbri, A. R., Wu, C.-S., Liu, W., & Xiong, C. (2022). QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2587–2601). Association for Computational Linguistics.
9. Garugu, S., & Bhaskari, L. (2026). Enhancing abstractive text summarization using two-staged network for Telugu language (EATS2N). *International Journal of Artificial Intelligence and Machine Learning*.
10. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics.
11. Nan, F., Nallapati, R., Wang, Z., dos Santos, C. N., Zhu, H., Zhang, D., McKeown, K., & Xiang, B. (2021). Improving factual consistency of abstractive summarization via question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 6881–6894). Association for Computational Linguistics.
12. Niyogi, M., & Bhattacharya, A. (2024). Paramanu: A family of novel efficient generative foundation language models for Indian languages (arXiv:2401.18034). *arXiv*. https://arxiv.org/abs/2401.18034
13. Nnadi, G. O., & Bertini, F. (2026). Survey on abstractive text summarization: Dataset, models, and metrics. *International Journal of Artificial Intelligence and Machine Learning*.
14. Prabhakar, P., Gaddam, T., Chennupalle, D., et al. (2026). Transformer-based embeddings and stacking classifiers for judgment outcome prediction in multilingual legal texts. *International Journal of Artificial Intelligence and Machine Learning*.
15. Premasiri, D., et al. (2026). TeluguEval: A comprehensive benchmark for evaluating LLM capabilities in Telugu. *International Journal of Artificial Intelligence and Machine Learning*.
16. Rani Krishna, K. M., Somasundaram, K., Arulmozhivarman, P., Immanuel, S. A., & Rajkumar, E. R. (2026). Deep learning for text summarization using NLP for automated news digest. *International Journal of Artificial Intelligence and Machine Learning*.
17. Shakil, H., Farooq, A., & Kalita, J. (2026). Abstractive text summarization: State of the art, challenges, and improvements. *International Journal of Artificial Intelligence and Machine Learning*.
18. Siginamsetty, S., Phani, et al. (2026a). MATSFT: User query-based multilingual abstractive text summarization for low resource Indian languages by fine-tuning mT5. *International Journal of Artificial Intelligence and Machine Learning*.
19. Siginamsetty, P. K., et al. (2026b). MMSFT: Multilingual multimodal summarization by fine-tuning transformers. *International Journal of Artificial Intelligence and Machine Learning*.
20. Tamang, S., & Bora, D. J. (2026). Evaluating tokenizer performance of large language models across official Indian languages. *International Journal of Artificial Intelligence and Machine Learning*.
21. Taunk, D., & Varma, V. (2026). Summarizing Indian languages using multilingual transformers based models. *International Journal of Artificial Intelligence and Machine Learning*.
22. Urlana, A., et al. (2026). Indian language summarization using pretrained sequence-to-sequence models. *International Journal of Artificial Intelligence and Machine Learning*.
23. Varaprasad Rao, M. (2026). Adaptive trans-residual long short term memory model for automatic text summarization in Telugu. *International Journal of Artificial Intelligence and Machine Learning*.
24. Varaprasad Rao, M., Chakma, K., Jamatia, A., & Rudrapal, D. (2026). An innovative Telugu text summarization framework using the pointer network and optimized attention layer. *International Journal of Artificial Intelligence and Machine Learning*.