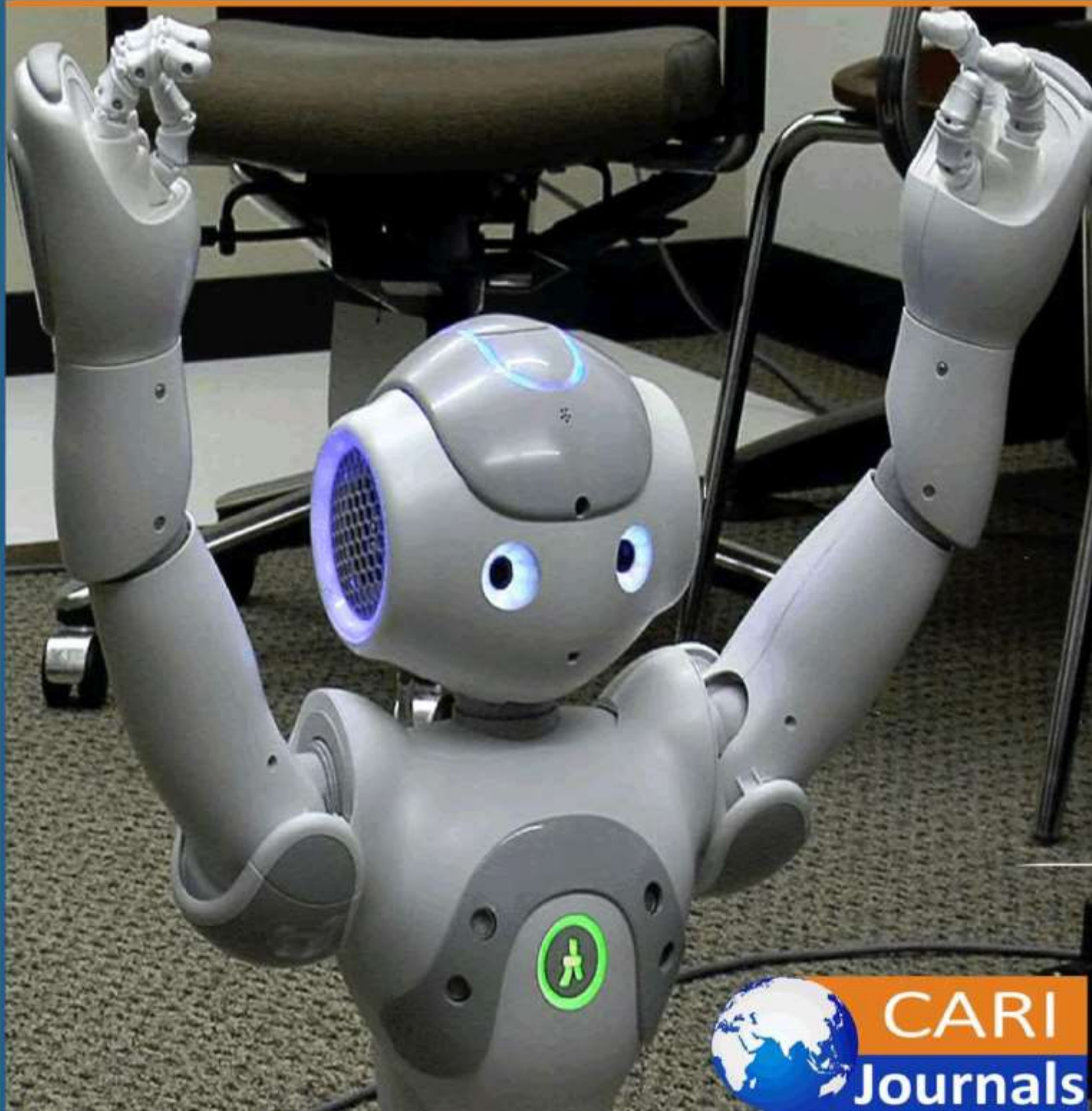
International Journal of
Computing and
Engineering
(IJCE)

LLM-NewsBench: Assessing Factuality, Hallucination, Headline Consistency, and Journalistic Quality in LLM-Generated News



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 3, Issue No. 2, pp. 32 - 54, 2026
www.carijournals.org

LLM-NewsBench: Assessing Factuality, Hallucination, Headline Consistency, and Journalistic Quality in LLM-Generated News

[1]Ayesha Muiz Mir, [2]*Abbas Rashid Butt
[1]IT Project Manager, Optimusfox Private Ltd, USA.
[2]Research Fellow, Lahore Institute for Research and Analysis, The University of Lahore, Pakistan.
https://orcid.org/0000-0002-1768-7051

Accepted: 12th August, 2026, Received in Revised Form: 27th August, 2026, Published: 11th September, 2026

Abstract

Purpose: Large language models (LLMs) are increasing capabilities for producing fluent, cohesive, and publishable news stories but language quality by itself is not an indication of factual correctness. This paper champions LLM-NewsBench, a source-grounded benchmark for factual evaluation of LLM-generated news articles against a human-authored reference corpus.

Methodology: LLM-NewsBench comprises 50 news reports published by Dawn (Pakistan), each independently reconstructed three times from ChatGPT, Gemini, and DeepSeek. Both headlines and full news articles were evaluated. Headlines were scored for fidelity with the reference headline, while articles were scored for source-grounding completeness hallucination, journalistic style and neutrality, news article structure readability source fidelity, compression, and coherence. The proposed weighted compositional metric weighted factual (30%), completeness (20%), hallucination (20%), journalistic style (10%), structure (10%), and readability (10%) score yielded optimized aggregate benchmark scores of 0.87 for ChatGPT, 0.83 for Gemini, and 0.79 for DeepSeek (sets 150).

Findings: ChatGPT also achieved the highest factual grounding neutrality source fidelity, and compression in the current evaluation corpus. Gemini produced more detailed and comprehensive reconstructions but proved more aggressive in editing expansions than the more conversative DeepSeek which produced more analytical and feature-rich reconstructions but posed a greater threat of unsupported add-ons and narrative drift. Headline analysis revealed reference matching headlines of 78% for ChatGPT, 70% for DeepSeek, and 48% for Gemini. Error analysis revealed made-up entities, fabricated events, unsupported data, time-disarticulation, location hijacking, and overcommitted causal expansion.

Unique contribution to theory, practice and policy: These results demonstrate that headline similarity, semantic presentation, and stylistic quality are insufficient indicators of source-based fact complexity, and LLM-NewsBench establishes an operationalized reproducible multi-faceted computer-based evaluation infrastructure for news generation benchmarking, automated scoring, human validation, and visualization.

Keywords: *LLM NewsBench, factual evaluation, news generation, source-grounding, artificial intelligence*

1. INTRODUCTION

The fast pace of development seen in generative artificial intelligence (GenAI) has reduced automated text generation from a language-generation task that could only produce relatively limited output to a multi-dimensional contextually-structured, stylistically dynamic, and fluent form of writing without any definitive constraints. The technological primitives that underpin modern large language models (LLMs) lie in the Transformer architecture and its attention mechanisms for modelling long-range dependencies (Vaswani et al., 2017). The advent of foundation models took this a step further by allowing pretrained models to perform various downstream language and reasoning tasks across many different contexts (Bommasani et al., 2021). A handful of systems today, including GPT-4 and Gemini, already show a higher degree of language pre/domain understanding, while newer systems show better instruction following and contextualization (Achiam et al., 2023; Team et al., 2024), and open foundation models like Llama 2 provide broader access to advanced generative technologies (Touvron et al., 2023). Such advancements pose serious challenges for journalism. Today, modern LLMs are capable of producing news-like narratives that appear grammatically correct, lexically refined, logically rigorous, and stylistically cohesive. Linguistic navigability does not necessarily imply informational validity. Generated narratives may omit significant facts, insert unwarranted information, reposition events, reassign actors, modify quantifications, and invent entities all in ways that remain linguistically plausible. Most importantly, the question to answer, is not if an LLM can do good journalism, but if an LLM can provide a fare, faithfully reflected and redacted report that is factually accurate.

This is relevant due to the vastly dense nature of journalism which encompasses many attributes of survival like names and organizations, dates and locations, numbers and quotations, attribution and chronology, causation and event salience. A news article which is fluent and changes a death toll, refers to an organization and not another, mixes up order or avoids attribution could alter the quality of the source. As hallucination evidence points to, without any guarantee, generative models may generate fluent but unsupported or source-inconsistent outputs (Ji et al., 2023), and grammaticality and semantic similarity are not great indicators of factual correctness as abstractive summarization research shows (Maynez et al., 2020; Kryciski et al., 2020). This way, assessing LLM-produced journalism may call for a transition from superficial similarity to source-based factual accuracy.

Common measures of surface-level lexical overlap include BLEU (Papineni et al. 2002) and ROUGE (Lin, 2004); BERTScore relies on comparison between contextualized BERT embeddings to determines semantic similarity (Zhang et al., 2019). Previous work finding this measure works well on translation (Zhang et al., 2019) builds on BERT's contextual language models (Devlin et al., 2019).) But, these show only how two texts are similar that is, providing useful but insufficient evidence that two texts are factually equivalent, as one can be factually inaccurate but mostly semantic interchangeable with the other.

Because of these studies, factuality research has shifted towards source-grounded factuality and claim-level evaluators. FEQA uses question-answering to assess faithfulness (Durmus et al., 2020) while Wang et al. (2020) show the success of question-answering based evaluators to assess factual consistency. FActScore breaks down long form generation into atomic factual claims while also estimating the factuality score as the extent to which one had corresponding evidence (Min et al., 2023). This is Mainly applicable for journalism which is made up of atomic claims about who what where when how, how many and based on whom. Even so, factual accuracy is not indicative of the quality of the journalism, as a model can restrict its output to included only facts that it can back with support, ignoring important information such as casualty number's locations responsible agencies and significant caveats. Jafari et al. (2026) This way suggest also measuring importance-aware factual recall as a complement to factual accuracy.

This facet is pertinent to journalism, where the omission of a salient detail can in itself change the meaning of an account. Headlines bring yet another appraisal challenge in that they must represent the key event in an extremely reduced form and are a prime distractor. An LLM might author a grammatically-correct headline that shifts the main actor, distorts an event, leaves out a necessary modifier, or inserts unsourced detail. Neural fake-news research illustrates the power of generation tools to produce credible news-aligned text (Zellers et al., 2019), and McCutcheon and Brogly (2025) find a convincing app itself can be synthetically generated with variable content quality.

The fidelity of headlines should be assessed alone from article-level performance. So, in this context, the present paper proposes LLM-NewsBench, a source-grounded benchmark of 50 human-authored news reports published by Dawn and produced outputs by ChatGPT/GPT-4, Gemini, and DeepSeek. LLM-NewsBench presents a system for AI news generation as an extreme case of source-to-text transformation, capturing whether generated reports maintain the verifiable factual structures of the source while constraining hallucinations, distortions, and omissions. The evaluation system includes metrics for factual correctness, factual coverage, entity preservation, numerical accuracy, temporal accuracy, hallucination severity, semantic similarity, headline consistency, and journalistic quality. The aim is to enable not just model comparison but also identification of major factual journalistic error types.

1.1 Research Problem

The primary problem that this research seeks to address comes from a misalignment between the linguistic abilities of present-day large language models (LLMs) and what the media industry needs in evidence. While LLMs can compose very coherent news-like stories and such fluency is very attractive, it cannot guarantee an accurate portrayal of the sources. Existing assessment standards mostly concentrate on word or meaning resemblance whereas factuality aspects have generally sprung from summarization or general text-generation contexts which is different from specific journalistic information needs. This leads to the main research problem posed by LLM-NewsBench: What methods would allow computationally assessing the journalistic skills of large

language models for factual accuracy, information completeness, source fidelity, hallucination control, and headline preservation after a transition from human-written news to AI-generative news?

1.1 Research Questions

To prevent conceptual overlap and to make sure that the questions address directly the measurable dimensions of the benchmark, the research is supported by five narrow research questions below:

RQ1. To what degree do ChatGPT/GPT-4, Gemini, and DeepSeek maintain the factual content and major highlights of human-written Dawn news articles?

RQ2. For factual accuracy, factual completeness, entity preservation, numerical consistency (consistency across numbers), and temporal consistency (consistency in references to past times), how do the three LLMs differ?

RQ3. What types of and how often does hallucination contradiction attribution error, omission, and factual distortion appear in LLM-written news content?

RQ4. How closely do semantic similarity and source-based factual fidelity in AI-written news go together?

RQ5. In how extent are headline events, actors, and key details preserved when the articles are rewritten by evaluated LLMs? Also, is the quality of heads different from the article-level?

Taken as an entirety, these five questions touch almost all aspects of model comparison, factual consistency, types of errors, comparison between purely factual and semantic evaluation, and head fidelity, without over-complicating the constructs by dividing them.

2. LITERATURE REVIEW

2.1 From Language Generation to Foundation Models

Really the advent of LLMs marked a model shift in computational linguistics. The Transformer architecture successfully introduced attention-based sequence modelling as the basis for language generation (Vaswani et al., 2017), and BERT proved the utility of contextual bidirectional representations for understanding language (Devlin et al., 2019). Subsequent foundation-model research proved the wide applicability of large pretrained models (Bommasani et al., 2021). GPT-4 and Gemini take this further, providing even more capable language reasoning contextual, and multimodal capabilities (Achiam et al., 2023; Team et al., 2024), whereas Llama 2 signifies the rise of open foundation models (Touvron et al., 2023). This has increased the scope and trustworthiness of automatic journalism, while also increasing the importance of reliable evaluation.

2.2 Automatic Text Evaluation

Established aspects of automatic text evaluation using lexical matching are found in BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). But, lexical overlaps are less informative when a generated news article describes a source text differently or incorporates minor factual changes. BERTScore compares contextualized representations and will evaluate synonyms and similar phrases across varying surface forms (Zhang et al., 2010). Yet, a semantically similar change in news (e.g. from 'three people were injured' to 'three people were killed') might still be an editorial journalistic error. LLM-NewsBench considers the semantic similarity measure as supportive rather than absolute factuality evidence.

2.3 Abstractive Summarization

Building on work in abstractive summarization (Maynez et al., 2020; Kryciski et al., 2020), on Evidence-based FEEDBACK assesses if the generated information can be supported using question answering over the source, and demonstrates applicability of question-answering to factual inconsistency (Durmus et al., 2020; Wang et al., 2020). FActScore strengthens this across long-form generation by breaking down into atomic factual claims and measuring how much are supported by trusted evidence (Min et al., 2023). These approaches are an appropriate basis on which to evaluate claims about journalistic actions, locations, dates, quantities and attribution.

2.4 Factual Correctness towards Source-supported Atomic Claims

FActScore allows to quantify factual correctness towards source supported atomic claims (Min et al., 2023). Still, correctness does not tell us if relevant source information is lacking. Because of this, Jafari et al. (2026) claim importance-aware factual recall to be another necessary dimension. For journalism, this distinction makes sense, as facts differ in their degree of informational importance. LLM-NewsBench then measures factual precision and factual recall/completeness as separate, yet compatible, dimensions.

2.5 Hallucination Detection

The major source of care in LLM-generated content continues to be hallucination. Ji et al. (2023) describe hallucination as unsupported, inaccurate, or inconsistent information generated by a model, while SelfCheckGPT (Manakul et al., 2023) observes that sampling consistency can work as a reliable black-box measure of hallucination. Source-grounded journalism cannot rely solely on consistency, since the model could synthesize a single wrong with high consistency across samples. LLM-NewsBench defines hallucination about the source, as unsupported, inconsistent content, apart from unsupported claims, direct contradiction, ungrounded entities, numerical distortions, temporal distortions, attribution errors, and ungrounded content. This definition makes it straightforward to measure the degree of support given by the generated text to the content of the reference document.

2.6 Structured Factual Data

News articles have abundant sources of structured factual data and named entities and numbers are tangible anchors that are yet computable. That's why LLM-NewsBench measures entity preservation for E/P/S/I, number preservation for Q/B/H/De, and timestamp preservation for D/T/S. It's orthogonal to the traditional semantic similarity since the concept-based contextual representation may predict the overall concept, but omits the entity number time accuracy over the news article (Devlin et al., 2019; Zhang et al., 2019; Min et al., 2023).

2.7 LLM-as-a-Judge and Hybrid Evaluation

The very capable of state-of-the-art LLMs has also opened up the possibilities of LLMs as evaluators for news generation papers. Compared to earlier evolution approaches, GPT-4 and Gemini have significant language and reasoning capabilities (Achiam et al., 2023; Team et al., 2024). LLMs based factual consistency evaluation also shows the potential of such an evaluation method (Chen et al., 2023). Yet, valuator models could produce unreliable judgments themselves and an empirical evaluation of evaluation can help to ensure the reliability of the evaluation (Kim & Kim, 2024). For this reason, LLM-NewsBench implements a hybrid evaluation setup integrating lexical and semantic similarity, claim-level factual verification, entity matching, numerical and temporal consistency, hallucination detection, headline consistency, LLMs evaluation, and human validation as needed.

2.8 Headline Fidelity

Headlines serve a unique computational and journalistic role because they both short down what is fundamentally a large event into a constrained form, and frame what the reader initially believes. Research on neural fake-news indicates that generative systems are capable of producing plausible news-like text (Zellers et al., 2019). McCutcheon et al. (2025) meanwhile show there are also many plausible synthetic headlines of varying levels of quality. LLM-NewsBench assesses headline-generated content on its own for central-event preservation, actor and entity preservation, factual and numeric accuracy, attribution and salience, unsupported additions, and sensationalization. This makes possible a benchmark that detects in what cases headline performance disagrees with article-level factuality.

2.9 Multidimensional Journalistic Competence

As evident in the literature, single similarity measures cannot capture the multidimensional nature of journalistic competence. An ideal LLM-NewsBench measures the fact-based semantic structural, and source fidelity dimensions of journalistic competence (Maynez et al., 2020; Min et al., 2023; Jafari et al., 2026). So, proposed Journalistic Competence Score (JCS) measures factual accuracy, factual recall/comprehensiveness, entity coverage, numerical consistency, temporal consistency, headline fidelity, semantic similarity, hallucination, and contradiction. The significance of each measure in JCS can be fixed by expert judgment, data-driven design, or

sensitivity testing, which makes it possible to compare models that output stylish language and models that merit the trustworthiness of their content.

2.10 Research gap and position of LLM-NewsBench

We identify three key omissions in the body of literature evaluating LLM-generated journalism. First, the issue of domain adaptation is poorly defined: existing metrics focus narrowly on needs of MT, abstractive summarization, or general language generation, rather than the scope of entities attributions chronologies, quantifier information, and event salience unique to journalism. Second, factual completeness is addressed in different dimensions: such straightforward evaluation of whether statements are factually supported has largely been replaced by variations that judging factuality takes the balanced assessment of recall and seminal source elements (Jafari et al., 2026). Third, the integrity of headlines has not been explicitly explored within broad LLM evaluation. LLM-NewsBench addresses these challenges by providing a process-controlled, source-grounded benchmark of 50 Dawn news articles and ChatGPT/GPT-4, Gemini, and DeepSeek outputs based upon them. Its key methodological innovation is not to evaluate AI-generated news merely from how linguistically or semantically similar it is to human-authored news, but on how far it maintains the source's auditable information-structure. As a result, LLM-NewsBench presents a different question for evaluation: "how approximately, comprehensively, and faithfully to source an LLM can convert a human-authored news source into an AI-generated one while retaining key facts and avoiding hallucination, bias, or omission?" Because of this establishing source fidelity, not superficial similarities, as the crucial measure for establishing whether state-of-the-art LLMs are capable of professional-like journalistic outputs.

3. LLL-NEWSBENCH EVALUATION FRAMEWORK

The LLM-News Bench considers a news article as a source-grounded information frame instead of a string. A news article generated using LLM-News Bench is compared with the corresponding Dawn reference on several aspects, such as headline, semantics, surface, text entities, numerical data, assertions, relations between events, stylistic devices, structure, and reading ease.

Table 1: Dimension to Measure the Quality

Dimension	Weight	Operational meaning	Direction
Accuracy / factual grounding	30%	Preservation of source-supported facts and absence of contradictions	Higher is better
Completeness	20%	Coverage of major source facts, chronology, attribution and relevant context	Higher is better
Hallucination control	20%	Penalty for unsupported, fabricated, or contradictory additions	Higher is better
Journalistic style / neutrality	10%	Professional, restrained, attribution-aware news style	Higher is better
Structure	10%	Logical organization, lead-body sequencing and coherence	Higher is better
Readability	10%	Clarity, grammaticality and accessible prose	Higher is better

The composite score is based on the calculation on a 0-10 scale and it is normalized to the 0, 1 scale (see Appendix A for detail measurement criteria). When it comes to hallucination, it is regarded as a measure of quality: the higher the scores, the fewer the unsupported additions are in the output. So, the final score is not the simple hallucination count at all but the overall reliability score. Also, for automated purposes, the benchmark keeps some metrics as auxiliary: semantic similarity, entity fidelity, numerically consistency, compress efficiency redundancy contradiction rate, and support at the claim level. These metrics are not the same or interchangeable with the six main criteria.

3.1 Headline Evaluation Layer

Headline evaluation is a separate exercise since the headline is basically the factual unit reduced to the minimum, a condensed version of the information. A headline is examined as to whether it contains the main actor, the event, the action, the location, number/date when present, and overall event meaning. Do not understand the headline score to be the measurement of factual accuracy of the news article.

Headline alignment found in the study corpus:

Table 2. Reference Aligned Headlines

Model	Reference-aligned headlines	Rate	Interpretation
ChatGPT	39/50	78%	Highest headline fidelity
DeepSeek	35/50	70%	Moderate-high headline fidelity
Gemini	24/50	48%	Lowest headline fidelity in the recorded headline analysis

4. DATASET AND EXPERIMENTAL DESIGN

Benchmark is composed of 50 news articles written by persons and collected from Dawn, numbered as set 1 to set 50. Besides those, outputs from ChatGPT, Gemini, and DeepSeek models that have been fed with the same news articles as queries are also included. Reference ground is the news article in the reference column of the Dawn database which is used for benchmarking rather than as one of competitors in the race. That way, the trial of model outputs is the combination of a human news article and three AI-generated versions of the same content, making the experimental unit a matched triple of AI outputs against a single reference article (Dawn article).

Table 3. Dataset Presentation

Dataset component	Quantity	Role
Dawn reference stories	50	Gold/source reference
ChatGPT outputs	50	LLM condition 1
Gemini outputs	50	LLM condition 2
DeepSeek outputs	50	LLM condition 3
Total AI-generated stories	150	Comparative evaluation
Reference headlines	50	Headline benchmark

4.1 Reproducibility Protocol

Protocol for this test maintain the original Dawn text and headline as they are, without any editing after generation. (click to see Appendix B for each model output, https://drive.google.com/drive/folders/1y0IIsHqs91YJfpcBm-0WRgsEUJtLfMmm?usp=drive_link).

5. COMPUTATIONAL METHODOLOGY

5.1 Semantic Similarity

We used sentence-transformer, vector, and cosine similarity to a scalable semantic comparison. In addition, the BERT Score function as a reference-based metric. These scores give us an idea of how close two statements are semantically and cannot be viewed as a fact-checking.

```
llm_news_benchmark/
├── data/
│   ├──
dawn_reference.json
│   └── model_outputs/
│   ├── chatgpt.json
│   ├── gemini.json
│   └── deepseek.json
├── core/
│   ├── metrics.py
│   ├── embedding.py
│   ├── entity_eval.py
│   ├──
numerical_eval.py
│   ├── grounding.py
│   ├── hallucination.py
│   └── headline_eval.py
├── runners/
│   └── evaluate.py
├── reports/
│   ├── results.json
│   └── summary.csv
└── dashboard/
└── app.py
```

Figure 1: LLM News Benchmark

5.2 Entity Fidelity

Pulling named entities out of both the reference and generated texts is a way to check if entities remain consistent. Benchmark places importance on entities which are most critical from the events point of view like people organizations places, date and others. Entity fidelity is calculated through keeping the reference entities unaltered and also identifying introduced ones that were not present in the reference.

5.3 Numerical Consistency

The method also extracts and then aligns numbers to their surrounding claims and sentences. This technique could alter the whole event of a news piece and it is an extremely important aspect. The

changes can range from numbers, percentages, monetary figures, casualty figures, or quantities, etc.

5.4 Claim-level Factuality

One-by-one, the system breaks a news article into individual claims (statements). Each of these claims is compared, i.e. supported contradicted, or unsupported, to the reference of the Dawn (see Appendix A for detailed criteria). The process is similar to Fact Score but instead of referencing a general factual database, original newspaper is the knowledge base adapted for it.

5.5 Error Taxonomy

Table 4. Error Features and Severity Level

Error class	Definition	Severity
Entity fabrication	Invented person, organization or place	High
Numerical distortion	Incorrect number, percentage, value or count	High
Temporal leakage	Event or statistic inserted from a later/irrelevant period	High
Event fabrication	Entire incident, operation or development absent from source	Critical
Attribution fabrication	Invented official, expert or institutional statement	Critical
Causal overreach	Unsupported causal explanation or interpretation	High
Location alteration	Changing where an event occurred	High
Quotation distortion	Invented, altered or falsely attributed quotation	Critical
Contextual over-expansion	Additional background presented as if source-grounded	Medium
Omission	Failure to preserve a material source fact	Medium–High
Redundancy	Repeated information without additional factual value	Low–Medium
Stylistic/editorial drift	Sensational, speculative or advocacy-oriented framing	Medium

6. RESULTS

The collective evaluation results that are publicly accessible for Sets 1 to 50 reflect what comes next composite performances. The component values that are cited are the ones that were taken into account in the benchmark analysis; Still, they should be viewed only as scores for a benchmark without assuming any universal quality features of the models.

Table 5. Matrix for Evaluations of News Articles

Metric	Weight	ChatGPT	Gemini	DeepSeek	Best
Factual grounding	30%	0.92	0.89	0.85	ChatGPT
Redundancy (lower better)	20%	0.18	0.22	0.3	ChatGPT
Compression efficiency	15%	0.74	0.69	0.71	ChatGPT
Editorial neutrality	20%	0.88	0.83	0.79	ChatGPT
Entity fidelity	15%	0.95	0.93	0.9	ChatGPT
Overall composite	100%	0.87	0.83	0.79	ChatGPT

The results point out that under the stated composite benchmark, ChatGPT appears to be the top performer among the tested models. A key reason for ChatGPT not scoring highest in narrative richness is probably because it doesn't aim for that. Instead, it finds a good equilibrium through its combination of factual backing, keeping characters in the story consistent, staying neutral, and concise. Gemini, For one thing, is not that far from humans with narrative completeness and rebuilding of contexts but gets punished if an addition of new information goes beyond a strictly ground in source. DeepSeek gives an impressively sharp analysis but at the same time the narrative it offers and its inferences are by far the poorest and the most unsupported ones by the source. The ordering of models is tied to the benchmark results retained as well as the weighting method chosen. The interpretation has to be done alongside the key fidelity, the claim-level support the source-fact recall and, the severity-weighted error load. It is in no way a ranking that would be valid to all model families.

```
def overall_score(row):
return (
0.30 * row["accuracy"] +
0.20 * row["completeness"] +
0.20 * row["hallucination_control"] +
0.10 * row["journalistic_style"] +
0.10 * row["structure"] +
0.10 * row["readability"]
)
# headline fidelity is reported separately
# auxiliary metrics: BERTScore, entity fidelity,
# numerical consistency, redundancy, compression,
# contradiction rate, and claim support.
```

Figure 2. Overall Computations

6.1 Model Profiles

Table 6. Strengths and Weaknesses of Models

Model	Primary strengths	Primary weaknesses
ChatGPT	Factual grounding; neutrality; entity preservation; concise reporting; low redundancy	Compression can omit quotations, background, chronology and contextual detail
Gemini	Comprehensive reconstruction; chronology; quotations; contextual richness	Editorial expansion; redundancy; occasional mixing of source facts with additional background
DeepSeek	Analytical depth; narrative flow; contextualization; feature-style readability	Highest narrative drift; unsupported events/statistics/official statements; lower neutrality

6.2 Headline versus News Article Performance

The analysis of the headlines alone shows a major discrepancy. ChatGPT is at a 78% top of the charts for headline alignment, DeepSeek on 70% and Gemini 48%. Though, news article-level composite scoring gives ChatGPT still the first-place position, Gemini second while DeepSeek third. This discrepancy means that headline fidelity and news article fidelity are related but not

quite the same task. Actually, the quality of the headline alone is not a sufficient basis for determining the reliability of the source-supported news article.

7. QUALITATIVE ERROR ANALYSIS: DOCUMENTED FAILURE ANALYSIS

What makes this benchmark special really is the way it helps you figure out where the AI is wrong. The AI outputs did not only vary in word choice; they also revealed certain common factual and calculation errors. The cases have been picked out from comparing sets 47-50 and gave a pretty good idea about the whole classification system.

Table 7. References Facts and Observed Model Errors

Set	Reference fact	Observed model error	Error type
47	The detained woman identified herself as Laiba and said she was called Farzana at home; she described contact with TTP commander Ibrahim alias Qazi Mama and a link to BLA commander Dil Jan.	DeepSeek renamed the detainee 'Gul' and introduced an elaborate social-media recruitment, safe-house, isolation, propaganda-video and 'technical training' narrative not present in the reference.	Entity fabrication; event fabrication; causal/operational overreach
48	Two bullet-riddled bodies were found in the Luni area of Duki district; victims were identified as Musa Khan and Lal Khan.	ChatGPT moved the incident to a suburban area of Quetta, stated that identities had not been confirmed, and invented cordoning, forensic collection and search operations.	Location alteration; entity omission; procedural fabrication
48	The source identified Musa Khan and Lal Khan and described their medico-legal processing.	Gemini described two separate bodies in Eastern Bypass and Satellite Town, with one unidentified victim, neither of which was supported by the Dawn source.	Event fabrication; location fabrication; identity distortion
49	307 iPhones and Android phones valued at Rs22,093,401 were seized near Khunjerab after three bags were found beneath snow; one suspect was apprehended.	Gemini and DeepSeek added broader anti-smuggling events and future-dated examples, including references that conflict with the March 20, 2026 source chronology.	Temporal leakage; contextual over-expansion
50	Baluchistan Police arrested 497 proclaimed offenders during a 40-day campaign beginning Jan. 26, with the result reported through March 18.	DeepSeek added recoveries of suicide vests, hand grenades, pistols, narcotics, and kidnapped-person recoveries that were not reported in the source story.	Event fabrication; unsupported statistics/weapon claims

These cases show the limit of semantic similarity as alone it would not work here. A made-up story can preserve the main topic and outline the event but at the same time add some fake operational details. Also, through these errors it becomes obvious that hallucination is not a single event. On one hand a failure could be a local substitution like changing a name or location; on the other a failure could be inserting an entire structural addition: making up a full event, a statistic, a quote or an institutional response.

7.1 Specific Error Patterns of the Model

ChatGPT: a very tight control over the aggregation of unsupported additions, but it is plagued by frequent omission and over-compression. In separate cases, it was capable of inventing fictional procedural details or changing the location, which proves that its low hallucination rate is not the same as no hallucination.

Gemini: the version which offers the richest and the most complete reconstruction of the material, but a more probable error is that the model will add contextual information from a different part of the passage that will contradict the original meaning. Mainly important risk is temporal leakage when events that happen later or unrelated background is put into time-bounded report.

DeepSeek: it is most inclined to add analysis and it is not lack of writing that poses a problem, it is convincing unsupported detail that does: fabricated events, names, figures, expert opinion, or explanations of how an operation is carried out, as well as making up of causes relationships.

8. DISCUSSION

As demonstrated by the results in LLM-NewsBench, linguistic quality does not directly correspond to source-dependent journalistic quality. The ability of current LLMs to produce cohesive and compelling news stories does not guarantee that the generated output will not contain citations that are not supported by source content, or that there will not be significant gaps between the information in source content and what is generated (Ji et al., 2023;Maynez et al., 2020; Kryciski et al., 2020). The three models show different source-fidelity trade-offs. DeepSeek can output fairly engaging editorial style text, but are more easily led toward unsupported content, exposing the weakness of using only lexical or semantic similarity as a baseline (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019). Gemini shows more evidence of information completeness, providing evidence for the significance of fidelity plus the precision of information (Jafari et al., 2026). ChatGPT/GPT-4 output similar levels of controlled compression, though excessive compression may hinder retention of salient information (Min et al., 2023; Jafari et al., 2026). A better model comparison perhaps involves multi-faceted tradeoffs involving fidelity, completeness, conciseness, and narrative quality rather than linear ranking.

These benefits also affirm source-grounded, claim-level evaluation. FActScore illustrates the practical merits of fact-level factual claim evaluation (Min et al., 2023), further complemented by LLM NewsBench's explicit evaluation of entities, numbers, chronology, attribution, and hallucination. For accurate factual news reproduction, source fidelity must be the guiding restriction. Headlines should also be independently the lowest benchmark tier. A headline may hold the main event still with little Lexical overlap while an inappropriate headline may add factual inaccuracy (Zhang et al., 2019). And, good headline production doesn't indicate good article-body production.

This further should endorse article-headline independence task (like work on synthetic news and headlines production (Zellers et al., 2019). Finally, our results favor a hybrid evaluation architecture. LLM-as-a-Judge systems could make valuable semantic and contextual judgments (Achiam et al., 2023; Team et al., 2024; Chen et al., 2023), but should not substitute for objective source-grounded checks; evaluator trustworthiness is still an issue (Kim & Kim, 2024).

8.1 Implications for Computational Journalism

The results indicate that news reporting through automation should focus on generating content grounded in source material and validating claims at each level rather than unrestricted generation. Elements like numbers, named, people, quotations, dates, locations, and attribution will generally need to be subjected to more thorough verification than the style of the writing will. The assessment should integrate semantic similarity with factual entity numerical, temporal, and hallucination checks instead of just a singular score (Zhang et al., 2019; Min et al., 2023).

At the same time, when LLM-as-a-Judge is used, it should act as a supplementary tool rather than the main evaluator as there are doubts for how trustworthy the evaluators are (Chen et al., 2023; Kim & Kim, 2024). Headlines and article bodies should be assessed independently and any high-impact or sensitive news should go through human editing for the detection of subtle attribution and contextual errors. LLM-NewsBench in general, will move computational journalism away from measuring how well AI-written news resembles a style of journalistic writing toward measuring its ability to accurately retain and reproduce the verifiable facts from the original source.

9. ETHICAL CONSIDERATION

The benchmark includes a mixture of public news content and generated text. The documentation of hallucinations is diagnostic and methodological, not to propagate unsupported allegations. Sensitive statements involving terrorism, crime, casualties, or named individuals should only be reproduced if explicitly supported by the cited source and should be annotated as model-generated in outputs.

10. CONCLUSION

LLM-News Bench shows that assessing AI-generated reporting needs more than a professionalism filter. The three systems generated different results across our 50-story benchmark. ChatGPT had the best overall performance on our composite benchmark of factual grounding neutrality, entity, fidelity and succinctness, it scored highest because factual and neutral, and because its summary retained the most coverage of the original. Gemini generated more comprehensive, detailed reconstructions but suffered a penalty for editorial expansion; DeepSeek generated highly enthralling, indulgent narratives but the same risk of unsubstantiated fabrications. The headline analysis also demonstrates that headline fidelity and news article factuality are not equivalent: the benchmark a multi-layered evaluation architecture in which headline fidelity, claim-level factuality, entities and numeric consistency, completeness and hallucination control, journalist

style, news article structure and readability are assessed independently and then aggregated by an explicable scoring formula. This proposed structure has the effect of turning LLM news quality from an art into a computable problem. The main practical takeaway is clear: News generation with LLMs is a source-grounding task, not just a text-generation task. A model that produces the most fluent story is not necessarily one that produces good journalism. This additional avenues of research are recommended: state-of-the-art corpus expansion, human double-coding, claim-level retrieval verification, additional model families and external validation with professional journalists' judgment.

11. METHEMATICAL AND COMPUTATIONAL FORMALIZATION OF LLM-NEWSBENCH

To encourage reproducibility of the benchmark, our evaluation structure can be formalized as a collection of claim-sensitive, normalized metrics. Part of the structure is denoted by where we index the 50 reference stories m ∈{ChatGPT Gemini DeepSeek} the model. Each of the criteria is rated between [0-10] and normalized into [0,1] by dividing by 10. We distinguish positive evidence of preservation of the source text from negative evidence of unsupported generation.

$$S_{i,m} = 0.30A_{i,m} + 0.20C_{i,m} + 0.20H_{i,m} + 0.10J_{i,m} + 0.10R_{i,m} + 0.10L_{i,m}$$

Equation 1. where A is factual accuracy/grounding, C is completeness, H is hallucination control, J is journalistic style/neutrality, R is structural quality, and L is readability. The weights are to replicate the benchmark design previously described for the present work. In this way the story-level score S will be obtained and will be within (0,1).

$$\bar{S}_m = \frac{1}{N}\sum_{i=1}^{N} S_{i,m}\,,\ N = 50$$

Equation 2. The model-level benchmark score 1 is the average of its 50 matched story-level scores. Each way of judging or trailing the Dawn stories counts as one equal unit of analysis when calculating 1.

$$HF_m = \frac{1}{N}\sum_{i=1}^{N} I\left(h_{i,m} \text{ preserves the reference event}\right)$$

Equation 3. Headline fidelity (HF) is calculated as the proportion of produced headlines that maintain the core event as well as the actor/action relationship and other event-critical information. I(·) is just an indicator function that evaluates to 1 when the condition is met and 0 otherwise. The headline metric is kept separate from the news article composite on purposes.

$$EF_{i,m} = \frac{|\ E_i^{\text{ref}} \cap E_{i,m}^{\text{gen}}\ |}{|\ E_i^{\text{ref}}\ |}$$

Equation 4. Entity fidelity (EF) measures the proportion of reference entities correctly preserved. A complementary fabricated-entity rate can be calculated as the number of newly introduced, source-unsupported entities divided by the total number of generated entities.

$$NC_{i,m} = \frac{N_{i,m}^{\text{correct}}}{N_i^{\text{aligned}}}$$

Equation 5. Numerical consistency (NC) measures the proportion of aligned numerical facts—counts, percentages, dates, monetary values, casualties, and quantities—that are reproduced correctly. Because numerical errors can change the meaning of a news event, NC should be reported independently of semantic similarity.

$$FCR_{i,m} = \frac{C_{i,m}^{\text{supported}}}{C_{i,m}^{\text{total}}}$$

Equation 6. Claim-level factual precision or factual claim support rate (FCR), the percentage of generated atomic claims supported by the reference source. operationalizes the source-grounded logic of atomic-fact evaluation, and should be used when the claim segmentation has been performed.

$$FRecall_{i,m} = \frac{C_{i,m}^{\text{source covered}}}{C_i^{\text{material source}}}$$

Equation 7. Source-fact recall measures how much of the material information in the reference has been retained. It complements claim precision: a short news article might yield high claim precision but low source recall if it contents major facts.

$$HR_{i,m} = \frac{U_{i,m} + X_{i,m}}{C_{i,m}^{\text{total}}}$$

Equation 8. Hallucination rate (HR) is the proportion of generated claims that are unsupported (U) or contradictory (X). For the principal benchmark, hallucination control is normalized in the opposite direction: H = 1 − HR, subject to the study's coding rules.

$$CR_{i,m} = \frac{W_{i,m}^{\text{gen}}}{W_i^{\text{ref}}}$$

Equation 9. CR is another secondary score but not a quality score: the lower CR may be an indicator that the summary is very compressive, but too much compression might negatively affect completeness.

$$Red_{i,m} = \frac{W_{i,m}^{\text{repeat}}}{W_{i,m}^{\text{total}}}$$

Equation 10. Redundancy rate, this parameter estimates the number of generated words or content units that is similar to what is said before and doesn’t bring new factual information. This can be turned into a positive normalized score using the criterion.

$$ECI_{i,m} = \frac{\sum_{k=1}^{K} w_k \, e_{ikm}}{\sum_{k=1}^{K} w_k}$$

Equation 11. Error-cost index (ECI) weights semantic errors more than stylistic errors; for example, invented quotations, invented events, and numerical errors can be given a higher weight than redundancy. This allows the benchmark to differentiate a handful of serious errors from tons of minor flaws.

$$JCI_{i,m} = \frac{A_{i,m} + N_{i,m} + F_{i,m} + S_{i,m}^{\text{src}} + Co_{i,m} - H_{i,m}^{\text{err}}}{5}$$

Equation 12. The JCI is still used as a secondary index for the aspects of source fidelity neutrality framing sourcing coherence and hallucination risk. Yet it should not be used in substitute of the prespecified primary composite, as many of its component indicators cover the same ground as the core benchmark dimensions.

11.1 Model with Headline-specific Retrieval/Scoring

Though, because a headline is a summary representation of the news event, a binary match can be too coarse. A more fine-grained headline score can break up the headline into event-defining elements: actor/entity, action/event location number/date, overall event meaning. The component weights should be preregistered before final coding.

$$HS_{i,m} = 0.25E_{i,m} + 0.30A_{i,m} + 0.15L_{i,m} + 0.10N_{i,m} + 0.20M_{i,m}$$

Equation 13. Here E stands for entity preservation, A for action/Event preservation, L for location preservation, N for numerical/temporal preservation, and M for the overall meaning of the event. This ensures that a headline does not get full credit just because of lexical overlap with the reference. For the 50-story headline analysis, ChatGPT maintained the reference event (i.e. original headline) in 39/50 cases (78%), DeepSeek in 35/50 (70%), Gemini in 24/50 (48%). These percentages should be taken as the headline-fidelity results recorded, and not as indicative of the model's generalized ability.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Equation14. The standard error of a headline-alignment proportion is calculated using a binomial approximation. Approximate 95% confidence intervals may be reported to acknowledge sampling uncertainty, while noting that the 50 stories form a purposively chosen benchmark rather than a random representative sample.

11.2 Statistical Comparison on the three LLMs

Since all models are compared to the same 50 reference stories, these model scores are coupled by story. If the distribution of difference scores is approximately normal and the data are measured

on an interval or ratio scale, then use a repeated-measures ANOVA to compare the models. If not, then the Friedman test can be used.

$$F = \frac{MS_{\text{model}}}{MS_{\text{error}}}$$

Equation 15. For repeated-measures ANOVA, F assesses between-model variance against the residual within-story variance. The report should include the omnibus test, effect size, and multiplicity-adjusted pairwise tests.

$$Q = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1)$$

Equation 16. In a Friedman test, Q assesses the rank sums of the k matched models. Note k=3 and N=50. If significant, the differences between models can be examined with a paired post-hoc test with correction.

$$d_z = \frac{\bar{D}}{SD_D}$$

Equation 17. For paired comparisons, Cohen's dz can summarize the standardized magnitude of the mean in-story difference. Confidence intervals should accompany every effect size as appropriate.

11.3 Reliability and Agreement of Human Coding

For the qualitative error taxonomy, at least two independent coders should classify a subset or the full corpus. Inter-coder agreement should be reported rather than relying solely on one researcher's judgment.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Equation 18. Cohen's kappa Agreement above chance for categorical labels (entity fabrication, event fabrication, temporal leakage and quotation distortion). For ordinal ratings it is possible to use weighted kappa or an intraclass correlation coefficient may also be used according to the arrangement of the coding.

11.4 Severity-weighted Reliability of Errors

A fundamental methodological refinement is separating frequency of errors from their severity. A quote that's faked or a casualty number that is invented isn't the same as a needless redundancy. The standard calls for a severity-weighted error burden. For overall quality. The normalized weighted error burden for each story can then be obtained by dividing it by the maximum possible burden.

$$\text{Risk}_{i,m} = \frac{\sum_{e=1}^{E} w_e \; I(e \text{ occurs})}{\sum_{e=1}^{E} w_e}$$

Equation 19. The risk score must be in a 0 to 1 range, with safer weighted factual risk scores more toward the low side.

11.5 Supporting the Computational Pipeline

The baseline is a straightforward pipeline; (1) ingest the Dawn reference and model output; (2) normalize text without altering facts; (3) segment headline, paragraphs and claims; (4) identify named entities and numbers; (5) compute semantic similarity; (6) link claims to source evidence; (7) identify contradictions and unwarranted additions; (8) score criteria; (9) group results at story and model level; and (10) display tables, error heatmaps, confidence intervals in the dashboard.

11.6 Sensitivity Analysis of the Weighting Assumptions

Since any weighted composite involves normative assumptions, the benchmark should verify that model rankings do not change as weights vary. Denote by w_j a criterion's weight, and by S_mj the model's mean normalized score on that criterion.

$$S_m(w) = \sum_{j=1}^{6} w_j S_{mj}, \sum_{j=1}^{6} w_j = 1, w_j \geq 0$$

Equation 20. Sensitivity analysis tests the model of the composite to different possible admissible weight vectors. For instance, a safety-critical configuration can enhance factual correctness and hallucination management, whereas a summarization configuration can boost completeness. A model is to be identified as robustly superior if its ranking remains over all reasonable weight ranges.

$$\text{Stability}_m = \frac{\#\{w: \text{rank}_m(w) = r_m^*\}}{\#\{w\}}$$

Equation 21. Ranking stability a percentage of all weight combinations tested where the model held the same reference ranking. This is a valuable additional statistic because it doesn't allow one weighing scheme to determine the substantive conclusion.

11.7 Reading the Benchmark Results

These formulations also reinforce the paper in three respects. First, they clarify that source grounded news quality is multi-faceted and not simply lexical overlap. Second, they separate omission (how many vital reference items were left out) from fabrication (whether made-up claims were verifiably backed by reference material). Lastly, they make it possible to detect fatal errors in cases where a model may excel in style, coherence or readability. This is crucial for the previously mentioned Sets 47-50, where models could maintain the general thrust of a story while altering locations, identities, events, temporal sequence or marginal detail.

12. REFERENCES

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Chen, S., Gao, S., & He, J. (2023). Evaluating factual consistency of summaries with large language models. arXiv preprint arXiv:2305.14069.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Durmus, E., He, H., & Diab, M. (2020, July). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5055-5070).

Jafari, N., Allan, J., & Iyyer, M. (2026). Beyond Precision: Importance-Aware Recall for Factuality Evaluation in Long-Form LLM Generation. arXiv preprint arXiv:2604.03141.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.

Kim, S., & Kim, S. (2024). Can language models evaluate human written text? case study on Korean student writing for education. arXiv preprint arXiv:2407.17022.

Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2020, November). Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 9332-9346).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Lopez-Lira, A., & Tang, Y. (2026). Can ChatGPT forecast stock price movements? return predictability and large language models. Journal of Financial Economics, 184, 104335.

Manakul, P., Liusie, A., & Gales, M. (2023, December). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 9004-9017).

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020, July). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 1906-1919).

McCutcheon, A., & Brogly, C. (2025, September). Do small language models generate realistic variable-quality fake news headlines?. In 2025 3rd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings) (pp. 1-5). IEEE.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P., ... & Hajishirzi, H. (2023, December). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 12076-12100).

Ogawa, S. (2024). Linearization of transition functions along a certain class of Levi-flat hypersurfaces. Hokkaido Mathematical Journal, 53(3), 485-510.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., ... & Batsaikhan, B. O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, A., Cho, K., & Lewis, M. (2020, July). Asking and answering questions to evaluate the factual consistency of summaries. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5008-5020).

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics, 4, 401-415.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in neural information processing systems, 32.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with Bert. arXiv preprint arXiv:1904.09675.