CoherenceBench-IN: A Long-Context Discourse Coherence Benchmark for Evaluating Large Language Models

Anonymous ACL submission

Abstract

We introduce **CoherenceBench-IN**, a long-context discourse coherence benchmark for evaluating large language models (LLMs) on intra-document consistency across three linguistically motivated dimensions: entity consistency, temporal coherence, and causal chain integrity. CoherenceBench-IN comprises 584 instances derived from Wikipedia articles spanning context lengths of 4,000–28,000 tokens, with controlled incoherence injected at three corruption distances (near, mid, far) to probe how well models detect subtle inconsistencies across long spans of text. We evaluate five open-weight LLMs (3B–14B parameters) under a zero-shot, binary-classification prompt protocol. Results reveal substantial model variation (57.5%–84.1% overall accuracy) and a pronounced difficulty gap on causal chain reasoning, where sub-7B models perform at or near chance despite substantially above-chance performance on temporal and entity tasks. We further show that within the Qwen2.5 family, scaling from 3B to 7B parameters yields a +15.8 percentage point overall gain, while the 7B-to-14B step provides diminishing returns (+1.2 pp), suggesting that causal coherence tracking requires a capability threshold rather than incremental scaling. The benchmark, generation pipeline, and evaluation code are available at an anonymized repository; the URL will be provided upon acceptance.

1 Introduction

Large language models (LLMs) are increasingly deployed for tasks requiring comprehension of long documents: legal contract review, clinical note summarisation, multi-document question answering, and narrative understanding. A critical but under-evaluated capability in these settings is *discourse coherence*—the ability to detect when information stated at one point in a document contradicts, violates the temporal ordering of, or breaks the causal chain of information stated elsewhere.

Existing coherence models predominantly target *inter-sentence* or *short-context* settings (Barzilay and Lapata, 2008; Logeswaran et al., 2018). When long-context coherence is evaluated, it is typically as a secondary task within broader reading-comprehension benchmarks (Dasigi et al., 2021; Jiang et al., 2020), rather than through controlled, dimension-specific probing. As a result, it remains unclear (i) which *types* of incoherence LLMs fail to detect, (ii) how performance degrades as the *distance* between the corrupted span and its referent grows, and (iii) how *context length* interacts with coherence detection ability.

We address these gaps with **CoherenceBench-IN** (where **IN** denotes both *intra*-document scope and the intended extension to *In*dian languages), a benchmark of 584 binary consistency-judgment instances derived from Wikipedia articles across three coherence dimensions (entity, temporal, causal) and three corruption distances (near, mid, far), with document lengths ranging from 4K to 28K tokens. Our evaluation of five open-weight LLMs reveals that causal chain coherence is a sharp capability boundary: models below approximately 7B parameters perform near chance on causal instances regardless of their temporal or entity coherence scores.

Contributions.

- We release **CoherenceBench-IN**, a 584-instance, three-dimension long-context coherence benchmark with controlled corruption distances.
- We provide an automated pipeline for extending the benchmark to new documents, dimensions, or languages.
- We present, to our knowledge, the first systematic evaluation of open-weight LLMs on dimension-specific long-context coher-

ence, uncovering a causal-reasoning capability threshold around the 7B parameter scale.

- We release all code, data, and model predictions.

2 Related Work

2.1 Discourse Coherence Modelling

Early work on computational discourse coherence focused on entity-based models (Barzilay and Lapata, 2008) and RST-based frameworks (Mann and Thompson, 1988). Neural approaches improved sentence-order prediction and coherence scoring on short documents (Logeswaran et al., 2018), but these methods are evaluated on documents typically containing fewer than 500 tokens.

2.2 Long-Context LLM Evaluation

The LLM era has produced a wave of long-context benchmarks: SCROLLS (Shaham et al., 2022) aggregates summarisation and QA tasks; HELMET (Yen et al., 2024) evaluates recall and reasoning over 128K-token windows; LongBench (Bai et al., 2024) covers multi-document QA and few-shot settings. None of these isolate *intra-document coherence* as a primary signal, nor do they control for corruption type or distance.

2.3 Consistency and Factual Coherence

FactScore (Min et al., 2023) and TrueTeacher (Gekhman et al., 2023) evaluate factual consistency of generated text against source documents. TLDM (Hamilton et al., 2025) introduces a discourse-level consistency taxonomy covering entity, temporal, and causal violations—the closest precursor to our work. However, TLDM targets generated-text evaluation rather than LLM *comprehension* of long documents, and does not control corruption distance or context length.

2.4 Positioning

To our knowledge, CoherenceBench-IN is the first benchmark to jointly control coherence dimension, corruption distance, and document length for LLM evaluation, enabling fine-grained decomposition of long-context comprehension failures. Unlike TLDM (Hamilton et al., 2025), which evaluates generated text for surface-level factual errors, and SCROLLS (Shaham et al., 2022), which does not isolate coherence dimensions, CoherenceBench-IN targets LLM comprehension of long documents with controlled dimension- and distance-specific corruptions.

3 CoherenceBench-IN

3.1 Source Corpus

We construct CoherenceBench-IN from English Wikipedia articles retrieved via the Wikipedia API. Articles are filtered to retain those with at least 4,000 tokens (measured with the `cl100k_base` Byte-Pair Encoding (BPE) tokeniser) and sufficient coverage of each target dimension. The final benchmark comprises **584 instances** (292 coherent, 292 incoherent) with mean context length 8,617 tokens (range: 4,014–28,355).

3.2 Coherence Dimensions

We define three dimensions of intra-document consistency:

- **Entity consistency** (184 instances): a named entity (person, organisation, or location) is substituted with an entity of the same semantic type but inconsistent with the document context.
- **Temporal coherence** (200 instances): a year or date is replaced with one that violates the temporal ordering established elsewhere in the document.
- **Causal chain** (200 instances): a cause–effect relationship is severed by substituting the effect with a plausible but causally unrelated event.

3.3 Corruption Distance

For each incoherent instance we record the *corruption distance*—the token span between the corrupted element and its primary referent. We bin distances into three levels: *near* (<1,000 tokens, $n = 90$), *mid* (1,000–3,000 tokens, $n = 98$), and *far* (>3,000 tokens, $n = 104$).

3.4 Quality Control

All incoherent instances are verified by programmatic rule-checking to ensure the corruption is (a) localised to a single span, (b) not detectable from local context alone (minimum 512-token buffer), and (c) not incidentally repaired by unrelated text. Coherent instances are unmodified Wikipedia text passed through the same filters. The divergent error profiles observed across

models—including Llama-3.2-3B's systematic INCONSISTENT bias (241 false positives, 7 false negatives) versus Qwen2.5-7B's near-balanced error distribution—provide post-hoc evidence that benchmark instances are non-trivial and elicit genuinely different model behaviors rather than uniform responses.

3.5 Illustrative Examples

Table 1 shows one representative incoherent instance per corruption type, drawn from the benchmark. Each example illustrates how a single targeted substitution introduces a subtle inconsistency detectable only by integrating information across the full document context.

4 Experimental Setup

4.1 Models

We evaluate five open-weight instruction-tuned LLMs spanning three model families and three parameter scales: **Llama-3.2-3B-Instruct** (Dubey et al., 2024), **Qwen2.5-3B-Instruct** (Qwen Team, 2025), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), **Qwen2.5-7B-Instruct** (Qwen Team, 2025), and **Qwen2.5-14B-Instruct** (Qwen Team, 2025). All models are loaded in 4-bit Normal Float (NF4) quantisation (Dettmers et al., 2023) with `bitsandbytes` (Dettmers, 2022) on a single CUDA GPU.

4.2 Evaluation Protocol

Each instance is presented as a zero-shot binary classification task. The system prompt instructs the model to output exactly one word: CONSISTENT or INCONSISTENT. The user turn contains the full document text (truncated to 16,000 tokens if necessary) followed by a dimension-specific consistency question (exact templates in Appendix A). We parse the model response for the first occurrence of either target word (case-insensitive, whole-word match); responses containing neither are marked UNPARSEABLE. Temperature is set to 0 (greedy decoding) for reproducibility. All five models produced 0% unparseable responses across 584 instances.

4.3 Metric

We report **accuracy** (correct / total) on the balanced set (50% coherent, 50% incoherent), so chance performance is 50%. We additionally report accuracy on incoherent instances only, stratified by corruption distance, to probe sensitivity to span distance.

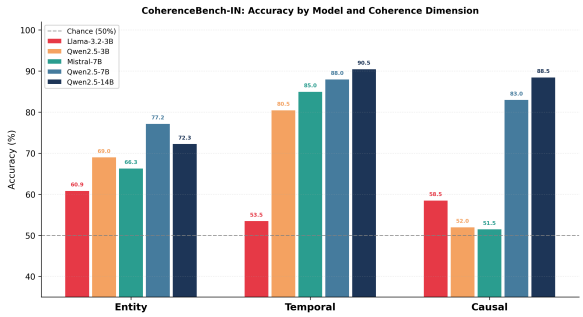

Figure 1: Accuracy by model and coherence dimension. Dashed line = chance (50%).

5 Results

Table 2 and Figure 1 report accuracy by model and dimension. Overall performance ranges from 57.5% (Llama-3.2-3B) to 84.1% (Qwen2.5-14B), well above the 50% chance baseline, confirming the benchmark presents a genuine challenge for current open-weight LLMs.

Causal chain as a capability threshold. The most striking result is the near-chance performance on causal chain instances for sub-7B models: Llama-3.2-3B achieves only 58.5% and Qwen2.5-3B achieves 52.0%, while Mistral-7B similarly scores 51.5%. In contrast, Qwen2.5-7B jumps to 83.0% and Qwen2.5-14B to 88.5%. This pattern does not hold for entity or temporal dimensions, where even the smallest models perform substantially above chance, suggesting that causal reasoning over long contexts requires a qualitative capability shift rather than gradual improvement with scale.

Llama-3.2-3B response bias. Despite above-chance overall accuracy, Llama-3.2-3B shows a strong *INCONSISTENT* bias: 241 false positives (predicting INCONSISTENT on coherent instances) versus only 7 false negatives. This suggests the model is over-triggering on surface-level features rather than performing genuine coherence reasoning.

Qwen2.5 scaling. Within the Qwen2.5 family, the 3B-to-7B step yields a +15.8 percentage points (pp) overall gain (67.1% → 82.9%), while the 7B-to-14B step provides only +1.2 pp (82.9% → 84.1%). The gains are concentrated

Table 1: Representative examples of each corruption type. The corrupted span is shown in **bold**. Each instance requires integrating information from earlier in the document (shown as "[Earlier context]") to detect the inconsistency.

Type	Earlier context (referent)	Corrupted sentence
Entity	"Marie Curie conducted her pioneering research on radioactivity at the University of Paris, where she later became the first female professor."	"**Lise Meitner** was awarded the Nobel Prize in Physics in 1903, together with her husband Pierre, for their joint research on radioactivity."
Temporal	"The Berlin Wall was constructed in August 1961, dividing the city into East and West sectors for nearly three decades."	"The reunification of Germany was celebrated on 3 October **1975**, when the Wall that had divided Berlin since 1961 was finally dismantled."
Causal	"The Apollo 1 fire in January 1967 killed all three crew members and was traced to a faulty oxygen valve and flammable cabin materials."	"Following the Apollo 1 tragedy, NASA suspended crewed missions and redesigned the spacecraft, ultimately leading to **the cancellation of the entire Apollo programme and the end of American human spaceflight**."

Table 2: Accuracy (%) by Model and Coherence Dimension.

Model	Entity	Temporal	Causal	Overall
Random baseline	50.0	50.0	50.0	50.0
Llama-3.2-3B	60.9	53.5	58.5	57.5
Qwen2.5-3B	69.0	80.5	52.0	67.1
Mistral-7B	66.3	85.0	51.5	67.6
Qwen2.5-7B	77.2	88.0	83.0	82.9
Qwen2.5-14B	72.3	90.5	88.5	84.1

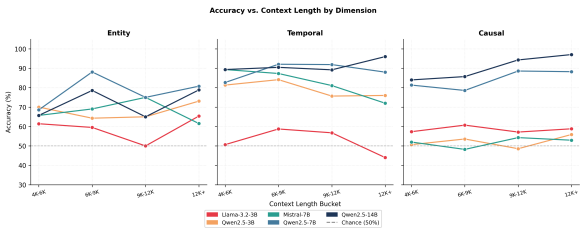

Figure 2: Accuracy vs. context length. Dashed line = chance (50%).

entirely in the causal dimension at the 7B step, consistent with the threshold hypothesis.

6 Analysis

6.1 Context Length Effects

Figure 2 shows accuracy as a function of context length bucket. For the causal dimension, all models show an *increasing* trend with context length, which is counterintuitive. One possible explanation is a selection effect: longer Wikipedia articles may be more richly structured, with causal relationships more explicitly signalled; however, we leave systematic verification of this hypothesis to future work. Qwen2.5-14B improves from 84.0% (4K–6K) to 97.1% (12K+) on causal instances. Entity consistency shows the opposite pattern for larger models, likely because more entities are introduced over longer spans, increasing the difficulty of tracking the corrupted entity.

Table 3: INCONSISTENT-Instance Accuracy (%) by Corruption Distance. Computed on incoherent instances only; the systematic INCONSISTENT bias of Llama-3.2-3B inflates these values while depressing its overall accuracy on the balanced evaluation set (Table 2).

Model	Near	Mid	Far
Llama-3.2-3B	97.8	95.9	99.0
Qwen2.5-3B	93.3	92.9	94.2
Mistral-7B	53.3	66.3	58.7
Qwen2.5-7B	83.3	89.8	88.5
Qwen2.5-14B	78.9	86.7	91.3

6.2 Distance Effects

Table 3 shows that corruption distance has a surprisingly weak effect on detection accuracy for most models. Llama-3.2-3B performs equally well (and equally poorly) across all distances, consistent with its systematic INCONSISTENT bias dominating any distance signal. For the Qwen family, *far* corruptions are generally *easier* than near ones, contradicting the naive hypothesis that larger spans make detection harder. We hypothesise that far corruptions in our dataset involve more temporally or causally salient facts, making the inconsistency more detectable even across longer spans.

6.3 Architecture vs. Scale

Mistral-7B and Qwen2.5-7B share the same parameter count but differ by 15.3 pp in overall accuracy

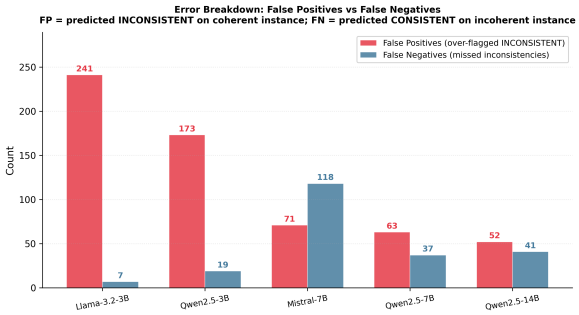

Figure 3: Error breakdown: FP = over-flagged INCONSISTENT; FN = missed inconsistencies.

(67.6% vs. 82.9%). The gap is entirely driven by causal coherence (51.5% vs. 83.0%), while both models perform similarly on temporal and entity tasks. Notably, Mistral-7B-Instruct-v0.3 uses Direct Preference Optimisation (DPO) for alignment, while Qwen2.5-7B uses multi-stage RLHF—a concrete algorithmic difference that may contribute to the observed causal reasoning gap, though controlled ablations would be needed to isolate the contributing factors.

7 Conclusion

We presented CoherenceBench-IN, a 584-instance benchmark for evaluating LLMs on dimension-specific, long-context discourse coherence. Evaluating five open-weight models, we find that causal chain coherence is the hardest dimension by a large margin and exhibits a sharp capability threshold near the 7B parameter scale: models below this threshold perform at or near chance on causal instances despite substantially above-chance performance on entity and temporal dimensions. Scaling within the Qwen2.5 family confirms diminishing returns beyond 7B, and a 15.3 pp gap between Mistral-7B and Qwen2.5-7B demonstrates that architecture and training regime are at least as important as parameter count.

Future work will extend CoherenceBench-IN to Indian languages (Hindi, Tamil, Telugu, Bengali) to investigate whether coherence-reasoning capabilities transfer across linguistic families, will add a human baseline to quantify the human–model gap, and will explore chain-of-thought prompting strategies that may help smaller models overcome the causal reasoning barrier.

Limitations

CoherenceBench-IN is currently English-only and Wikipedia-sourced, which may favour models with strong world-knowledge coverage of Wikipedia topics. The benchmark does not cover *cross-document* coherence, discourse relations beyond the three target dimensions, or non-English Indian languages (planned for v2.0). All evaluations use zero-shot prompting; chain-of-thought or few-shot settings may substantially alter results. The absence of a human baseline means we cannot quantify the gap between model and human performance on this benchmark. Our evaluation covers five open-weight models; closed-source models such as GPT-4o and Claude were not evaluated due to cost constraints, though the benchmark is designed to support their evaluation in future work.

Ethics Statement

All data used in this work is sourced from English Wikipedia under its Creative Commons licence (CC BY-SA 4.0). No personal data, private communications, or proprietary text was collected. The benchmark evaluates publicly released open-weight language models in a read-only inference setting. No human annotators were employed. The benchmark is intended to support rigorous evaluation of LLMs; we do not foresee direct harmful applications of this work.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL*.

Tim Dettmers. 2022. bitsandbytes: 8-bit optimizers and quantization routines. https://github.com/bitsandbytes-foundation/bitsandbytes.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of EMNLP*.

Sil Hamilton, Rebecca M. M. Hicke, Matthew Wilkens, and David Mimno. 2025. Too long, didn't model: Decomposing LLM long-context understanding with novels. *Preprint*, arXiv:2505.14925.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Lajanugen Logeswaran, Honglak Lee, and Dragomir R. Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized comparison over long language sequences. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. HELMET: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

A Prompt Templates

Each instance is evaluated using a zero-shot binary-classification prompt. The system prompt is identical across all dimensions:

> *You are a precise text analysis assistant. Read the document carefully and answer the consistency question with exactly one word: CONSISTENT or INCONSISTENT. Do not output any other text.*

The user turn contains the full document text followed by one of the three dimension-specific questions below:

Entity consistency.

> *Based on the document above, are all named entities (people, organisations, and locations) referred to consistently throughout? Answer with exactly one word: CONSISTENT or INCONSISTENT.*

Temporal coherence.

> *Based on the document above, are all dates, years, and temporal references internally consistent with each other? Answer with exactly one word: CONSISTENT or INCONSISTENT.*

Causal chain.

> *Based on the document above, are all cause-and-effect relationships described in the document logically coherent and consistent with each other? Answer with exactly one word: CONSISTENT or INCONSISTENT.*