Who Remembers What? Tracing Information Fidelity in Human-AI Chains

Suvojit Acharjee[1,2*] **Utathya Aich**[2,3*] **Diptarka Mandal**[2] **Asfak Ali**[2†]

[1]CSE Department, Institute of Engineering and Management, Kolkata,
[2,4]Deep Duo Foundation, Kolkata, India, [3]CNH Industrial, India

Abstract

In many real-world settings like journalism, law, medicine, and science communication, information is passed from one person or system to another through multiple rounds of summarization or rewriting. This process, known as multi-hop information transfer, also happens increasingly in workflows involving large language models (LLMs). But while summarization models and factuality metrics have improved, we still don't fully understand how meaning and factual accuracy hold up across long chains of transformations, especially when both humans and LLMs are involved.

In this paper, we take a fresh look at this problem by combining insights from cognitive science (Bartlett's serial reproduction) and information theory (Shannon's noisy-channel model). We build a new dataset of with more than 200 original source paragraph and 700 five-step transmission chains that include human-only, LLM-only, mixed human–LLM, and cross-LLM settings across a wide range of source texts. To track how meaning degrades, we introduce three new metrics: Information Degradation Rate (IDR) for semantic drift, Meaning Preservation Entropy (MPE) for uncertainty in factual content, and Cascaded Hallucination Propagation Index (CHPI) for how hallucinations accumulate over time. Our findings reveal that hybrid chains behave asymmetrically. When a human summary is refined by a language model, the final output tends to preserve meaning well, suggesting that models can improve upon human-written summaries. The code and data will be avilabe at : `https://github.com/transtrace6/TransTrace.git`.

1 Introduction

”ChatGPT Is a Blurry JPEG of the Web”

Ted Chiang
The Newyorker
February, 2023

The flow of information is a fundamental driver of techno-scientific progress in contemporary society. From policy making and industry decisions to medical advancements, educational progress, and artificial intelligence achievements, the seamless transmission of information drives nearly every aspect of innovation. As large language models (LLMs) increasingly mediate how knowledge is processed and communicated (Shahzad et al., 2025), concerns arise about the integrity and continuity of information within these systems. These concerns are magnified in multi-agent environments, where multiple LLMs or AI agents exchange and build upon shared information. With each communicative iteration, subtle distortions, omissions, or reinterpretations can lead to cumulative information loss. This phenomenon poses significant challenges for tasks that require sustained reasoning, coordination, or consensus among agents. Therefore, a systematic understanding of how LLMs retain, transform, or degrade information across dynamic multi-agent communication pathways is critical for advancing model reliability, interpretability, and resilience in complex, real-world deployment scenarios.

Information often changes as it is transmitted across people or machines. In cognitive psychology, Bartlett's (1932) “serial reproduction” experiments demonstrated that repeated retelling of a story progressively simplify content and shift meaning (Roediger III et al., 2014), a phenomenon now known as information degradation. This phenomenon is reflected in contemporary language practices, where text is frequently summarized, paraphrased, or reformulated through successive

[*]These authors contributed equally.
[†]Corresponding author: `asfakali.etce@gmail.com`

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2718–2726
December 20-24, 2025 ©2025 Association for Computational Linguistics

stages of human editing or machine-assisted processing workflows.

This phenomenon can be formally characterized as compound information loss. Let S_0 denote the original input and $T(.)$ be the transformation function applied at step i. The transformation function may represent a variety of operations, such as summarization, paraphrasing, or rephrasing, among others. These operations are lossy in nature. As a result, nuanced information can be compressed or omitted during encoding, leading to a potential degradation of the information. After n number of sequential transformations, the final output can be represented using Equation 1.

$$S_n = T_n(T_{n-1}(...(T_1(S_0)))) \qquad (1)$$

The information retention between the original sequence and the sequence after n^{th} transmission can be represented as Equation 2

$$R_n = I(S_0, S_n) \qquad (2)$$

where I measures the mutual information. Consequently, the compound information loss is defined by Equation 3

$$L_n = 1 - R_n \qquad (3)$$

LLMs, systems fundamentally designed to generate and transform language, operate on the Transformer architecture, which, while powerful, is inherently lossy in nature due to its reliance on fixed-size token representations (Vaswani et al., 2017). This compression can lead to subtle distortions or omissions during encoding and generation, especially when transformations are performed sequentially. Previous research in NLP has predominantly focused on single-step fidelity in tasks such as summarization (Maynez et al., 2020) or iterative model refinement for editing and summarization (Chen et al., 2023). In parallel, multistage workflows, spanning human-AI collaboration and multi-agent LLM communication, have become increasingly prevalent in real-world applications, such as document drafting, revision, and cascading summarization pipelines (Park et al., 2023).

Despite this trend, the multi-hop dynamics of information flow, which examine how meaning evolves, degrades, or is potentially recovered as it passes through chains of humans, LLMs, or their combinations, remain critically underexplored. These transformation chains resemble Bartlett's (1932) 'serial reproduction' effect. In the context of LLMs and hybrid pipelines, such degradation can affect semantic consistency, with implications for reliability in downstream tasks.

To address this gap, we present the first systematic investigation of information degradation across sequential transformation chains. Our main **contributions** are as follows:

- We present a systematic study of information transfer across *human-only, LLM-only, hybrid human-LLM, and cross-LLM chains*, drawing connections between cognitive theories of memory decay and contemporary NLP workflows.

- We introduce *TransTrace*, a large-scale dataset comprising *700 summarization chains* constructed from *more than 200 diverse source paragraphs*. This dataset enables controlled and systematic investigation of semantic degradation, fidelity loss, and transformation patterns across sequential summarization steps in both human and LLM-mediated settings.

- We introduce three metric such as, *Information Degradation Rate* (IDR) for calculate semantic drift per hop, *Meaning Preservation Entropy* (MPE) for uncertainty in factual retention calculation, and *Cascaded Hallucination Propagation Index* (CHPI) for persistence of hallucinated content.

- We characterize the differences in information degradation across humans, LLMs, hybrid human-LLM setups, and cross-LLM chains, uncovering patterns of compression, semantic drift, and hallucination propagation. These findings inform the design of *fidelity-aware summarization pipelines*, robust human-AI collaboration workflows, and stable multi-agent LLM systems.

2 Theoretical Foundations

Information transfer is fundamentally constrained by the nature of the communication channel. In Shannon's information theory, a message S is encoded and transmitted over a noisy channel C, resulting in a received message $\hat{S}$. The fidelity of transmission is quantified by the mutual information $I(S, \hat{S})$ which decreases with noise and transformation steps as mentioned in Equation 4. Here we assume that the sequence of transformations $S- > \hat{S}_1- > ...- > \hat{S}_n$ forms a first order Markov Chain. Under this assumption, the data processing inequality (Cover, 1999) guarantees that mutual information is non-increasing across these transformations. This framework provides a foundational understanding of how fidelity can be lost during communication, particularly through repeated transformations.

$$I(S;\hat{S}_n) \leq I(S;\hat{S}_{n-1}) \leq ... \leq I(S;S) \quad (4)$$

We emphasize that this inequality asserts non-increase, not necessarily strict decrease; strict contraction generally requires additional assumptions on the transformation channels (Raginsky, 2016; Polyanskiy and Wu, 2017). This insight extends naturally to both human and machine-mediated language processing. In modern NLP workflows, repeated paraphrasing, summarization, and re-encoding of text can compound semantic drift, factual inconsistency, and hallucinations. These effects mirror the behavior of noisy channels, where transformations act as potential points of degradation. This iterative transformation amplifies entropy and reduces the semantic overlap with the original input. For modeling convenience, we treat the sequence of transformations as a first order Markov process, where each transformed message depends only on its immediate predecessor, as in Equation 5.

$$P(\hat{S}_n \mid S) = \prod_{i=1}^{n} P(\hat{S}_i \mid \hat{S}_{i-1}) \quad (5)$$

This factorization is intended to express first order dependency for analytical clarity. This formulation in Equation 5 enables us to interpret human and LLM-driven language processing as a probabilistic communication chain. Each transformation $\hat{S}_i = T_i(\hat{s}_{i-1})$ not only contributes to possible information loss but also alters the message distribution over time. As the number of hops n increases the cumulative entropy $H(\hat{s_n})$ is often observed to increase, while the mutual information $I(S;\hat{S}_n)$ is expected to decrease under these assumpions.

$$\frac{dI(S;\hat{S}_n)}{dn} \leq 0 \quad \text{and} \quad \frac{dH(\hat{S}_n)}{dn} \geq 0$$

This decay in fidelity is critical in both theoretical and practical terms, especially when information is passed across multiple agents (i.e. human, machine, or a combination).

Given this framework, we propose the following hypothesis regarding multi-hop information transmission chains.

Hypothesis Statement

Null Hypothesis (H_0): The fidelity of transmitted information, as measured by semantic similarity or factual consistency, remains constant regardless of the number or type of transformation steps in the communication chain.

Alternative Hypothesis (H_1): The fidelity of transmitted information degrades with increasing transformation steps, and the rate of degradation varies depending on whether transformations are performed by humans, LLMs, or mixed-agent chains.

3 Proposed Metrics for Modeling Fidelity Degradation

To formalize the cognitive and information-theoretic foundations of information transmission, we propose three novel fidelity metrics: *Information Degradation Rate (IDR)*, *Meaning Preservation Entropy (MPE)*, and *Cascaded Hallucination Propagation Index (CHPI)*. These metrics capture distinct but complementary aspects of fidelity loss in iterative language transformations, providing a principled framework for analyzing degradation in human, machine, and hybrid communication chains.

For example, consider the fact: *"Lionel Messi scored the winning goal for Argentina in the 2022 World Cup final."* After one iteration, a paraphrase such as *"Messi secured Argentina's victory in the 2022 World Cup final with a goal"* maintains the original meaning with minimal drift, resulting in a low IDR and low MPE, indicating strong semantic alignment and factual preservation. In the next hop, *"Messi helped Argentina win the 2022 World Cup"* omits the specific detail of the goal, introducing semantic uncertainty that increases both IDR and MPE. By the third hop, the statement becomes *"Messi scored a hat-trick in the final,"* introducing a hallucinated event. If this fabrication remains unchanged in a subsequent hop, CHPI increases, reflecting the persistence and amplification of hallucinated content over time.

Unlike conventional metrics such as ROUGE (Lin, 2004) or BERTScore (Zhang et al.), which provide static, pairwise similarity estimates, the proposed metrics are designed to model the temporal dynamics of fidelity degradation across successive transformations. IDR captures semantic drift grounded in schema-driven memory decay. MPE draws from Shannon's information theory to quantify uncertainty in factual retention. CHPI identifies and quantifies the propagation of hallucinations across hops—an effect critical to understanding cascading failures in generation pipelines. These metrics are particularly relevant for applications such as multi-hop summarization, serial paraphrasing, and staged information relay, where fidelity loss is gradual, cumulative, and context-dependent.

Information Degradation Rate (IDR): Bartlett's serial reproduction theory posits that memory undergoes schema-driven compression, progressively omitting or distorting details over retellings. We operationalize this effect as the *average semantic drift per hop*. Given a chain of N texts $\{S_0, S_1, \ldots, S_{N-1}\}$ where S_0 is the original, we define:

$$\mathrm{IDR} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left[1 - \mathrm{BERTScore}\left(S_{i-1}, S_i\right)\right] \tag{6}$$

where $\mathrm{BERTScore}(\cdot, \cdot)$ measures semantic similarity between adjacent hops. A higher IDR indicates faster semantic drift and thus lower preserves fidelity.

Two-hop Chain and IDR Computation

Statements:

- S_0: "COVID-19 is caused by the SARS-CoV-2 virus."
- S_1: "The coronavirus SARS-CoV-2 is responsible for COVID-19." (BERTScore ≈ 0.95)
- S_2: "COVID-19 stems from various viruses including SARS-CoV-2." (BERTScore ≈ 0.78)

Computation:

$$\begin{aligned}\mathrm{IDR} &= \frac{1}{2}\left[(1-0.95)+(1-0.78)\right] \\ &= \frac{1}{2}(0.05+0.22) = 0.135\end{aligned}$$

Here, the first hop retains meaning closely (low drift), but the second hop introduces factual distortion, increasing the overall IDR.

Meaning Preservation Entropy (MPE): Shannon's information theory models communication as a *noisy channel*, where entropy quantifies uncertainty in transmitted information. In the context of multi-hop summarization or transformation, we view the preservation of meaning as a probabilistic process. Let $p_i^{(h)}$ denote the probability that atomic fact i is preserved at hop h, which is basically the BERTScore of the same model. We define:

$$\mathrm{MPE} = \frac{1}{N-1} \sum_{h=1}^{N-1} \Big[- \sum_{i=1}^{F} p_i^{(h)} \log p_i^{(h)} \Big] \tag{7}$$

where F is the number of atomic facts extracted from the original text. A higher MPE reflects increased uncertainty (entropy) in fact preservation across hops, consistent with the *noisy channel* view of information degradation.

Two hop Fact Preservation and MPE Computation

Original Fact Set (from S_0): $\{$(COVID-19, caused_by, SARS-CoV-2), (SARS-CoV-2, is_a, coronavirus)$\}$

After Two Hops:

- S_1: "The coronavirus SARS-CoV-2 is responsible for COVID-19." (Fact preservation probabilities: $p^{(1)} = [0.95,\ 0.92]$)
- S_2: "COVID-19 stems from various viruses including SARS-CoV-2." (Fact preservation probabilities: $p^{(2)} = [0.60,\ 0.30]$)

Computation:

$$\begin{aligned}\mathrm{MPE} = \frac{1}{2}\Big[&- \big(0.95 \log 0.95 + 0.92 \log 0.92\big) \\ &- \big(0.60 \log 0.60 + 0.30 \log 0.30\big)\Big] \\ &\approx 0.397.\end{aligned}$$

Here, hop 1 retains facts with low uncertainty (low entropy), while hop 2 shows high uncertainty, reflecting significant information loss.

Cascaded Hallucination Propagation Index (CHPI): Hallucinations refer to synthetic or spurious content introduced during iterative paraphrasing or summarization, which may persist and propagate across successive hops. To model this phenomenon, we redefine CHPI by leveraging entailment-based (Ge et al., 2023) scoring to quantify the persistence of hallucinations relative to the original source text S_0. Formally:

$$\mathrm{CHPI} = \frac{1}{N-1} \sum_{h=1}^{N-1} \left[1 - P(S_h \models S_0)\right], \quad (8)$$

where $P(S_h \models S_0)$ denotes the probability that the statement at hop h, S_h, is logically entailed by the original text S_0. The term $1 - P(S_h \models S_0)$ thus measures the degree of hallucination at each hop. CHPI aggregates this measure across the transformation chain, capturing how hallucinated content is preserved or amplified over time.

Entailment Drop across Hops and CHPI Computation

Statements:
Original Statement (S_0): "COVID-19 is caused by the SARS-CoV-2 virus."
Hopwise Paraphrases:

- Hop 1 (S_1): "The coronavirus SARS-CoV-2 is responsible for COVID-19." (Entailment: 0.95)
- Hop 2 (S_2): "COVID-19 is caused by influenza viruses." (Entailment: 0.10)

Computation:

$$\begin{aligned}\mathrm{CHPI} &= \frac{1}{2}\left[(1 - 0.95) + (1 - 0.10)\right] \\ &= \frac{1}{2}(0.05 + 0.90) = 0.475\end{aligned}$$

Here, hallucination introduced in hop 2 significantly increases CHPI, indicating amplified factual divergence.

Together, IDR, MPE, and CHPI provide a principled lens for quantifying how meaning degrades through chains of transformations. IDR operationalizes *schema-driven semantic drift*, MPE captures *information-theoretic uncertainty*, and CHPI models *error cascades* analogous to amplification in a noisy communication channel. This unified formulation enables a rigorous investigation of fidelity loss across human cognition, LLM-only workflows, and hybrid human–LLM pipelines, thereby bridging classical theories of memory and communication with contemporary multi-hop evaluation in NLP.

4 Proposed TransTrace Dataset

We introduce *TransTrace*, a large-scale dataset designed to support systematic investigation of semantic degradation and fidelity loss in iterative summarization processes. The dataset comprises **700 summarization chains**, each constructed from **more than 200 diverse source paragraphs** spanning a wide range of topics and writing styles. The dataset encompasses a diverse range of domains, including news articles, medical abstracts, legal case summaries, Wikipedia passages, and scientific reports, thereby enhancing the overall coverage and representational breadth. TransTrace enables fine-grained analysis of meaning drift, hallucination propagation, and transformation dynamics across multiple hops of summarization under different communication regimes.

The dataset is organized into four experimental settings as shown in Figure 1, each targeting a distinct mode of transformation. In the first setting, designed to capture fidelity degradation in human-to-human communication, each of the source paragraphs was assigned to a unique participant with the instruction to summarize it in 50–60 words. Their summaries were then passed sequentially to four additional participants, each unaware of the original source, who repeated the summarization task based only on the previous hop's output. This results in five-hop human-only summarization chains for source texts.

In the second setting, we replicate the same five-hop summarization chain structure using large language models (LLMs), where each LLM receives the previous hop's output and produces a 50–60 word summary. This allows us to evaluate fidelity degradation in LLM-only pipelines under controlled conditions. To explore hybrid communication dynamics, the remaining two settings alternate between human and LLM agents. In the *LLM-to-Human* setting, the initial summary is generated by an LLM and subsequently passed through one human participant for summarization. Conversely, in the *Human-to-LLM* setting, the process begins with a human-generated summary followed by only one LLM summarization. These configurations enable comparative analysis of error propagation patterns across mixed human–LLM interfaces.

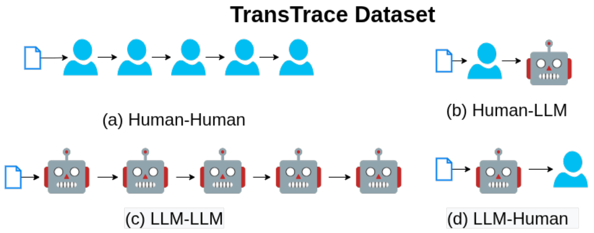

Figure 1: Overview of the four experimental settings in the *TransTrace* dataset. (a) *Human–Human*: A source paragraph is summarized iteratively by five different human participants, each with access only to the previous summary. (b) *Human–LLM*: The first summary is written by a human, followed by one successive summarizations by a large language model. (c) *LLM–LLM*: All five summarization hops are performed by an LLM, each conditioned only on the previous hop's output. (d) *LLM–Human*: The process begins with an LLM-generated summary and proceeds through one human summarizers in sequence.

TransTrace provides a unified testbed for studying meaning preservation, semantic drift, and hallucination behavior across iterative summarization. It bridges controlled experimentation with real-world communication dynamics, supporting evaluation of both cognitive theories and modern LLM behavior in multi-hop settings.

5 Results and Interpretation

To evaluate how meaning is preserved across different summarization chains, we compute six complementary metrics, summarized in Table 1. Three of these, IDR, MPE, and CHPI—are specifically designed to capture semantic shifts and distortions during iterative summarization. The remaining metrics, ROUGE, BERTScore, and Entailment are standard in the summarization and NLP literature.

To quantify semantic alignment, we use BERTScore and Entailment, both computed using the `deberta-xlarge-mnli` model (He et al., 2021). The entailment metric leverages natural language inference (NLI) to measure whether the final summary logically follows from the original source. Together, these metrics provide a comprehensive view of fidelity, hallucination, and pragmatic drift across conditions. We organize our findings across two paradigms: single-agent summarization chains, in which either a human or an LLM performs all summarization hops; and hybrid chains, in which a human and LLM alternate roles across hops.

5.1 Single-Agent Summarization

When five humans summarize in sequence without seeing the original text there's a sharp drop in meaning preservation. This is reflected in high CHPI scores, meaning hallucinations (added or imagined content) grow as the summary is passed along. This finding aligns with cognitive theories suggesting that human memory and language processing introduce bias and noise over time, especially when information is recalled and rewritten without full context.

LLM-only chains, like those generated by `LLaMA3-8B` (Grattafiori et al., 2024) and `Mistral-Saba` (Jiang et al., 2023), performed better in keeping the original meaning intact. These models had lower IDR and CHPI, which suggests that even without understanding in the human sense they can repeat and rephrase content more consistently than humans in chain-like tasks. This may be due to their ability to operate with perfect memory of the previous input, avoiding the cognitive "telephone effect" humans experience.

Among all models, the `Compound Beta` (Upadhye et al., 2019) system stood out. While its ROUGE score wasn't the highest, it had the lowest error in meaning (MPE), low hallucination propagation (CHPI), and the highest entailment scores. In other words, it may not match words as closely as others, but it captures the core meaning better. This supports the idea from cognitive pragmatics that shallow similarity (e.g., repeating the same words) doesn't always mean true understanding or faithful retelling. The `Qwen2.5` (Yang et al., 2025) and `GEMMA` (Team et al., 2024) amodels landed somewhere in the middle. `Qwen2.5` seemed more prone to letting hallucinations build up over time, while `GEMMA` preserved meaning a bit more reliably, but wasn't as consistent as the top performers. Tresults show that large language models especially those with hybrid designs—can actually outperform humans when it comes to preserving meaning over multiple summarization steps. This reflects a growing realization in cognitive science and AI research: humans are flexible but noisy communicators, while models, though not truly understanding, can offer high-fidelity transmission under certain conditions.

Table 1: Comparison of fidelity and standard evaluation metrics across Human and several LLM summarization chains. Lower IDR, MPE, and CHPI indicate better semantic preservation and reduced hallucination; higher ROUGE, BERTScore, and Entailment suggest stronger alignment with the source content.

Metric	Human	LLaMA3	Mistral-Saba	Compound Beta	Qwen2.5	GEMMA
IDR ↓	0.2431	**0.1434**	0.1130	0.2062	0.1842	0.1676
MPE ↓	0.0657	0.0572	0.0921	**0.0312**	0.0528	0.0391
CHPI ↓	0.5434	0.2660	0.2969	**0.1777**	0.3372	0.3002
ROUGE ↑	0.2484	0.2846	**0.4070**	0.1428	0.2068	0.2348
BERTScore ↑	0.6213	0.7014	**0.7950**	0.6260	0.6814	0.6967
Entailment ↑	0.4566	0.7351	0.7031	**0.8223**	0.6628	0.6998

Table 2: Metric comparison between hybrid summarization directions: Human→LLM vs LLM→Human.

Metric	Human→LLM	LLM→Human
IDR ↓	0.2296	0.2453
MPE ↓	0.0906	0.0771
CHPI ↓	0.1978	0.1807
ROUGE ↑	0.3876	0.3989
BERTScore ↑	0.7704	0.7547
Entailment ↑	0.8022	**0.8193**

5.2 Asymmetries in Human–LLM Hybrid Summarization Chains

We uncover a marked asymmetry in semantic fidelity across hybrid summarization chains involving alternating human and machine summarizers. In this experiment, we employ LLaMA3-8B as the representative LLM. We compare two configurations: **Human→LLM**, where the model revises a human-written summary, and **LLM→Human**, where a human revises a model-generated summary. While both chains yield similar surface-level ROUGE scores (0.388 vs. 0.399), more semantically grounded metrics reveal distinct behavioral profiles.

The Human→LLM chain achieves a higher BERTScore (0.770) and entailment score (0.802), indicating strong preservation of source meaning. Although this setting also shows a relatively elevated IDR (0.230) and the highest MPE (0.091), these distortions appear to be mitigated by the model's capacity for semantic repair. This pattern suggests that LLMs function effectively as semantic "correctors" when presented with compressed or pragmatically altered human input. From a cognitive standpoint, this aligns with the notion that LLMs, unconstrained by limitations in working memory or attentional bandwidth, can restore omitted content by leveraging distributional priors over language and meaning.

In contrast, the LLaMA3→Human configuration shows a higher IDR (0.245) and slightly lower BERTScore (0.755), despite achieving the highest entailment score (0.819) among all conditions. This suggests that human summarizers, when presented with model-generated input, tend to reinterpret underspecified or ambiguous language in ways that introduce plausible—but potentially ungrounded—semantic content. The lower CHPI score (0.181) indicates relatively fewer hallucinations, yet the increased IDR reflects lexical or paraphrastic drift. This mirrors phenomena observed in cognitive science, where humans—particularly under ambiguity or lossy input—rely on pragmatic inference and prior world knowledge, leading to expansions that may not strictly preserve the source intent.

These findings reveal cognitively motivated asymmetries in hybrid summarization pipelines. LLaMA3 demonstrates robustness to human-induced compression, whereas humans may over-interpret model outputs, leading to subtle semantic drift. Notably, both hybrid configurations surpass human-only and model-only baselines on entailment and hallucination metrics, suggesting that alternating human and LLM contributions can leverage complementary inductive biases.

6 Ablation Study

Since the proposed metrics (IDR, CHPI, and MPE) rely on the underlying embedding model used in BERTScore, we conducted an ablation study by substituting the original BERTScore encoder with RoBERTa-large and Sentence-BERT encoders. The Pearson correlation between BERTScore (BERT-base encoder) and RoBERTa-large was 0.90, indicating very high agreement, with a one-tailed paired t-test yielding a p-value nearly equal to 0. Similarly, the correlation between BERTScore and Sentence-BERT was 0.82, also indicating strong agreement, with a corresponding p-value = 1×10^{-9}. These results demonstrate that the proposed metrics are robust to the choice of encoder.

7 Limitation

Our study comes with several limitations. First, we used only one LLM architecture (LLaMA3-8B) across all model-based steps. This may limit generalizability, as different models (or scales) could behave differently. Second, the summarization tasks were performed in isolation and lacked real-world constraints like time pressure, user intent, or discourse context, all of which can influence how humans and models summarize. Third, while our metrics capture semantic drift and hallucination persistence, they rely on automated proxies and may miss subtler dimensions of meaning such as pragmatic nuance, discourse coherence, or cultural framing. Additionally, we treated humans and models as fixed agents without modeling variation within each group—different humans or model prompts may yield significantly different trajectories.

8 Conclusion & Future Work

In this work, we presented the first systematic study of meaning preservation in multi-hop summarization chains involving humans, LLMs, and their hybrids. Using over 700 five-step transformation chains and introducing three fidelity-focused metrics such as IDR, MPE, and CHPI we offer a deeper understanding of how information decays through iterative rewriting. Our findings reveal that hybrid human–LLM chains behave asymmetrically: while LLMs can effectively clean and sharpen human summaries, humans tend to reinterpret or drift from LLM outputs. LLM-only chains show the highest semantic drift, while human-only chains strike a more stable middle ground.

In future work, we plan to extend this framework to interactive summarization and co-editing scenarios, where humans and LLMs iteratively revise content together. We also hope to evaluate multilingual and multimodal settings, and incorporate human judgments to better align fidelity metrics with real-world expectations. Ultimately, understanding how meaning evolves across human–LLM chains can help design more trustworthy and controllable summarization workflows for high-stakes domains like healthcare, law, and journalism.

References

Xiuying Chen, Guodong Long, Chongyang Tao, Mingzhe Li, Xin Gao, Chengqi Zhang, and Xiangliang Zhang. 2023. Improving the robustness of summarization systems with dual augmentation. *arXiv preprint arXiv:2306.01090*.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Jiaxin Ge, Hongyin Luo, Yoon Kim, and James Glass. 2023. Entailment as robust self-learner. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13803–13817.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Yury Polyanskiy and Yihong Wu. 2017. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer.

Maxim Raginsky. 2016. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389.

Henry L Roediger III, Michelle L Meade, David A Gallo, and Kristina R Olson. 2014. Bartlett revisited: Direct comparison of repeated reproduction and serial reproduction techniques. *Journal of Applied Research in Memory and Cognition*, 3(4):266–271.

Tariq Shahzad, Tehseen Mazhar, Muhammad Usman Tariq, Wasim Ahmad, Khmaies Ouahada, and Habib Hamam. 2025. A comprehensive review of large language models: issues and solutions in learning environments. *Discover Sustainability*, 6(1):27.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Neelesh S Upadhye and 1 others. 2019. On the compound beta-binomial risk model with delayed claims and randomized dividends. *arXiv preprint arXiv:1908.03407*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.