LLM as a Broken Telephone: Iterative Generation Distorts Information

Amr Mohamed[1]†, Mingmeng Geng[2], Michalis Vazirgiannis[1,3], Guokan Shang[1]†

[1]MBZUAI, [2]SISSA, [3]Ecole Polytechnique

†Correspondence: {amr.mohamed, guokan.shang}@mbzuai.ac.ae

Abstract

As large language models are increasingly responsible for online content, concerns arise about the impact of repeatedly processing their own outputs. Inspired by the "broken telephone" effect in chained human communication, this study investigates whether LLMs similarly distort information through iterative generation. Through translation-based experiments, we find that distortion accumulates over time, influenced by language choice and chain complexity. While degradation is inevitable, it can be mitigated through strategic prompting techniques. These findings contribute to discussions on the long-term effects of AI-mediated information propagation, raising important questions about the reliability of LLM-generated content in iterative workflows.

0th	A lorry driver has been fined after his load of slabs fell off his vehicle on a bend, writing off a passing car worth £50,000.
2nd	A lorry driver was fined after a stone slab he was transporting fell off his vehicle at a bend, causing damage to a passing car worth up to £50,000.
10th	A bus received a $50,000 fine after a large rock dislodged from it at a bend and forced passing cars to swerve off the road.
50th	A bus received a compensation of $50,000 after a large rock struck the bus when the bus changed lanes on the city road, causing damage and an explosion on the road.
100th	A small car received a compensation of $50,000 after a large rock collided with the car, causing an accident and an explosion on the road.

Table 1: Example of iterative translations of an English news article using *Llama-3.1-8B-Instruct*, with Thai as the intermediate language, highlighting the distortions introduced over the different iterations.

1 Introduction

Large Language Models (LLMs) are becoming an integral part of our daily lives, helping us process, comprehend, and convey information via text, while also expanding their support to additional areas (Yin et al., 2023). Consequently, an increasing amount of online content is now model-generated or assisted (Geng and Trotta, 2024), and such content is almost indistinguishable from human-produced data (Uchendu et al., 2023).

This prompts us to consider the question: what effects arise when the same piece of information is repeatedly processed by LLMs through multiple iterations? This procedure is analogous to the telephone game in human communication, a widely known children's game in which a message is passed sequentially from one player to the next, with the final version often differing significantly from the original, usually with amusing or humorous effect. This happens because players often act as *broken telephones*, where information is gradually distorted as it is passed along the chain of individuals, highlighting how repeated transmission can lead to the accumulation of errors, omissions, or unintended alterations (Whallon et al., 2011).

Investigating these effects for LLMs is becoming increasingly crucial in the present era, because LLMs are not only consuming human-supplied information at one time, but also processing their own outputs in an iterative way. Therefore, our study focuses on exploring whether LLM also acts as a broken telephone, when the same content is continuously refined, paraphrased, or reprocessed, and particularly when the generated output becomes the input for subsequent model iterations. We expect to observe an effect similar to that of human information distortion through iterative generation.

In our study, we simulate the LLMs' telephone game primarily through the task of translation, as iterative translation serves as a critical and tangible testbed for examining how meaning and form degrade over repeated generations. This setup reflects real-world scenarios—for instance, cross-lingual news transmission—where content is repeatedly

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7493–7509
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

translated across languages. The extent of distortion, shaped by the interplay between a language's representation in the training data and its linguistic similarity to the source, is not unique to translation but illustrates a broader phenomenon also observed in other iterative LLM tasks, such as rephrasing, which we investigate under three experimental setups.

As illustrated in Figure 1, within each iteration, a document in English is subsequently translated into one or more different languages, then back to English, by leveraging LLMs. We compare the back-translated English version with the initial English version at every iteration with textual relevance and factuality measures, to investigate whether and how information distortion accumulates. Our results show that over time, small alterations in phrasing, meaning, or factual details can accumulate, leading to a progressive drift from the original source, as illustrated by the example in Table 1. Code and data are publicly available[1]. Our main findings include:

- The degree of information distortion in translation chains depends on the choice of intermediate languages, influenced by their linguistic similarity to the source language and their prevalence in the model's pre-training and post-training corpora.
- Greater chain complexity, whether by adding languages or models, often amplifies distortion, with longer chains introducing more degradation regardless of the type of iterative chain.
- Although distortion is unavoidable, it can be mitigated through temperature control and restricted prompting, which restrict the LLM from deviating significantly from the original text.

Our research echoes the ongoing conversation about the long-term impact of the widespread use of LLM-generated content on models themselves, humans, and society at large—often termed *model* and *knowledge* collapse (Guo et al., 2024b; Peterson, 2024). Our findings raise concerns about the reliability of AI-mediated information dissemination over the long term and in an iterative way.

2 Related Work

Model Collapse. Iterative training on synthetically generated data induces model collapse, a phenomenon characterized by systematic erosion of the long-tail components of the original data distribution (Shumailov et al., 2023). Theoretical analyses further elucidated how self-consuming training loops alter intrinsic scaling laws, thereby intensifying this collapse (Fu et al., 2024; Dohmatob et al., 2024), complementing earlier findings on distributional distortions (LeBrun et al., 2022). Furthermore, Guo et al. (2024b) demonstrated that iterative training on synthetic text does not preserve the nuanced richness of human language, particularly in creative tasks, underscoring the broader challenges of maintaining linguistic diversity in iteratively generated content.

Iterative Generation and Information Evolution. Iterative generation can trigger model collapse, whereby the diversity of real-world information degrades over time—a process that Peterson (2024) defines as knowledge collapse. Research on language evolution offers a framework for analyzing these degradations (Markov et al., 2023), aligning with broader perspectives on cultural evolution (Mesoudi and Whiten, 2008; Caldwell and Millen, 2008). In the context of LLMs, Perez et al. (2024) analyzed text properties evolution in rephrasing, continuation, and inspiration-taking tasks. Their work, however, overlooked translation—a key LLM application—and focused solely on chains involving a single model. Our work overcomes these shortcomings by investigating how iterative information translation accelerates distortions, explores heterogeneous model chains, and extends the analysis to higher complexity rephrasing chains, providing a broader view of iterative generation's impact on information evolution.

LLM Agents. We consider the implications for multi-agent settings, where communication frameworks leverage collaborative interactions between multiple LLMs (Park et al., 2023; Wu et al., 2023; Li et al., 2024b). These frameworks enable agents to iteratively refine outputs through debate-style interactions (Helm et al., 2024) or cooperative task decomposition (Pham et al., 2023), often improving accuracy in mathematical and logical tasks (Zhang et al., 2024). As introduced by Park et al. (2023), generative agents showcase the potential for creating interactive simulacra of human behavior through memory, reflection, and planning. However, such architectures implicitly assume that iterative exchanges preserve or enhance information fidelity—a premise challenged by our findings in translation chains. While prior work focuses on emergent problem-solving capabilities (Chan et al., 2024), our study reveals how these same iterative

[1] https://github.com/amr-mohamedd/LLM-as-a-Broken-Telephone

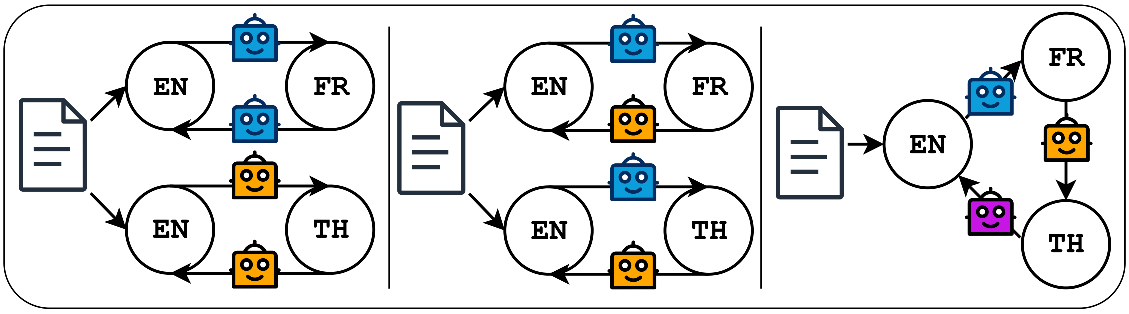

Figure 1: Overview of examples from our three experimental setups. ***Left***: Bilingual Self-Loop—A single model iteratively translates a document from English (EN) to French (FR) or Thai (TH) and back to English. ***Middle***: Bilingual Two-Player—Two different models collaborate within the same chain on translating between English and French or English and Thai. ***Right***: Multilingual Multiplayer—A more complex translation chain involving multiple languages and models, designed to examine how increasing the variety of languages and models accelerates information distortion over iterative generations.

mechanisms accelerate information distortion, particularly in scenarios where translation ambiguities compound through successive agent handoffs.

Evaluation of LLM Outputs. In addition to the multi-agent perspective, it is essential to scrutinize how LLM outputs are evaluated. Existing research predominantly relies on metrics such as token similarity (Hu and Zhou, 2024), output diversity (Guo et al., 2024a; Shaib et al., 2024), and factuality (Wang et al., 2023; Iqbal et al., 2024; Min et al., 2023; Chern et al., 2023). However, these evaluations are generally confined to single iterations and fail to capture the cumulative degradation introduced by iterative generation—a critical aspect of the translation chains under investigation. Although previous studies have explored variations in toxicity, positivity, difficulty, and length in iterative LLM transmission chains (Perez et al., 2024), they have overlooked the systematic assessment of textual similarity and factuality. Our work addresses this gap by providing a rigorous analysis of the deterioration of these properties over successive iterations in both translation and rephrasing tasks.

3 Methodology

In this section, we formalize the telephone game procedure with machine translation, noting that the broken telephone effect may occur with any generative task when carried out iteratively.

3.1 Notations and Definitions

Let $\mathcal{D} = \{d_i\}_{i=1}^{I}$ denote a set of I *documents*, $\mathcal{L} = \{l_j\}_{j=1}^{J}$ as a set of J natural *languages*, and $\mathcal{M} = \{m_k\}_{k=1}^{K}$ for a set of K *models*.

We define a *translation chain* as a sequence of N translation iterations that progressively transform a document. For iteration $t \geq 1$, let $d_{i,l_{\text{source}}}^{(t-1)}$ be the i-th document in the source language at iteration $t-1$. At iteration t, an ordered language chain is constructed by selecting a permutation $\pi^{(t)}$ of $J-1$ languages from $\mathcal{L}$ and forming the sequence

$$\mathcal{L}^{(t)} = (l_1^{(t)}, l_2^{(t)}, \ldots, l_{J-1}^{(t)}, l_J^{(t)}) \qquad (1)$$

with the requirement that $l_J^{(t)} = l_{\text{source}}$ (ensuring that the final translation returns to the source language). Simultaneously, a model sequence

$$\mathcal{M}^{(t)} = \left(m_1^{(t)}, m_2^{(t)}, \ldots, m_J^{(t)}\right) \qquad (2)$$

is defined, where each $m_k^{(t)}$ is sampled uniformly from $\mathcal{M}$ (allowing repeats; if $|\mathcal{M}| = K = 1$, the same model is used throughout).

Let $\mathcal{T}_{a \leftarrow b}^{m}(\cdot)$ denote the translation operator that converts an input from language b to language a using model m. The composed operator for iteration t is then

$$\mathcal{T}^{(t)} = \mathcal{T}^{m_J^{(t)}}_{l_J^{(t)} \leftarrow l_{J-1}^{(t)}} \circ \cdots \circ \mathcal{T}^{m_2^{(t)}}_{l_2^{(t)} \leftarrow l_1^{(t)}} \circ \mathcal{T}^{m_1^{(t)}}_{l_1^{(t)} \leftarrow l_{\text{source}}} \qquad (3)$$

so that the updated document is given by

$$d_{i,l_{\text{source}}}^{(t)} = \mathcal{T}^{(t)}\left(d_{i,l_{\text{source}}}^{(t-1)}\right). \qquad (4)$$

Starting with $d_{i,l_{\text{source}}}^{(0)} = d_i$, the process yields the sequence $(d_{i,l_{\text{source}}}^{(0)}, d_{i,l_{\text{source}}}^{(1)}, \ldots, d_{i,l_{\text{source}}}^{(N)})$, where N is the total number of iterations.

3.2 Experimental Settings

Languages. We selected *English* (*EN*) as l_{source} for all experiments and *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*) as the *bridge* (intermediate) languages in the translation chains. Within each iteration, a document in English is subsequently translated into one or more bridge languages, then back to English. This set creates varying degrees of semantic, lexical, and syntactic similarities between the source language and the bridge languages, which may differentially influence the extent of distortion introduced within the translation chains (Marchisio et al., 2020; Guerin et al., 2024).

Datasets. We utilized three datasets that span distinct domains: *BookSum* (Kryściński et al., 2021), *ScriptBase-alpha* (Gorinski and Lapata, 2015), and *(BBC)News2024* (Li et al., 2024a), from which we select articles published in 2024 to minimize the chances of data exposure that may result in biases amplification over the iterations (Luo et al., 2024; Li et al., 2024a). For our experiments, we randomly select 150 documents from each dataset, with each document containing between 100 and 200 words long.

Models. We primarily used two models, LLAMA-3.1-8B-INSTRUCT (*Llama*) (Dubey et al., 2024) and MISTRAL-7B-INSTRUCT-V0.2 (*Mistral*) (Jiang et al., 2023), for our main experiments. Additionaly, GEMMA-2-9B-IT (*Gemma*) (Team et al., 2024) is incorporated into Experiment 3 (Section 4.3) to evaluate higher complexity chains.

Decoding Parameters and Translation Prompt. Each model was used for inference with its default decoding parameters. We capped the maximum number of newly generated tokens at 8000 to encourage open-ended generation. This high limit allows translations, which can vary in length across different languages, to conclude naturally rather than being prematurely truncated. Models within the main experiments were prompted to translate documents from a source to a target language with a moderately constrained prompt. The full translation prompt can be found in Appendix E.

3.3 Evaluation Metrics

To comprehensively assess the impact of iterative generation on text quality, we employ two complementary sets of evaluation metrics: textual relevance and factuality preservation. The former quantifies the lexical, syntactic, and semantic deviations introduced at each generation step, while the latter evaluates the degree to which the generated text remains faithful to the original information.

Textual Relevance. We used **BLEU** (Papineni et al., 2002) to detect incremental errors, **ROUGE-1** (Lin, 2004) to quantify word-level omissions and subtle deviations, **CHR-F** (Popović, 2015) for capturing character-level deviations and errors accumulation, **METEOR** (Banerjee and Lavie, 2005) for being adept at capturing paraphrastic variations and subtle semantic shifts, and finally **BERTScore** (Zhang et al., 2019) for its focus on nuanced contextual and semantic relationships beyond traditional n-gram overlap-based methods.

Factuality Preservation. FActScore (Min et al., 2023) decomposes long-form text into atomic units and verifies each against a trusted reference using a dedicated judge model. In this study, we assume that the original text is factually correct and use FActScore to assess the rate of factuality degradation over the different iterations by comparing each model generation with its original text, then employ *Claude 3.5 Sonnet* to be the judge model.

4 Experiments

4.1 Experiment 1: Bilingual Self-loop

Setup. We fix the language set to

$$\mathcal{L} = \{\text{EN}, \, l_{\text{bridge}}\} \tag{5}$$

where $l_{bridge} \in \{\text{FR, DE, NL, VN, ZH, TH}\}$. We consider the case when $|\mathcal{M}_1| = |\mathcal{M}_2| = 1$, with $\mathcal{M}_1$ and $\mathcal{M}_2$ containing *Llama* and *Mistral* respectively. We also consider the three datasets: *BookSum*, *ScriptBase-alpha* , and *News2024*. For each dataset $\mathcal{D}$, every document $d_i^{(0)} \in \mathcal{D}$ undergoes $N = 100$ translation iterations with an iteration of the form:

$$\text{EN} \;\rightarrow\; l_{\text{bridge}} \;\rightarrow\; \text{EN}. \tag{6}$$

All translations within a single chain are performed by a single model. Concretely, at iteration t, the translation operator

$$\mathcal{T}^{(t)} = \mathcal{T}^{m_1}_{\text{EN}\leftarrow l_{\text{bridge}}} \circ \mathcal{T}^{m_1}_{l_{\text{bridge}}\leftarrow\text{EN}} \tag{7}$$

is applied to produce

$$d_i^{(t)} = \mathcal{T}^{(t)}\big(d_i^{(t-1)}\big). \tag{8}$$

This yields the sequence $(d_i^{(0)}, d_i^{(1)}, \ldots, d_i^{(100)})$ for each document $d_i^{(0)} \in \mathcal{D}$.

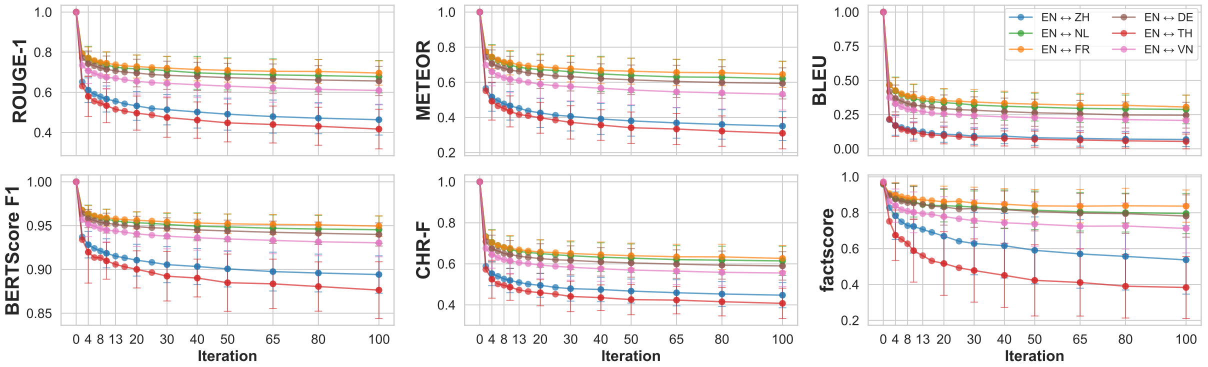

Figure 2: Results of *Llama* in the Bilingual Self-loop Experiment showing metrics evolution across translation iterations over the *News2024* dataset for *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*).

Hypothesis 1 (H1) *We hypothesize that iterative translation chains better preserve relevance and factuality when the bridge language shares lexical overlap, script, and syntax with the source language. In contrast, languages markedly dissimilar from the source language are expected to introduce greater distortion over iterations.*

Results. Figure 2 presents *Llama*'s iterative translation outcomes on the *News2024* dataset. Across all language pairs, there is a gradual decline in both factuality and relevance. Notably, language pairs exclusively using Latin script—with bridge languages such as French, German, and Dutch—demonstrated superior preservation of these qualities compared to those employing non-Latin script bridge languages, which exhibited more pronounced distortions over successive iterations. A similar trend was observed for *Llama* in the other datasets, while *Mistral* showed an even more severe decline across all three datasets. Comprehensive results for the remaining datasets and models are provided in Appendix B.1.

The average gradient values of FActScore in Table 2 quantify the rate of factuality loss across translation iterations. For language pairs composed solely of Latin script languages, gradients remain close to zero across all datasets and LLMs, indicating minimal degradations. For instance, in the *News2024* dataset, the average gradients for **EN** ↔ **FR** are -0.004 (±0.003) with *Llama* and -0.007 (±0.004) with *Mistral*, while for **EN** ↔ **DE** they are -0.005 (±0.003) and -0.011 (±0.006), respectively. In contrast, chains involving non-Latin scripts—particularly Thai—exhibit significantly faster factuality loss. In the *BookSum* dataset, the **EN** ↔ **TH** gradient is -0.026 (±0.014) with *Llama* and -0.040 (±0.025) with *Mistral*. This pattern is consistently observed across all evaluated language pairs, datasets, and models, with Thai demonstrating the highest rates of factual degradation.

4.2 Experiment 2: Bilingual Two-player

Setup. We fix the language set to

$$\mathcal{L} = \{\text{EN},\, l_{\text{bridge}}\} \tag{9}$$

where $l_{bridge} \in \{\text{FR}, \text{TH}\}$. Following the results presented in Section 4.1, we selected **EN** ↔ **FR** and **EN** ↔ **TH** for Experiment 2, as they demonstrated the lowest and highest levels of information distortion, respectively. We consider a model set $\mathcal{M}$ that includes both *Llama* and *Mistral*.

For this experiment, we used the *News2024* dataset because, as shown in Section 4.1, the choice of dataset did not significantly influence the observed trends, and to further mitigate data exposure (Luo et al., 2024; Li et al., 2024a).

Unlike Experiment 1, where a single model was used for both translation directions, we allow each translation step to potentially use a different model. At iteration t, we define a two-component model sequence:

$$\mathcal{M}^{(t)} = \left(m_1^{(t)}, m_2^{(t)}\right), \tag{10}$$

where $m_1^{(t)}$ is the model used for the translation from English to l_{bridge}, and $m_2^{(t)}$ is the model used for the translation from l_{bridge} to English. Each component is sampled uniformly from $\mathcal{M}$.

The translation operator at iteration t is then defined as:

$$\mathcal{T}^{(t)} = \mathcal{T}^{m_2^{(t)}}_{\text{EN}\leftarrow l_{\text{bridge}}} \circ \mathcal{T}^{m_1^{(t)}}_{l_{\text{bridge}}\leftarrow \text{EN}}. \tag{11}$$

		EN ↔ DE		EN ↔ FR		EN ↔ NL		EN ↔ TH		EN ↔ VN		EN ↔ ZH	
Dataset	**Model**	**Avg Grad.**	**Std Err.**	**Avg Grad.**	**Std Err.**	**Avg Grad.**	**Std Err.**	**Avg Grad.**	**Std Err.**	**Avg Grad.**	**Std Err.**	**Avg Grad.**	**Std Err.**
BookSum	Llama	-0.006	0.003	**-0.005**	0.003	-0.006	0.003	<u>-0.026</u>	0.014	-0.009	0.005	-0.021	0.011
	Mistral	-0.018	0.009	**-0.014**	0.007	-0.016	0.008	<u>-0.040</u>	0.025	-0.031	0.018	-0.028	0.015
News2024	Llama	-0.005	0.003	**-0.004**	0.003	-0.005	0.003	<u>-0.018</u>	0.009	-0.008	0.005	-0.011	0.006
	Mistral	-0.011	0.006	**-0.007**	0.004	-0.011	0.006	<u>-0.038</u>	0.022	-0.027	0.015	-0.024	0.012
ScriptBase-alpha	Llama	**-0.005**	0.003	-0.006	0.004	-0.005	0.004	<u>-0.015</u>	0.009	-0.011	0.007	-0.013	0.008
	Mistral	-0.010	0.006	**-0.008**	0.005	-0.009	0.006	<u>-0.039</u>	0.023	-0.027	0.015	-0.021	0.011

Table 2: Comparison of average gradient and standard error values of FActScore for the different models and language pairs across datasets.

The output document at iteration t is then determined as shown in Equation 8. This yields the sequence $(d_i^{(0)}, d_i^{(1)}, \ldots, d_i^{(100)})$ for each document in the *News2024* dataset.

Hypothesis 2 (H2) *We hypothesize that the coexistence of two different models in the same translation chain will add more distortions to the original information, thereby causing the original information to degrade over the successive iterations.*

Results. Figure 3 shows distinct patterns in the collaborative performance of *Llama* and *Mistral* across different languages. In French, the joint translation chain did not enhance the preservation of factuality or textual relevance relative to the models operating independently; instead, the collaboration introduced additional distortions that further degraded all evaluation metrics. Conversely, in Thai, the collaboration of *Llama* and *Mistral* resulted in reduced distortion compared to *Mistral* alone, though it still exhibited greater degradation than *Llama* in isolation.

We further quantified the average gradient of FActScore. For French, the collaborative chain exhibited an average gradient of -0.007 (±0.004), confirming minimal factuality degradation, though slightly worse than the standalone performances of *Llama* and *Mistral*. In contrast, for Thai, the collaborative chain showed a lower average gradient of -0.035 (±0.019) when compared to the standalone chain of *Mistral*. However, despite this lower decline, it was still outperformed by the standalone performance of *Llama*, where factual degradation was less pronounced.

4.3 Experiment 3: Multilingual Multiplayer

Setup. In this experiment, we design three settings of increasing complexity, each incorporating at least two bridge languages and at least two models within the same translation chain. The objective is to examine whether introducing a greater number of languages or models accelerates distortion.

Setting 1. We fix

$$\mathcal{L} = \{\text{EN}, \text{FR}, \text{TH}\} \tag{12}$$

and define $\mathcal{M}$ to contain both *Llama* and *Mistral*.

At each iteration t, we sample a permutation $\mathcal{L}^{(t)} = \pi^{(t)}(\mathcal{L})$ that enforces a cyclic translation path:

$$\text{EN} \to l_1^{(t)} \to l_2^{(t)} \to \text{EN},$$

with $l_1^{(t)}$ and $l_2^{(t)}$ drawn from $\{\text{FR}, \text{TH}\}$ and satisfying $l_1^{(t)} \neq l_2^{(t)}$. The corresponding model sequence is

$$\mathcal{M}^{(t)} = \big(m_1^{(t)}, m_2^{(t)}, m_3^{(t)}\big), \tag{13}$$

with each $m_k^{(t)}$ sampled uniformly from $\mathcal{M}$. The translation operator at iteration t is composed as:

$$\mathcal{T}^{(t)} = \mathcal{T}^{m_3^{(t)}}_{\text{EN}\leftarrow l_2^{(t)}} \circ \mathcal{T}^{m_2^{(t)}}_{l_2^{(t)}\leftarrow l_1^{(t)}} \circ \mathcal{T}^{m_1^{(t)}}_{l_1^{(t)}\leftarrow \text{EN}}, \tag{14}$$

which is applied iteratively to generate:

$$d_i^{(t)} = \mathcal{T}^{(t)}\Big(d_i^{(t-1)}\Big). \tag{15}$$

This produces $(d_i^{(0)}, d_i^{(1)}, \ldots, d_i^{(N)})$, where $d_i^{(0)}$ is the original document and $N = 100$.

Setting 2. We here retain $\mathcal{L}$ and the translation chain structure from Setting 1, utilizing the same translation operator as defined in Equation 14, while expanding $\mathcal{M}$ with an additional model, *Gemma*, to assess the impact of adding more models of similar size into the chain.

Setting 3. We extend the language set to:

$$\mathcal{L} = \{\text{EN}, \text{FR}, \text{TH}, \text{ZH}, \text{DE}\}, \tag{16}$$

and hold $\mathcal{M}$ fixed from Setting 1. The translation operator is then defined as:

$$\mathcal{T}^{(t)} = \mathcal{T}^{m_5^{(t)}}_{\text{EN}\leftarrow l_4^{(t)}} \circ \cdots \circ \mathcal{T}^{m_1^{(t)}}_{l_1^{(t)}\leftarrow \text{EN}}, \tag{17}$$

applied to generate $d_i^{(t)}$.

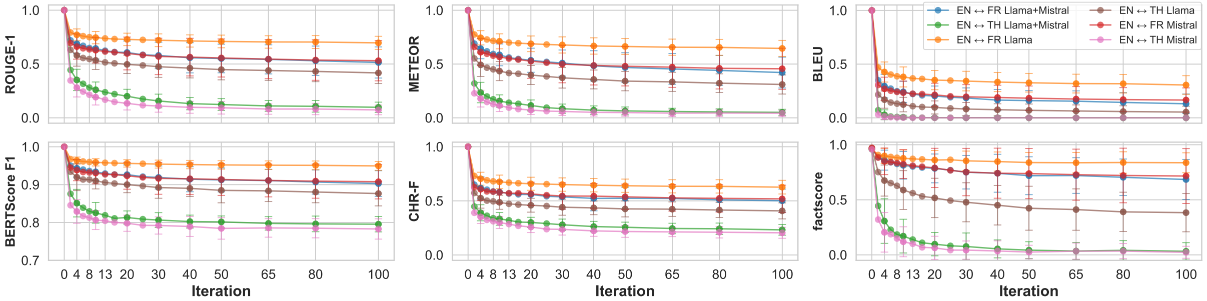

Figure 3: Comparison of metrics for the Bilingual Two-Player Experiment on the *News2024* dataset, illustrating the interaction effects between *Llama* and *Mistral* on translation chains for EN ↔ FR and EN ↔ TH, contrasted with their individual performances in the Bilingual Self-loop Experiment.

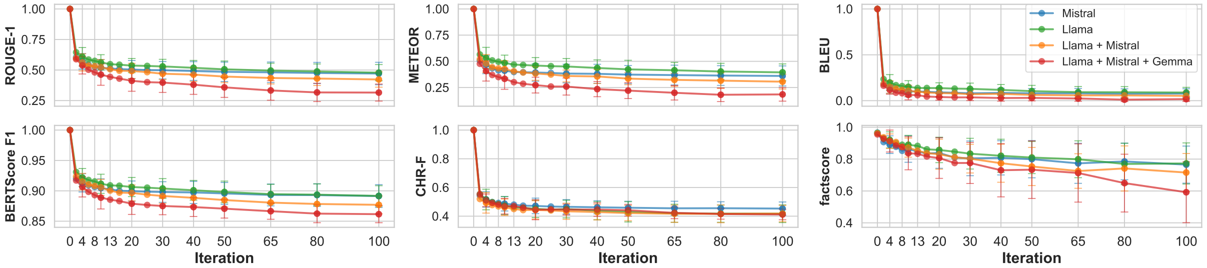

Figure 4: Results of rephrasing experiments on 30 randomly sampled documents from the *News2024* dataset. The figure compares the performance of individual models (*Llama*, *Mistral*, and *Gemma*) and their collaborative combinations over 100 rephrasing iterations.

Hypothesis 3 (H3) *We hypothesize that higher complexity translation chains cause higher factual degradation of the source document.*

Results. As shown in Appendix B.2, all three experimental settings indicated a comparable degree of factual, lexical, and semantic degradation by the 100^{th} iteration across all evaluation metrics. However, differences emerged in the rate at which this degradation occurred. Specifically, Setting 3 exhibited the steepest decline in factual accuracy, with an average FActScore gradient of -0.038 ± 0.02. By the 10^{th} generation, Setting 3's factuality had dropped to 0.054, and by the 100^{th} generation, it further declined to 0.04. Setting 1 followed closely, with an average gradient of -0.036 ± 0.02, showing a factuality score of 0.063 at the 10^{th} generation and 0.04 at the 100^{th}. Setting 2 exhibited the slowest rate of factual degradation, with an average gradient of -0.034 ± 0.02, reaching 0.075 at the 10^{th} generation and 0.04 at the 100^{th}.

5 Ablation Studies

5.1 Other Tasks: Rephrasing

Building on our findings in section 4, we extend our experiments to explore whether information distortion manifests in other types of iterative generation chains. Inspired by the work of Perez et al. (2024), who examined the evolution of toxicity, positivity, difficulty, and length in rephrasing as well as in continuation and inspiration-taking chains, we further probe the effects of information distortion in more complex rephrasing chains.

In this task, the model is instructed to rephrase a given document while preserving its full meaning (the full rephrasing prompt can be found in Appendix E). We randomly sampled 30 documents from *News2024* and conducted four experiments based on the setups from Sections 4.1, 4.2, and 4.3 (Setting 2). These experiments tested standalone rephrasing chains, the collaborative effects of *Llama* and *Mistral*, and an extended setup incorporating *Gemma* into the chain.

Rephrasing results are presented in Figure 4. Textual relevance metrics reveal rapid degradation of lexical and semantic properties over iterations. Among individual models, *Llama* shows the slowest divergence in textual relevance, with *Mistral* following. When these models collaborate, the degradation in textual relevance increases, and combining *Llama*, *Mistral*, and *Gemma* leads to the steepest decline, particularly after 100 iterations. The same order was observed when evaluating fac-

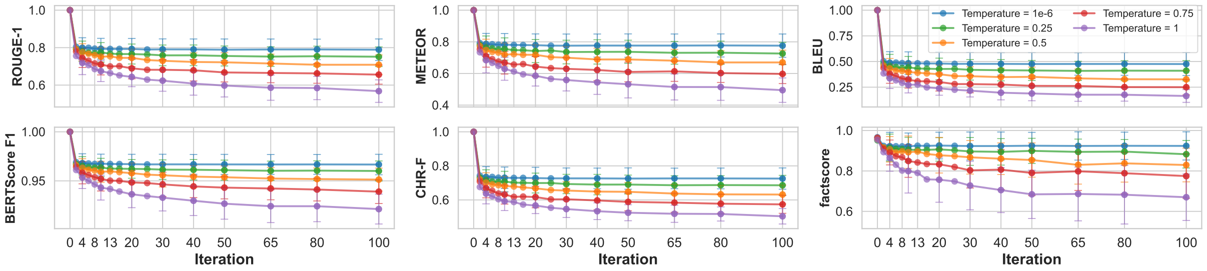

Figure 5: Impact of temperature variation on *Llama* outputs for 30 randomly sampled documents from the *News2024* dataset, evaluated on the EN ↔ FR translation chain.

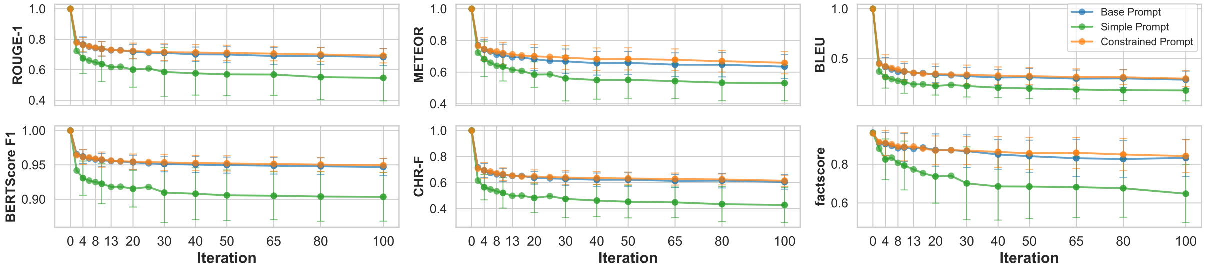

Figure 6: Impact of the prompt's level of constraint on noise accumulation measure on each of the metrics for *Llama* on 30 randomly sampled documents from the *News2024* dataset in iterative translation

tuality, although the loss was steadier, without clear convergence at the 100^{th} iteration.

5.2 Temperature Variation Affects Outputs

To further investigate the impact of decoding parameters on the models' outputs, we conducted several experiments using *Llama* across a spectrum of temperature parameter values, including 1×10^{-6}, 0.25, 0.5, 0.75, and 1.0 on 30 randomly sampled documents from *News2024*.

From Figure 5, higher temperature settings lead to greater factual and semantic degradation. At extremely low temperatures (1×10^{-6}), factuality drops slightly in the first two iterations but stabilizes thereafter. As temperature increases, stability diminishes, and factuality gradually diverges. Higher temperatures exacerbate this trend, with maximum temperature (1.0) causing the steepest decline, showing continuous divergence. Additional examples can be found in Appendix F.

5.3 Sensitivity of Iterative Translation Outputs to the Chosen Prompt

We subsequently investigated the influence of the translation prompt on the outputs produced by the iterative process. To this end, 30 documents were randomly sampled from the *News2024* dataset, and *Llama* was tasked with translating them using three distinct prompts characterized by varying levels of constraint: simple, base (used in all our experiments), and constrained. The complete prompts are provided in Appendix E.

Figure 6 illustrates that the level of constraint imposed by the prompt markedly affects the model's generation. Specifically, more constrained prompts were found to result in higher levels of relevance and factuality preservation.

6 Discussion and Conclusion

As LLMs increasingly shape online content, the likelihood that they re-process their own outputs continues to rise. This study confirms that such iterative generation leads to progressive information distortion, akin to the "broken telephone" effect in human communication. Our findings from translation-based experiments are multifaceted.

Effect of intermediate language(s) on information distortion. As found in Experiment 1, different language chains have varying levels of sensitivity to information distortion. As presented in Figure 2, we found that transmitting information between English and a highly similar language significantly reduces the distortion effect, while transmitting through a dissimilar language results in a more pronounced distortion. We suggest that this variation in information retention and distortion stems from the proportion of each language encountered during the models' training, with underrepresented languages experiencing greater distortion.

Chains of higher complexity may result in higher levels of distortion. Experiments 2 and 3 showed that increasing the levels of complexity of chains can result in higher levels of distortion. Figure 3 illustrates how the combination of *Llama* and *Mistral* amplified the distortion in the chain when French served as the bridge language. However, when Thai was used as the bridge language, their collaboration helped reduce distortion—likely due to the stronger model (*Llama*) and the weaker model (*Mistral*) interacting with an intermediate language that may have been underrepresented in *Mistral*'s training compared to *Llama*. Moreover, we observed that increasing the number of languages in the translation chain amplifies information distortion, likely due to the cumulative effects of longer generation sequences. In contrast, incorporating *Gemma* into the chain improved information retention, which we hypothesize stems from its larger parameter count—one to two billion more than *Llama* and *Mistral*. We leave the broader impact of model scaling for future work.

Information distortion can be reduced through temperature control and constrained prompting. Our findings suggest that while information distortion is unavoidable, it can be significantly mitigated through careful control of the model's generation temperature. Figure 5 shows that higher temperature values lead to greater distortion in the outputs, which we attribute to increased model creativity. A higher temperature encourages the generation of atypical tokens that may not fully preserve the meaning of the source document. Additionally, our analysis of prompt effects revealed that less constrained prompts contribute to greater noise accumulation over multiple iterations, resulting in higher divergence from the original meaning.

These findings underscore the need for strategies to mitigate such degradation and ensure the reliability of AI-generated content.

Limitations

While our study utilizes datasets from three distinct domains—book summaries, movie scripts, and news articles—these sources share similar characteristics and may not reflect the rare or long-tailed information found in specialized domains. Moreover, due to computational resource limitations, our experiments are restricted to models with 7–9 billion parameters, using their default generation settings. Future work should investigate whether incorporating datasets from specialized domains, employing larger models, or varying generation strategies (e.g., greedy decoding) impacts the degree of information distortion in iterative generation chains.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Christine A. Caldwell and Alan E. Millen. 2008. Studying cumulative cultural evolution in the laboratory. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1665–1670.

Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

I-Chun (Steffi) Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv:2307.13528.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. 2024. Towards theoretical understandings of self-consuming generative models. *arXiv preprint arXiv:2402.11778*.

Mingmeng Geng and Roberto Trotta. 2024. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

Nicolas Guerin, Shane Steinert-Threlkeld, and Emmanuel Chemla. 2024. The impact of syntactic and semantic proximity on machine translation with back-translation. *arXiv preprint arXiv:2403.18031*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024a. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024b. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Hayden Helm, Brandon Duderstadt, Youngser Park, and Carey E. Priebe. 2024. Tracking the perspectives of interacting language models. arXiv preprint arXiv:2406.11938.

Taojun Hu and Xiao-Hua Zhou. 2024. Unveiling llm evaluation focused on metrics: Challenges and solutions. *ArXiv*, abs/2404.09135.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of LLMs. arXiv preprint arXiv:2408.11832.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Benjamin LeBrun, Alessandro Sordoni, and Timothy J O'Donnell. 2022. Evaluating distributional distortion in neural language modeling. *arXiv preprint arXiv:2203.12788*.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024a. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18600–18607.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024b. Improving multi-agent debate with sparse communication topology. arXiv preprint arXiv:2406.11776; arXiv:2407.02030.

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.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Ilia Markov, Kseniia Kharitonova, and Elena L Grigorenko. 2023. Language: Its origin and ongoing evolution. *Journal of Intelligence*, 11(4):61.

Alex Mesoudi and Andrew Whiten. 2008. The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B*, 363(1509):3489–3501.

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. arXiv preprint arXiv:2305.14251.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Jérémy Perez, Corentin Léger, Grgur Kovač, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. 2024. When llms play the telephone game: Cumulative changes and attractors in iterated cultural transmissions. *arXiv preprint arXiv:2407.04503*.

Andrew J Peterson. 2024. Ai and the problem of knowledge collapse. *arXiv preprint arXiv:2404.03502*.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. 2023. Let models speak ciphers: Multiagent debate through embeddings. arXiv preprint arXiv:2310.06272.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. 2024. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2023. Factcheck-GPT: End-to-end fine-grained document-level fact-checking and correction of LLM output. arXiv preprint arXiv:2311.09000.

Robert Whallon, Robert K Hitchcock, and William A Lovis. 2011. *Information and Its Role in Hunter-Gatherer Bands*. Cotsen Institute of Archaeology Press at UCLA.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. arXiv preprint arXiv:2308.08155.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A Additional Details

All experiments were conducted using NVIDIA A100 (40GB VRAM) and A10 (24GB VRAM) GPU clusters. The compute allocation totaled 54 GPU-days, comprising 36 GPU-days on 8×A100 nodes and 18 GPU-days on 4×A10 nodes.

B Experiments Results Visualizations

We hereby present the complement of the visualizations of results from sections 4.1 and 4.3

B.1 Experiment 1: Bilingual Self-loop

B.1.1 Llama

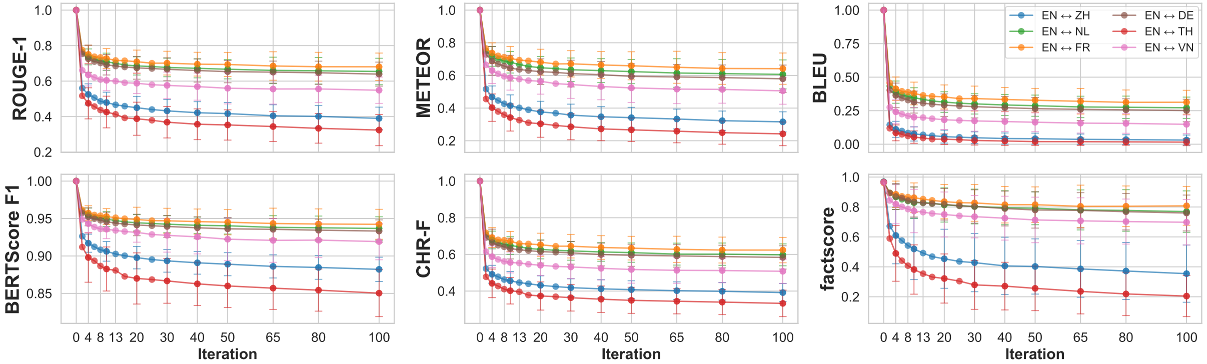

Figure 7: Results of *Llama* in the Bilingual Self-loop Experiment showing metrics evolution across translation iterations over the *BookSum* dataset for *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*)

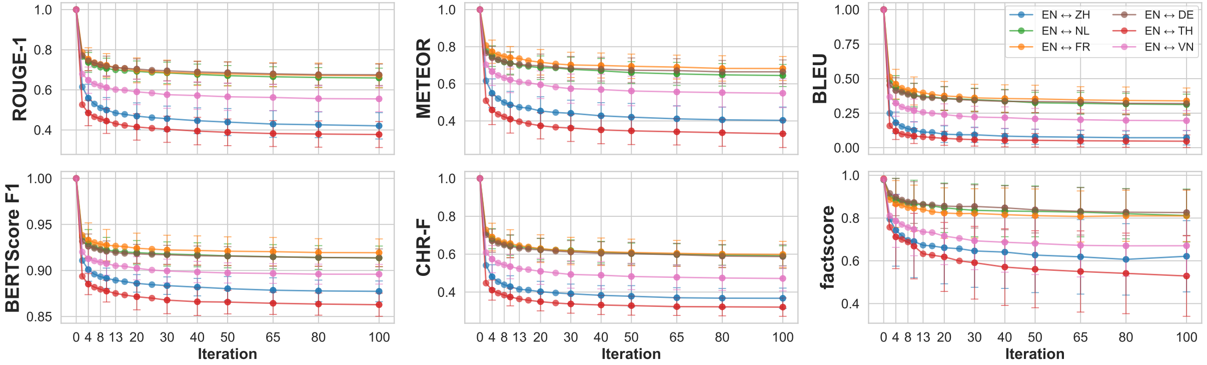

Figure 8: Results of *Llama* in the Bilingual Self-loop Experiment showing metrics evolution across translation iterations over the *ScriptBase-alpha* dataset for *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*)

B.1.2 Mistral

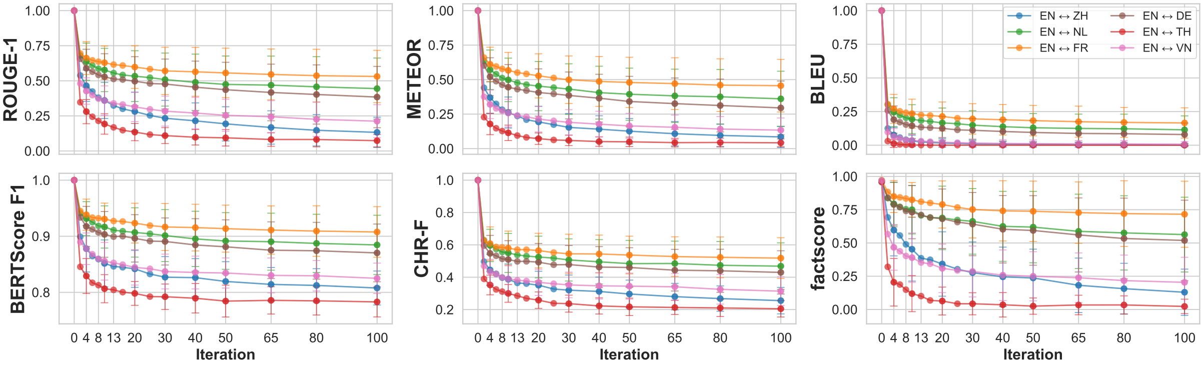

Figure 9: Results of *Mistral* in the Bilingual Self-loop Experiment showing metrics evolution across translation iterations over the *News2024* dataset for *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*)

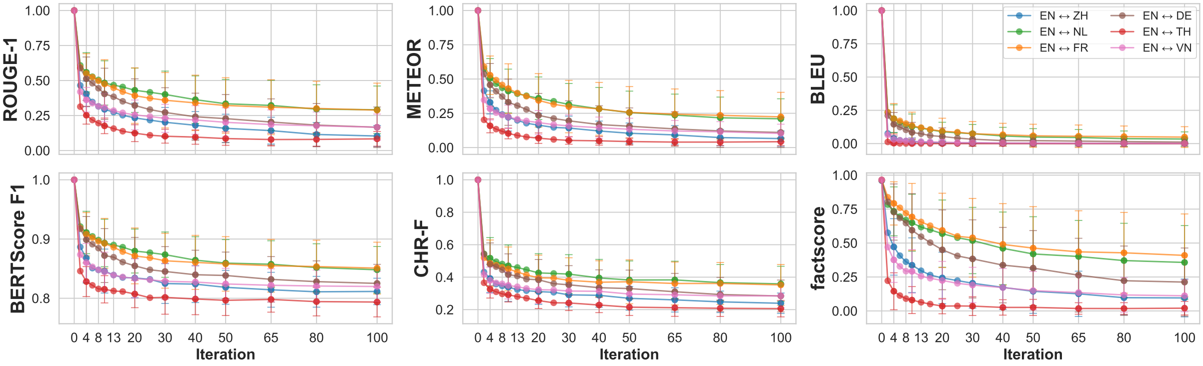

Figure 10: Results of *Mistral* in the Bilingual Self-loop Experiment showing metrics evolution across translation iterations over the *BookSum* dataset for *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*)

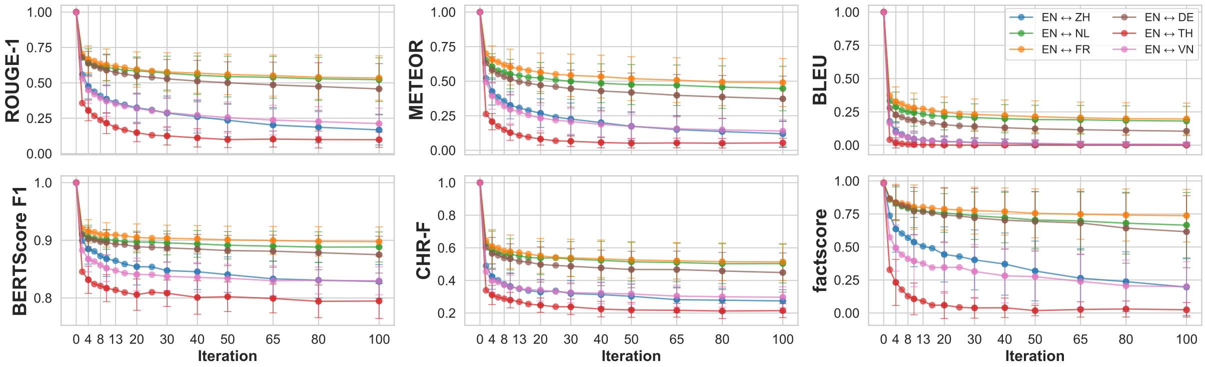

Figure 11: Results of *Mistral* in the Bilingual Self-loop Experiment showing metrics evolution across translation iterations over the *ScriptBase-alpha* dataset for *French* (*FR*), *German* (*DE*), *Dutch* (*NL*), *Vietnamese* (*VN*), *Chinese* (*ZH*), and *Thai* (*TH*)

B.2 Experiment 3: Multilingual Multiplayer

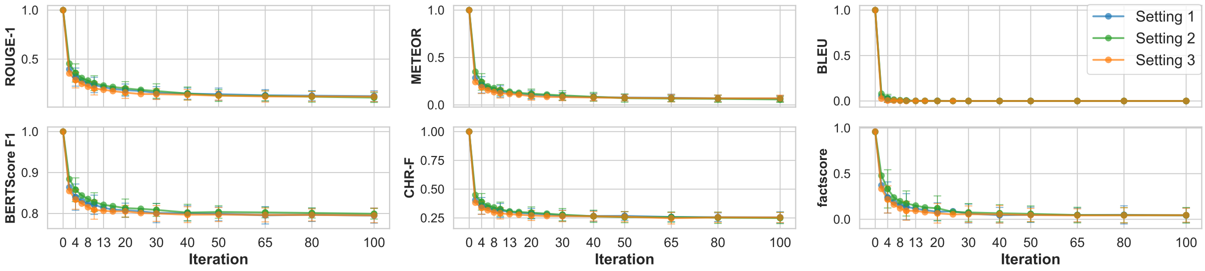

Figure 12: Results of Experiment 3 showing metrics evolution across translation iterations over the *News2024* dataset. ***Setting 1*** presents the translation chain of English with the bridge languages French and Thai. ***Setting 2*** maintains the same chain as *Setting 1*, while adding *Gemma* to the set of models. ***Setting 3*** further add the bridge languages Chinese and German to the chain from Setting 1.

C Generation Settings and Decoding Parameters

Each model was used for inference with its default decoding parameters as specified in their respective official documentation. We capped the maximum number of newly generated tokens at 8000 to encourage open-ended generation and ensure translations could conclude naturally. The choice of default parameters was made to reflect common practical usage scenarios where models are often deployed with these settings. The default sampling-based decoding parameters used for each model in our experiments are detailed in Table 3.

Model	Temperature	Top-p
Llama-3.1-8B-Instruct	0.6	0.9
Mistral-7B-Instruct-v0.2	0.0	N/A
Gemma-2-9B-it	1.0	0.95

Table 3: Default decoding parameters used for *LLama*, *Mistral*, and *Gemma*.

Further exploration of diverse decoding strategies, such as greedy decoding or beam search, remains an avenue for future work.

D Additional BLEURT Score Evaluations

To further reinforce our findings, we computed BLEURT scores (Sellam et al., 2020)—known for their high correlation with human judgments—across the full outputs of our primary experiments and rephrasing tasks. While main evaluations assessed FActScore for factual preservation alongside traditional textual relevance metrics (BLEU, ROUGE, METEOR, BERTScore) to capture changes in linguistic form over successive iterations, the BLEURT scores presented below (Tables 4 to 7) further substantiate our conclusions regarding information distortion. Higher BLEURT scores indicate better quality and closer semantic similarity to the original text.

D.1 Bilingual Self-loop BLEURT Scores

Table 4 presents the BLEURT scores for the Bilingual Self-loop experiment, corresponding to the *Llama* model on the *News2024* dataset (as detailed in Section 4.1 and visualized for other metrics in Figure 2).

Language Pair/	Iteration											
Iteration	0	3	6	10	15	20	30	40	50	65	80	100
EN ↔ FR	0.949	0.727	0.714	0.704	0.696	0.693	0.688	0.683	0.678	0.676	0.675	0.670
EN ↔ NL	0.949	0.721	0.704	0.689	0.679	0.670	0.664	0.657	0.654	0.649	0.645	0.642
EN ↔ DE	0.949	0.711	0.688	0.677	0.669	0.660	0.654	0.645	0.641	0.635	0.630	0.625
EN ↔ VN	0.949	0.681	0.659	0.643	0.632	0.625	0.614	0.606	0.603	0.596	0.589	0.586
EN ↔ ZH	0.949	0.618	0.587	0.570	0.551	0.537	0.522	0.514	0.508	0.498	0.494	0.489
EN ↔ TH	0.950	0.584	0.537	0.514	0.496	0.482	0.462	0.454	0.447	0.440	0.434	0.426

Table 4: BLEURT scores for the Bilingual Self-loop experiment using *Llama* on the *News2024* dataset. Scores show the evolution of text quality over 100 iterations for different language pairs.

Similar to our previous results for textual relevance and factuality in this experimental setup, the BLEURT scores exhibit the same trends. Specifically, there is a consistent decline in scores across all iterations, indicating progressive degradation of text quality. This degradation is less severe for language pairs where the bridge language uses a Latin script and shares more similarities with English (e.g., EN ↔ FR, EN ↔ NL, EN ↔ DE), which show higher BLEURT scores throughout the iterations compared to pairs involving non-Latin scripts or more distant languages (e.g., EN ↔ VN, EN ↔ ZH, and particularly EN ↔ TH). This observation aligns with Hypothesis 1, which posited that lexical and script similarity would influence the degree of information distortion.

D.2 Bilingual Two-Player BLEURT Scores

Table 5 shows BLEURT scores for the Bilingual Two-Player experiment on the *News2024* dataset (detailed in Section 4.2 and Figure 3), involving *Llama* and *Mistral*.

Language Pair/	Iteration											
Iteration	0	3	6	10	15	20	30	40	50	65	80	100
EN ↔ FR	0.949	0.684	0.661	0.640	0.622	0.614	0.597	0.592	0.575	0.570	0.564	0.547
EN ↔ TH	0.949	0.392	0.351	0.330	0.327	0.318	0.288	0.284	0.270	0.271	0.262	0.264

Table 5: BLEURT scores for the Bilingual Two-Player experiment on the *News2024* dataset.

D.3 Multilingual Multiplayer BLEURT Scores

Table 6 shows BLEURT scores for the Multilingual Multiplayer experiment on the *News2024* dataset (detailed in Section 4.3 and Appendix B.2, Figure 12).

Setting /	Iteration											
Iteration	0	3	6	10	15	20	30	40	50	65	80	100
Setting 1	0.949	0.385	0.348	0.334	0.333	0.329	0.311	0.295	0.305	0.304	0.302	0.295
Setting 2	0.949	0.414	0.360	0.330	0.328	0.312	0.312	0.292	0.284	0.275	0.274	0.283
Setting 3	0.949	0.397	0.350	0.326	0.312	0.307	0.309	0.303	0.298	0.296	0.287	0.300

Table 6: BLEURT scores for the Multilingual Multiplayer experiment on the *News2024* dataset.

D.4 Rephrasing Task BLEURT Scores

Table 7 shows BLEURT scores for the rephrasing task on the *News2024* dataset (detailed in Section 5.1 and Figure 4).

Model Combination /	Iteration											
Iteration	0	3	6	10	15	20	30	40	50	65	80	100
Llama	0.945	0.645	0.619	0.602	0.591	0.587	0.580	0.567	0.557	0.551	0.542	0.537
Mistral	0.945	0.593	0.577	0.562	0.552	0.545	0.538	0.535	0.533	0.533	0.528	0.521
Llama + Mistral	0.945	0.614	0.585	0.572	0.558	0.548	0.538	0.518	0.512	0.502	0.492	0.490
Llama + Mistral + Gemma	0.945	0.587	0.552	0.531	0.507	0.496	0.480	0.465	0.457	0.443	0.438	0.427

Table 7: BLEURT scores for the rephrasing task on the *News2024* dataset.

These BLEURT evaluations, due to their high alignment with human judgment, reinforce our original analysis and confirm that the pattern of iterative degradation observed in our experiments is robust. The decline in BLEURT scores over iterations across different tasks and settings provides further evidence for the "broken telephone" effect in LLM-based iterative generation.

E Generation Prompts

```
You are a translation expert. Given a passage, a source language, and a target language,
translate the passage from the source language to the target language while preserving
all the original meaning and without losing any context.
Do not write an introduction or a summary. Return only the translated passage.
Translate the following text from {source_language} to {target_language}: {document}
```

Figure 13: Base (main) translation Prompt for the translation chains experiments

```
Translate the following text from {source_language} to {target_language}: {document}
```

Figure 14: Simple translation Prompt for the prompt ablation

```
You are a translation expert. Please follow these instructions carefully:
- Task: You will receive a paragraph in {source_language}.
- Objective: Translate the paragraph into {target_language}.
- Guidelines:
  - Do not write an introduction or a summary.
  - Preserve the original meaning entirely; ensure no information is lost or altered.
  - Do not add, omit, or modify any details from the original paragraph.
  - Maintain the tone and style as closely as possible.
- Paragraph: {document}
```

Figure 15: Constrained translation Prompt for the prompt ablation

```
Given a passage, rephrase it while preserving all the original meaning and
without losing any context.
Do not write an introduction or a summary. Return only the rephrased passage.

Rephrase the following text: {document}
```

Figure 16: Rephrasing prompt used for the rephrasing task

F Examples Analysis

Temperature 1e-16: UEFA has imposed fines on the English Football Association and the Football Association of Ireland after their national anthems were booed before Ireland played England in the Nations League in September.
Temperature 0.25: The Union of European Football Associations (UEFA) has imposed fines on the England Football Association and the Football Association of Ireland after their national anthems were booed before Ireland played England in the Nations League in September.
Temperature 0.5: The UEFA Football Federation has sanctioned the England Football Association and the Football Association of Ireland after their national anthems were insulted before Ireland played England in the Nations League in September.
Temperature 0.75: The UEFA Football Federation has sanctioned the Football Association (FA) of England and the Football Association of Ireland (FAI) after their national anthems were deemed offensive before the UEFA Nations League match between England and Ireland in September.
Temperature 1: The Football Association (FA) of England and the Football Association of Ireland (FAI) have been sanctioned by the Union of European Football Associations (UEFA) due to incidents that occurred prior to the UEFA Nations League match between England and Ireland in September.

Table 8: An example of a news article highlighting the effect of temperature variation on the iterative translation process with English as the source language and French as the bridge language using *Llama* after 100 iterations. Color key: Entities, Financial Actions, Cultural Elements, Controversial Outcomes, Events/Locations, Emergent Details.