CSCI5541- S25- Final Report
A Study on Information Transmission and Transformation in LLM-Based Multi-Agent Systems for Effective Collaboration

Beñat Froemming-Aldanondo, Joshua Dickinson, Daniel Bielejeski, Isaac Ash-Johnson
InterAgent Communication Lab, University of Minnesota

Abstract

While significant research has focused on the behavior of individual large language models (LLMs), less attention has been paid to how information is transmitted and transformed between them. In human communication, it is well known that stories evolve as they are passed down through speech and writing. Inspired by this phenomenon, we investigate whether LLMs preserve message integrity in a simulated "telephone game" setting. Our study adopts a multi-agent framework motivated by swarm robotics, where distributed agents must share tasks without losing essential information. Experiments using GPT-4.1 reveal that message degradation occurs in distinct ways depending on prompt design.

1 Introduction

Throughout history, human communication has demonstrated that information rarely remains unchanged as it passes from person to person. Stories, legends, and historical events often evolve over time, shaped by the biases, interpretations, and conflicting memories of those who transmit them. A well known example is the legend of King Arthur and the Knights of the Round Table, where this tale is told over and over again, adapted by many authors. This phenomenon, where information subtly shifts with each retelling, highlights how cultural transmission and biases can alter narratives, sometimes distorting facts or introducing entirely new elements. Understanding these processes is crucial not only for studying human history but also for examining how artificial intelligence models, like large language models (LLMs), handle information transmission. Just like human stories change over time, AI systems may also transform data during repeated communication. This study draws inspiration from human cultural evolution to investigate whether LLMs can preserve message integrity during iterated information transmission. In this study, we conduct a series of experiments using LLMs to simulate the human "telephone game" to explore how information is transmitted and transformed across models. In the traditional telephone game, participants form a linear chain, passing a message from one person to the next with the goal of preserving the original message's integrity. The whispering element ensures that participants can't hear the message in advance, but this constraint doesn't apply to LLMs, as we can directly control input-output sequences along the chain. Unlike the traditional setup, we introduce multiple successors in our study to better simulate how a message propagates within a population. Each agent has a single parent and one or more successors.

We specifically focus on the application of swarm robotics. For instance, imagine a swarm of robots collaborating on a common task, and only one robot receives the initial task message. We assume that each robot is powered by its own LLM model, and the message must propagate through the swarm. If the message is altered along the way, it could disrupt the collaboration, as different robots would have different objectives in mind. For this reason, the messages transmitted in our experiments are detailed instructions.

For evaluation, we use both content-based and semantic-based metrics. In the first set of experiments, we compare the original message (root) with the message received by every other agent of the communication chain. In the second set, we ask the models to perform the received tasks themselves and measure success rate.

2 Literature Review

Extensive research has explored the behavior of individual LLMs, but studies on their interaction are relatively limited. However, there is significant research on human communication, particularly in the context of information transmission and the effects of repeated retelling, such as in the "telephone game". Bartlett laid the foundation for understanding memory and cultural transmission through factual story retelling in 1932 [Bartlett, 1932]. Mesoudi [Mesoudi et al., 2006] explored how stories with emotional content are more successfully transmitted than purely factual details. Additionally, Reisenzein [Reisenzein, 2000] investigated how surprise, as an affective component, aids in the retention and transmission of stories. These studies highlight the importance of emotional and surprising content in stabilizing story transmission over time. However, LLMs don't have emotions in the literal way that humans do.

There is limited research on the interaction between large language models (LLMs) without human intervention. One study explores the collaborative potential of multiple LLMs interacting to perform tasks, specifically in navigating from one Wikipedia page to another using the links provided on each page [Cheng et al., 2024]. Five methods of LLM interaction were tested and compared with a traditional graph-search algorithm and a single LLM as baselines. The findings show that effective collaboration is limited by the used prompts. Another research integrates Large Language Models (LLMs) with robot swarms to improve robot control and programming [Strobel et al., 2023]. The study proposes two approaches: indirect integration, where LLMs validate controllers before deployment, and direct integration, where each robot uses its own LLM for collaboration.

There are a few prior studies that examines how information changes in iterated interactions between LLMs using a "telephone game" approach. The first [Perez et al., 2024] study tracks how text properties like toxicity, positivity, difficulty, and length evolve through iterated interactions between LLMs. It reveals that biases and attractor effects are influenced by factors like the initial text, instructions, language model, and model size. Open-ended instructions lead to stronger distortions, and text properties like toxicity are more susceptible to these effects than length. The findings emphasize the need to consider multi-step transmission dynamics for better understanding LLM cultural behaviors. The second study includes a multilingual version [Mohamed et al., 2025] and demonstrates that distortion increases with time and complexity but can be mitigated with strategic prompting. The results also raise concerns about the reliability of LLM-generated content in iterative workflows.

3 Method

3.1 Novelty

Firstly, our research builds on existing studies of the traditional linear chain "telephone game" by incorporating multiple successors and one-to-many relationships, aiming to better simulate message propagation across a population or swarm of agents. Secondly, prior studies only focus on rewording or paraphrasing the message in each step of the propagation. Our research includes 10 different propagation strategies, including direct, explanatory, and adversarial prompts. This simulates an agent that needs to recall and explain the task that it was given correctly, and not just rephrase a text.

3.2 Materials and Data

We ran the experiments in Google Colab PRO using an A100 GPU. The model we used was the GPT-4.1 [OpenAI, 2023] through the API provided by OpenAI. Two smaller open source models were also tested using Huggingface, including Llama 3.2-3B and Mistral 7B Instruct. In the experiments, we extracted random tasks from two datasets: HowTo100M [Miech et al., 2019], a collection of YouTube tutorial transcripts; and RecipeBox [Eight Portions, 2020], a set of recipe instructions.

3.3 Pipeline and First Experiment

The first step of the pipeline involves randomly selecting individual tasks from the dataset to use as input in each of the experiments. A sample task we used is the following:

> "Fill your washing machine with hot water. You can do this by beginning a wash cycle, allowing the water level to rise. Pause your machine when it is sufficiently full, or when it enters its wash phase. Pour in 1 cup of bleach. Use the bleaching dispenser on your machine if it has one, otherwise add the bleach as you would in any other container. Mix it with the water by resuming the wash cycle. Let the water churn for 5 minutes before pausing again. Add your jacket to the machine. Place your jacket so that all parts are below the water line and leave it to soak for up to an hour. Finally, let the wash cycle run to completion, and run a second wash to ensure all traces of bleach are gone. Once dry, see if the new tone suits you by trying your jacket on. You may wish to run an extra cycle after extracting your jacket."

Then, we initialize a random communication tree, starting with a single root node. We set some parameters including `depth=8`, `num_agents=80`, and `max_children=3`. This allows us to build a tree

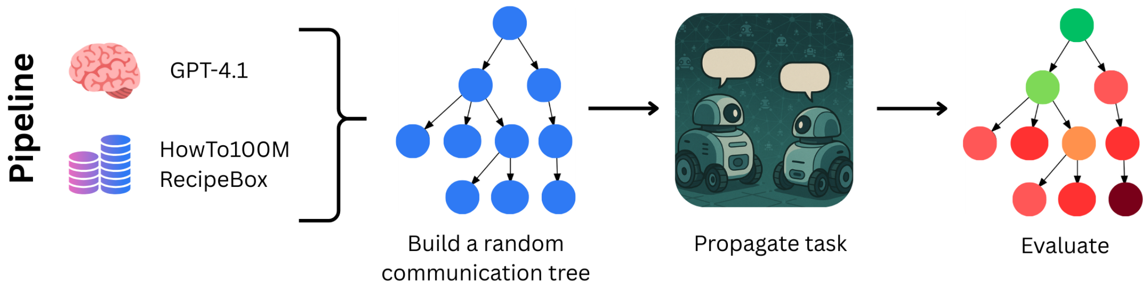

Figure 1: Pipeline.

with 8 levels and 80 nodes, each with a single parent (except the root) and 0-3 children. Each node in the tree corresponds to an agent and stores the message it received from its parent. Starting from the root, in a breadth-first manner, the task is transmitted from each node to its children using different prompting strategies: copy and paste, reword, explain, summarize, change tone to casual, optimize, expand, distort, explain to a child, and break into steps. Iteratively, the output of one node becomes the task of the next node. Once all agents received the task, we evaluate each node by comparing the original task with the task received by each agent using 3 metrics: length ratio and ROUGE score [Lin, 2004], to measure how much the text itself changed; and the cosine similarity of the average embeddings [Mikolov et al., 2013], to measure semantic drift.

We conducted experiments on five distinct randomly generated tasks and tree structures. For each depth level in the transmission chain, we averaged the results and visualized them in Figures 2–4. Figure 2 illustrates how the average message length changes with depth. As expected, copy-and-paste prompts preserved the original length. Prompts that involved rewording, breaking the message into steps, or simplifying for a child slightly altered the length. Expanding the task led to significant length increases, while distortion and explanation caused moderate growth. In contrast, prompts like casual tone conversion, summarization, and optimization consistently shortened the message. Figure 3 shows that shared word sequences drop sharply after the first transmission and then stabilize in subsequent steps. Distortion and expansion prompts altered the wording most significantly. Figure 4 tracks semantic drift across iterations. Rewording retained meaning most closely, which aligns with findings from prior work. However, we observed that prom-

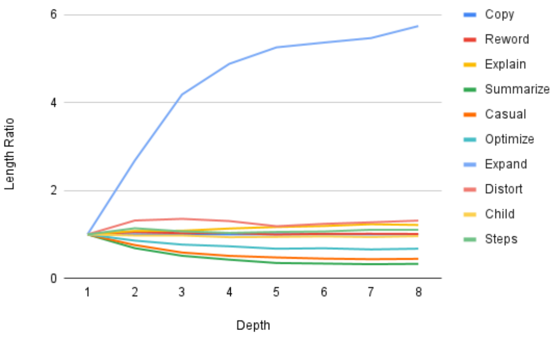

Figure 2: Average Length Ratio by depth.

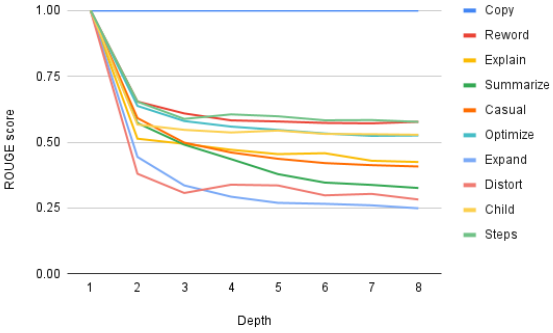

Figure 3: Average ROUGE score by depth.

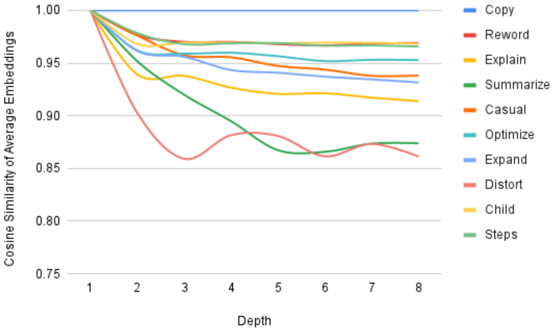

Figure 4: Average embeddings similarity by depth.

pts like summarization and explanation, despite seeming simple, resulted in considerable semantic loss, suggesting potential loss of key information. We repeated these experiments using LLaMA 3.2-3B and Mistral 7B Instruct. Both models showed rapid degradation, primarily due to formatting and prompt interpretation issues. Since each model's output became the input task for the next, consistent formatting was critical. Attempts to guide formatting through prompts often failed, leading to confusion and hallucinations. In contrast, GPT-4.1 performed reliably: adding a simple directive like "only output the new version in paragraph format" sufficed to maintain consistency across agents.

3.4 Second Experiment

In the second set of experiments, we again examine message propagation, but with a key difference: the tasks are simple problems that the LLMs can reliably solve individually with 100% accuracy. We weren't able to find any tasks that a large number of LLMs could do together collaboratively. An example task is:

> "Sarah has 3 boxes of markers. Each box contains 8 markers. She gives 5 markers to her friend. How many markers does she have left?"

During the propagation, we ask the LLMs not to give hints or solve the problem, but just reword the text or explain the problem statement. Each agent in the chain is then instructed to solve the problem it receives. If the message remains intact throughout propagation, the final answer should match the ground truth for all of them. This setup not only allows us to track semantic drift but also evaluates whether the transformed message remains interpretable and functional. Table 1 presents the average outcomes from 3 different problems similar to the one above.

Outcome	Reword	Explain
Correct	80	38
Incorrect	0	3
Not enough info	0	39

The results show that rewording preserved accuracy perfectly (80/80 correct), while explaining led to a performance drop, with an average of only 38 correct, 3 incorrect, and 39 responses lacking enough information. This suggests that explanatory prompts introduce more semantic drift than simple paraphrasing and may omit key details. A major limitation of this experiment is that even if a reformulated problem closely matches the original, the model might still fail to solve it especially if the task is complex, so outputs must be evaluated in context with the transformed problem itself. We observed that explanatory prompts often generalized the scenario, frequently omitting critical information such as numerical values. Overall, we found that prompt design and formatting had a significant impact on the reliability of message transmission.

4 Conclusion and Future Work

We investigated how messages change as they propagate through a network of large language models, using a branching communication structure and diverse prompting strategies. Our results show that even simple prompt variations can lead to significant semantic drift, affecting both message clarity and task success. Although some strategies preserved meaning better than others, no method was immune to degradation over multiple generations besides direct copy-paste. Additionally, smaller models like LLaMA 3.2–3B and Mistral 7B struggled with formatting and maintaining message integrity, often leading to confusion or hallucination. These findings highlight the difficulty of ensuring reliable communication in multi-agent LLM systems. Our experiments are fully reproducible using Google Colab and openly available models in Huggingface or by purchasing more advanced models like the OpenAI API. All code can be accessed on GitHub. The datasets we used are publicly available.

While our work does not pose immediate societal risks, in the future, as swarm robotics becomes more advanced, unreliable communication among LLMs could lead to failures in critical applications, such as automated decision-making or robotic control. Our study is limited to textual message propagation and individual execution, which simplifies the complexities of real-world agent interaction. Future directions include exploring dynamic feedback, bidirectional communication, and deploying these methods in physical multi-agent environments, such as robot swarms engaged in collaborative tasks. LLMs have already shown promise in enhancing human-robot interaction by enabling robots to interpret and act on natural language commands. This research represents a step toward realizing robotic systems that can communicate and collaborate with each other with the fluidity, adaptability, and reliability of human teams.

References

Federic C Bartlett. 1932. Remembering: An experimental and social study/frederic c. bartlett.

Grant Cheng, Oliver Wang, Alyssa Adams, Martin Biehl, Luc Caspar, and Olaf Witkowski. 2024. Interacting llms: A dive into collaborative ai. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 152–155.

Eight Portions. 2020. Recipebox dataset.

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.

Alex Mesoudi, Andrew Whiten, and Robin I. M. Dunbar. 2006. A bias for social information in human cultural transmission. *British Journal of Psychology*, 97(3):405–423.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Amr Mohamed, Mingmeng Geng, Michalis Vazirgiannis, and Guokan Shang. 2025. Llm as a broken telephone: Iterative generation distorts information.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jérémy Perez, Grgur Kovač, Corentin Léger, 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.

Rainer Reisenzein. 2000. Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition and Emotion*, 14(1):1–38.

Volker Strobel, Marco Dorigo, and Mario Fritz. 2023. Llm2swarm: Robot swarms that responsively reason, plan, and collaborate through llms. *arXiv:2410.11387v2*.