Use Case Demonstration @ DocEng2025: Conversation-Driven Multi-LLM Framework for Web Document Sentiment Analysis

Dominik Opitz
dominik.opitz@dlr.de
Institute of Software Technology, German Aerospace Center (DLR)
Germany

Andreas Hamm
andreas.hamm@dlr.de
Institute of Software Technology, German Aerospace Center (DLR)
Germany

Abstract

In this use case demonstration we show how a system of collaborative Large Language Models (LLMs) can be applied to the task of analyzing the sentiment of online news articles. The emergence of LLMs has proven to be highly valuable in interpreting unstructured text, offering nuanced and context-aware insights. While they can not fully replace traditional machine learning approaches for sentiment analysis, our approach illustrates how collaborative LLM architectures can enrich the explainability and trustworthiness of the outcomes.

CCS Concepts

• **Computing methodologies** → **Natural language generation**; • **Information systems** → *Information systems applications*; Sentiment analysis.

Keywords

Collaborative Language Models, Agent Discussion

ACM Reference Format:
Dominik Opitz and Andreas Hamm. 2025. Use Case Demonstration @ DocEng2025: Conversation-Driven Multi-LLM Framework for Web Document Sentiment Analysis. In *ACM Symposium on Document Engineering 2025 (DocEng '25), September 2–5, 2025, Nottingham, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3704268.3748678

1 Introduction

The vast number of news published daily online presents a challenge when it comes to staying informed in critical domains. Consequently, this can lead to information overload and exposes readers to inherent bias in many articles, which can influence opinions and hinder a neutral perspective.

Meanwhile, the emergence of Large Language Models (LLMs) has significantly advanced the automatic processing ability of unstructured text, yielding substantial improvements in challenges such as Question Answering and Sentiment Analysis (e.g. [2, 7]). So far, approaches often rely on a single LLM to perform the majority of the task. In social sciences, it is well established that groups of individuals can make better decisions than individuals alone [5]; this principle can be transferred to the field of Computer Science by enabling collaborative LLM architectures instead of single LLMs. Shanahan et al. [3] show how the integration of different perspectives into LLMs can provide a deeper understanding of the LLMs behavior. The ability of role-playing was further successfully applied by Sun et al. [4] to the task of Question Answering. The concept of collaboration has since been extended into other domains [6] and improved to solve complex tasks like code generation [1].

In this work, we showcase how a group of multiple, small-sized (3B - 8B) LLMs can be applied on a sentiment analysis task for digital news articles. While sentiment analysis is most often framed as a classification problem, the demonstrated approach treats it in a **more open form to find out finer nuances in opinions**. Our setup allows a group of LLMs to start a human-like discussion on the given task - analyzing the stance of a provided web article with respect to a given topic. As a result, the LLM-system outputs a list of identified stances which not only informs readers about potential biases upfront, but also enables clustering of news content based on shared perspectives, facilitating more structured topic analysis.

2 Architecture Design

Our approach integrates multiple LLMs (agents) into a single discussion (chat). They are given a task to discuss and solve collaboratively through the exchange of chat messages. Their goal is to return a final answer in a specific output format (e.g. a JSON). The task, the output format and the number of LLM agents along with their specific parameters are adjustable. We differentiate between *participants* - agents that actively generate messages in the conversation - and a moderator, who is prompted to quietly monitor the conversation and decide when to end it (e.g. when the goal is reached). To facilitate an effective conversation, after each message we let an independent LLM assign the next speaker based on the latest response, which typically contains clues as to who should speak next. We opt for small-sized LLMs for two specific reasons: First, it allows us to show that in collaboration, small-sized LLMs can reach results of a quality comparable to individual, large-sized LLMs. Second, since large models are often inaccessible due to hardware limitations, the ability of smaller models to collectively match their performance can offer a more practical and widely usable alternative.

3 Use Case Demonstration

We demonstrate our system on a specific use case: given an article[1] about Generative AI in Aerospace, we prompt three LLMs (`Agent`

[1]Article Source: https://a2globalelectronics.com/defense-aerospace/generative-ai-taking-off-in-the-aerospace-industry/, access: June 7th 2025

DocEng '25, Nottingham, United Kingdom

ACM ISBN 979-8-4007-1351-4/2025/09
https://doi.org/10.1145/3704268.3748678

A, Agent B, Agent C, each a Llama3 LLM with 8 billion parameters) to collaboratively discuss the article's viewpoint (stance) towards the focus topic *Space Exploration*. They are assigned to take an *optimistic*, a *pessimistic* and a *follow-along* personality trait, respectively. Our aim is to identify nuanced positions and potential biases that the article expresses towards Space Exploration. The conversation is initiated with a system prompt that instructs the agents on how to collaborate effectively, followed by a task description guiding them to identify and explain the key attitudes expressed toward the focus topic by examining both the content and tone of the article.

Conversation Demo. The conversation is opened by Agent A with a brief task summary, not suggesting any specific sentiments just yet and instead introducing potential areas to analyze (e.g. "... break down the article into sections and examine each one separately"). Most interestingly, Agent A ends their initial message by directly asking Agent B and Agent C about their thoughts. Agent B responds by approving Agent A's plan and expresses that the article is "overly promotional, lacks objectivity" and "too enthusiastic [ignoring] the potential risks or challenges". While these points of criticism are justified, it is worth noting that this initial, rather negative assessment comes from the only agent that was assigned a *pessimistic* personality trait. Agent A, which is assigned an optimistic personality, then agrees to Agent B's observation but attempts to shift the focus back to a more neutral/positive aspect ("... I suggest we explore how the benefits [...] are presented ...").

In subsequent messages, the agents highlight several moments where the article's overly positive tone is especially evident. They argue that the article lacks balance, overlooking potential drawbacks or limitations of AI in aerospace; the agents focus centers on critiquing the article's tone. However, a shift occurs when Agent B reminds the group that the task is "not to nitpick," but rather to "critically evaluate the article's claims." This message indeed shifts the focus of the conversation from the article's language style to the articles use of credible source and evidence of claims made.

After eight rounds of messages the moderator concludes the conversation and reports the article to be "optimistic", "promotional" and "uncritical". The reasoning for this evaluation expresses that the article "highlights the benefit and potential of generative AI in space exploration" but "lacks addressing any limitations or challenges, and ignores or downplays potential risks [or] drawbacks".

Conversation Insights. The conversation provides several insights into the agents' capabilities and limitations. In further experiments involving longer conversations (8+ messages), we observed that semantic drift tends to become noticeable around that point, likely due to the use of smaller-sized LLMs, which causes the agents to gradually deviate from the original topic. With growing number of chat messages, the LLMs increasingly struggle to correctly adhere to the original task, deviating from the topic. An early hint of potential drift appears to be the increased repetition of language patterns, the more messages are exchanged. This is especially apparent for the smaller 3B parameter sized models, which then tend to repeat paragraphs in the same format of "By doing X, we can achieve Y" (e.g. "By refining our approach, we can create [a richer analysis]"). In our framework, such early indicators of semantic drift are detected automatically using simple heuristics and intervened with a predefined message, bringing the conversation back on track.

For Llama models with 3B parameters there also appears to be a slight lack of self-awareness. Even though all agents are explicitly told about their own name and role within the conversation, they occasionally refer to themselves in third person (e.g. Agent A: "I appreciate Agent A and B's contributions.") and confuse messages written by themselves vs. other agents. This behavior indicates some level of confusion about their own "self" although it does not seem to occur frequently with models sized 8B or larger.

In our experiments, the use of multiple agents proved to be able to explore the given task from a range of different topics, which can greatly enhance the explainability and trustworthiness of the final results. The use of collaborative architectures does not necessarily improve the validity or integrity of the results - but it provides more nuanced, carefully considered perspectives, that contribute to a more elaborated outcome. For comparison, for the **same task description**, GPT-4o evaluated the same article as having an "overwhelmingly supportive stance toward space exploration" with a "hopeful, solution-driven, and celebratory tone [...] signaling belief that the fusion of AI and space science is not only beneficial but inevitable". While this assessment is accurate and properly reasoned by the model, it mirrors the article's tone without showing critical distance, although refining the prompt could likely lead to more critically evaluated results. In contrast, our multi-agent system demonstrates a heightened sensitivity to the text's rhetorical framing, explicitly showing greater awareness of how such enthusiasm may mask limitations or introduce bias.

Future Research. Our demonstration focuses on using the Llama-LLM family. Yet, it is worth exploring a mixture of different model families to increase the diversity of discussions. When using small sized models, further research should focus on mitigating semantic drift and optimizing for use cases beyond subjective sentiment analysis, such as logical reasoning and mathematical problems.

References

[1] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. Dynamic llm-agent network: an llm-agent collaboration framework with agent team optimization. (2023). doi:10.48550/arXiv.2310.02170{\#}.

[2] Luiz Rodrigues, Cleon Xavier, Newarney Costa, Hyan Batista, Luiz Felipe Bagnhuk Silva, Weslei Chaleghi de Melo, Dragan Gasevic, and Rafael Ferreira Mello. 2025. Llms performance in answering educational questions in brazilian portuguese: a preliminary analysis on llms potential to support diverse educational needs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (LAK '25). Association for Computing Machinery, 865–871. ISBN: 9798400707018. doi:10.1145/3706468.3706515.

[3] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623, 7987, 493–498. doi:10.1038/s41586-023-06647-8.

[4] Hongda Sun, Yuxuan Liu, Chengwei Wu, Haiyu Yan, Cheng Tai, Xin Gao, Shuo Shang, and Rui Yan. 2024. Harnessing multi-role capabilities of large language models for open-domain question answering. In *Proceedings of the ACM on Web Conference 2024*. Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw, (Eds.) ACM, New York, NY, USA, 4372–4382. ISBN: 9798400701719. doi:10.1145/3589334.3645670.

[5] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

[6] Frank Xing. 2025. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Trans. Manage. Inf. Syst.*, 16, 1, Article 5, (Feb. 2025), 24 pages. doi:10.1145/3688399.

[7] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: a reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Kevin Duh, Helena Gomez, and Steven Bethard, (Eds.) Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 3881–3906. doi:10.18653/v1/2024.findings-naacl.246.