Simulating Complex Immediate Textual Variation with Large Language Models

Fernando Aguilar-Canto
Centro de Investigación en Computación
Instituto Politécnico Nacional
faguilarc2021@cic.ipn.mx

Alberto Espinosa-Juárez
Centro de Investigación en Computación
Instituto Politécnico Nacional
aespinosaj2021@cic.ipn.mx

Hiram Calvo
Centro de Investigación en Computación
Instituto Politécnico Nacional
hcalvo@cic.ipn.mx

Abstract

Immediate Textual Variation (ITV) is defined as the process of introducing changes during text transmission from one node to another. One-step variation can be useful for testing specific philological hypotheses. In this paper, we propose using Large Language Models (LLMs) as text-modifying agents. We analyze three scenarios: (1) simple variations (omissions), (2) paraphrasing, and (3) paraphrasing with bias injection (polarity). We generate simulated news items using a predefined scheme. We hypothesize that central tendency measures—such as the mean and median vectors in the feature space of sentence transformers—can effectively approximate the original text representation. Our findings indicate that the median vector is a more accurate estimator of the original vector than most alternatives. However, in cases involving substantial rephrasing, the agent that produces the least semantic drift provides the best estimation, aligning with the principles of Bédierian textual criticism.

1 Introduction

According to Textual Criticism and Communication Theory, texts tend to be modified when they are transmitted from one sender to a receiver (Blecua, 1983; Pajares and Hernández, 2010). While most variations—random changes, omissions, and additions—have been extensively studied in Computational Textual Criticism, other changes, such as rephrasing or bias injection, are more difficult to address and introduce greater alterations than unintentional modifications.

For example, oral traditions are generally more challenging to reconstruct than written ones, and to our knowledge, none of the existing computational methods ((Fitch, 1971), RHM (Roos et al., 2006), Hoenen's algorithms (Hoenen, 2015, 2018), UR (Koppel et al., 2016)) effectively handle oral transmission. What happens if texts are modified through rephrasing rather than by textual errors? Moreover, what happens if the text is altered not only by paraphrasing but also by the injection of sociocultural bias?

While computational reconstruction techniques struggle in complex or non-textual scenarios, basic inferences may be possible under the simplified case of Immediate Textual Variation (ITV), *i.e.*, when one text is directly transmitted to a receiver. This scenario is simpler because it bypasses the need to reconstruct the *stemma*, a common step in text reconstruction.

A preliminary hypothesis for ITV states that the hyparchetype's embedding lies near the mean or median of its transmitted variants. Does this hypothesis accurately capture the ITV process? If so, the hyparchetype in vector space could be estimated directly from the observed copies using these statistical metrics.

In this paper, we experimentally evaluate whether basic statistics—specifically the mean and median of the corresponding text vectors—can approximate the ITV hyparchetype. To do so, we simulate one-step text transmissions with a known ground truth in a controlled environment. We propose using Large Language Models (LLMs) as agent-writers to generate the variant texts. Although human and machine text production differ, this setup serves as a preliminary step toward more realistic transmission simulations.

The paper is organized as follows: Section 2 reviews related work. Section 3 describes our methodology. Section 4 presents the results, and Section 5 discusses the key findings. A summary appears in Section 6.

2 Related Work

As noted in the introduction, computational approximations for complex text transformations re-

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 25–31, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-002

main scarce. Nevertheless, existing studies include: (1) computational simulations of social behavior; (2) computational simulations of text transmission; and (3) simulations of text transmission using Large Language Models (LLMs), a subtopic of (2). This work is part of the broader effort to use computational simulations with LLMs to model social phenomena. Some of these approaches have been applied to textual data, but there is a lack of recent literature on more complex forms of textual transmission, such as paraphrasing or bias injection.

2.1 Computational simulations of social behavior

Agent-based simulation is one of the core concepts in computational social science (Hox, 2017). Recent advances in the development of LLMs have influenced agent-based simulations in computational social science studies (Thapa et al., 2025), a field also known as "automated social science" (Manning et al., 2024). For instance, (Gao et al., 2023) studied the propagation of information in the form of opinions and emotions.

Although LLMs have been adopted with enthusiasm (see, for example, (Ferraro et al., 2024; Zhang et al., 2025)), some findings highlight their limitations. One major constraint is that LLMs may not function as individual agents but rather as a "superposition of perspectives," effectively acting as a community (Kovač et al., 2023; He et al., 2024).

Despite these limitations and the lack of evaluations with real-world data (Larooij and Törnberg, 2025), LLMs have been used to test several hypotheses in social sciences, generating potentially weak but insightful observations about human behavior (Ma et al., 2024). Additionally, these experiments shed light on the nature of LLMs and have been employed to detect bias (Qi et al., 2025).

2.2 Computational simulations of text transmission

In the field of Computational Textual Criticism, algorithm testing has been carried out on experimentally created datasets, using either human subjects or computational methods. In the latter case, some authors implement random changes and word substitutions in a given text (Koppel et al., 2016; Gelein, 2021).

2.2.1 Simulations of text transmission using Large Language Models

Simulations of text transmission using large language models constitute a specialized branch within computational studies of textual diffusion. Despite the expansion of NLP research, few works address this specific application. Marmerola *et al.* (Marmerola et al., 2016) employ traditional NLP techniques—such as part-of-speech tagging—to generate text variations. More recently, Zammit (Zammit, 2024) adopted a similar experimental framework and leveraged the T5 transformer to introduce paraphrasing as a form of text alteration.

3 Methodology

3.1 Main experiments

In this paper, we explore the following scenarios of textual drift:

1. *Omissions*: These are closely related to the textual variation phenomena traditionally studied in Textual Criticism.

2. *Paraphrasing*: The phenomenon of paraphrasing has been scarcely examined in Textual Criticism, as it generates alterations that are difficult to trace.

3. *Bias injection*: A more complex form of textual drift linked to bias injection, where differing cultural backgrounds introduce significant changes during text transmission.

The principal question we address is the following: **Given different scenarios of textual drift, can we approximate the original text representation?**

In this study, we focus solely on immediate variations, without modeling full transmission processes such as hierarchical clustering. In other words, if we have access to multiple immediate variants of a single text, can we reconstruct its original representation?

To answer this question, we propose an experimental approach based on computational simulations. Alternatively, similar experiments could be conducted with human subjects. In both cases, extrapolation to uncontrolled environments remains implausible. However, these simulations may offer insights into textual drift in sociocultural contexts, where establishing a ground truth for comparison is challenging. For instance, one might use parallel corpora reflecting different perspectives on the same phenomenon, but it is unrealistic to rely on them to recover a hypothetical original text. In all scenarios, we generate a source text and apply various perturbation schemes to experimentally assess the feasibility of approximating its original representation.

In all experiments, we implement different LLMs $m_1, \ldots, m_l$ which act like *agents*. A different LLM m_0 act like a text generator. All the generated texts with m_0 were used as ground truth for experiments. In the first empirical setup (*omissions*), we performed the variations (6) by randomly deleting one sentence. For *paraphrases* and *bias injection*, we apply six different LLMs to produce the necessary changes.

Finally, we embedded all texts into a vector space using the Jina-v3 sentence transformer m_T (Sturua et al., 2024), which differs from the models used earlier. For visualization, we applied Uniform Manifold Approximation and Projection (UMAP) (Healy and McInnes, 2024) to reduce the embeddings to two dimensions. We then computed cosine similarities between each text representation and the central tendency vectors (mean and median) in the original latent space.

3.2 Data generation

Data generation was performed primarily with Claude 3.5 Haiku (as of 2025-07-18) (Anthropic, 2025). We generated 100 distinct texts that follow the structure of abstract news items by using the following rule:

$$H \rightarrow W_1\ W_2\ W_3\ W_4\ W_5\ W_6\ W_7 \qquad (1)$$

where H is the headline, W_1 is the actor, W_2 is the action, W_3 is the object, W_4 is the method, W_5 is the reason, W_6 is the location, and W_7 is the time. Each element W_i has ten predefined possibilities to enhance diversity. To avoid generation failures and ethical or legal issues, we employed abstract entities. For instance, actors include "the provisional council" or "a legislative junta." Headlines were sampled from a uniform distribution. For example,

> (1) A bipartisan committee repealed economic sanctions via back-channel negotiations to strengthen alliances within the federal archives amid rising tensions.

While the headline was randomly generated, the body of the news item was produced using Claude 3.5 Haiku with a Chain-of-Thought setup. For the prompts used in this article, please refer to the Code and Prompt Availability statement.

Model	Similarity
Mean	0.9977
Median	**0.9988**

Table 1: Average similarities from mean and median values to the original vector.

3.3 Variation induction

For paraphrases and bias injection, we used the following models:

1. Gemma3 (Team et al., 2025).
2. Llama 3 (Grattafiori et al., 2024).
3. Gemini 2.5 Flash (Comanici et al., 2025).
4. DeepSeek V3 0324 (Liu et al., 2024).
5. GPT-4o mini (Hurst et al., 2024).
6. Phi-4 mini instruct (Abouelenin et al., 2025).

We used 20 different prompts for paraphrases using a zero-shot setup to promote diversity.

In the case of bias injection, we considered only one abstract scenario: polarity. In the prompt, we ask the models to rephrase the text from either a negative or positive point of view. Three models (Gemma3, Llama3, and Gemini) were used to generate negative perspectives, while the remaining models generated positive rewritings.

4 Results

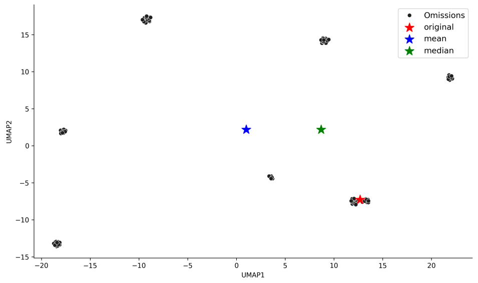
4.1 Omissions

Table 1 presents the average cosine similarities between the original text's latent representation and both the mean and median vectors. The median vector achieves a slightly higher similarity, and this difference is statistically significant (Mann–Whitney U test, $p < 0.05$).

Figure 1 visualizes 90 variations of the same text. Because it comprises only six sentences, only six unique variations (including the original non-variation) are displayed[1]. Unexpectedly, even in this constrained setting, the original representation does not lie exactly at the center of the resulting distribution.

[1] The visualization also includes the unaltered text.

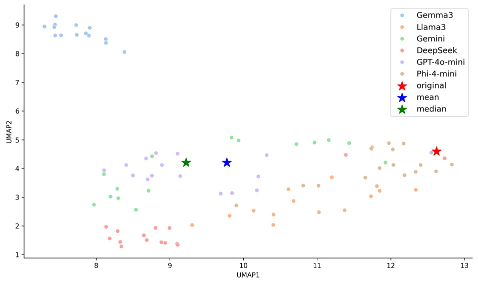
Figure 1: Visualization of variants of text (id: 0) using omissions.



4.2 Paraphrasing

Preliminary results with paraphrases (see figure 2) show that in some cases the original text might not be centered, and in the text drift might not diffuse in all directions as a classical physical system. Both the mean and median vectors are not the best estimators to the original text vector. In this particular case, the agent with less drift (Phi-4) was a best estimator.

Figure 2: Visualization of 90 variants of text (15 each model) using paraphrases and UMAP. In this particular case, the original text was generated with Qwen3-8B (Yang et al., 2025).



The quantitative analysis in the original latent space (see Table 2) indicates that, across all tested texts, both the mean and median vectors closely approximate the original representation. We observed significant differences between the mean/median vectors and every other model's outputs except those from Llama3 (Mann–Whitney U test, $p < 0.05$), while the mean and median themselves did not differ significantly (Mann–Whitney U test, $p = 0.6156$).

After reducing dimensionality to three via UMAP (cosine metric), texts generated by Llama3 exhibited even higher cosine similarities to the mean and median vectors (0.999935 and 0.999934, respectively) compared to 0.999721 in the original latent space. Figures 3 and 4 visualize specific cases with two dimensions.

Model	Similarity
Mean	0.9601
Median	**0.9609**
Gemma3	0.9290
Llama3	0.9540
Gemini	0.8454
DeepSeek	0.8467
GPT-4o-mini	0.8987
Phi-4-mini	0.8946

Table 2: Average cosine similarities from mean and median values to the original vector using paraphrases.

Figure 3: Visualization of variants of text (id: 3) using paraphrases.

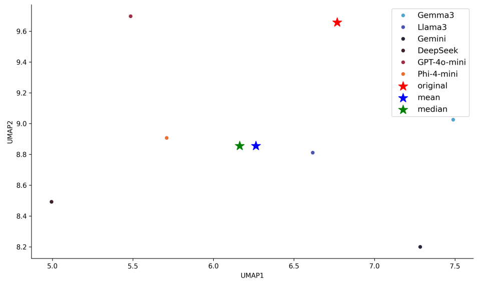
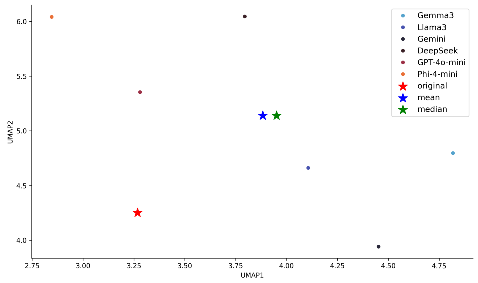


Figure 4: Visualization of variants of text (id: 16) using paraphrases.



4.3 Bias injection

In the bias induction experiments, the Phi-4 Mini model exhibited the least drift (see Table 3) and the highest average similarity compared to the centrality metrics, although its difference from the median vector was not statistically significant (Mann–Whitney U test, $p < 0.05$). We also observed significant differences between the mean and median vectors and all other groups, as well as between the mean and median themselves—except in the mean vs. GPT-4o Mini and median vs. Phi-4 Mini comparisons (Mann–Whitney U test,

Model	Similarity
Mean	0.9763
Median	0.9801
Gemma3	0.9291
Llama3	0.9447
Gemini	0.8699
DeepSeek	0.9382
GPT-4o-mini	0.9627
Phi-4-mini	**0.9807**

Table 3: Average cosine similarities from mean and median values to the original vector using bias injection.

$p < 0.05$). Figures 5 and 6 illustrate selected examples.

Figure 5: Visualization of variants of text (id: 22) using bias injection.

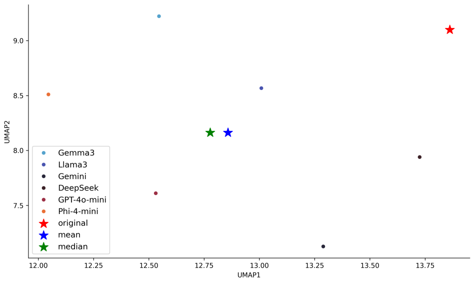
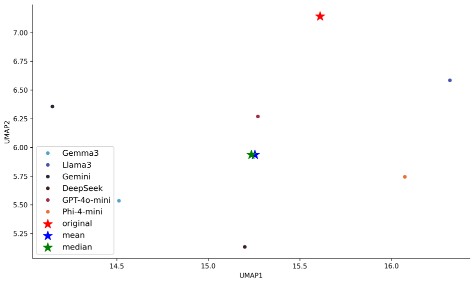


Figure 6: Visualization of variants of text (id: 44) using bias injection.



5 Discussion

This work constitutes a preliminary empirical approach to studying the process whereby a given text undergoes complex changes, in order to verify whether it is possible to retrieve information about the original from its immediate copies.

The main results discussed in section 4 provide a clear picture of the TIV simulation. In all cases, the median vector was a better estimator of the original vector than the mean. Both metrics outperformed most individual models. However, in some scenarios, we identified one model whose output drifted less and whose vector lay closer to the original than the median vector.

These findings offer two complementary insights. On one hand, the mean and median vectors can surpass most models in approximating the original text: by observing all surviving copies, we can reconstruct an approximation of the source, even in the presence of bias. On the other hand, it appears that one agent may introduce minimal variation, effectively serving as the best estimator of the original text. Identifying this agent may therefore be the optimal strategy for complex transmissions, in line with the Bédérian “best text” approach (Koppel et al., 2016).

The main results discussed in section 4 provide a clear picture of the TIV simulation. In all cases, the median vector was a better estimator of the original vector than the mean. Both metrics outperformed most individual models. However, in the paraphrasing and bias injection scenarios, we identified one model whose output drifted less and whose vector lay closer to the original than the median vector.

These findings offer two complementary insights. On one hand, the mean and median vectors can surpass most models in approximating the original text: by observing all surviving copies, we can reconstruct an approximation of the source, even in the presence of bias. On the other hand, it appears that one agent may introduce minimal variation, effectively serving as the best estimator of the original text. Identifying this agent may therefore be the optimal strategy for complex transmissions, in line with the Bédérian “best text” approach (Koppel et al., 2016).

6 Conclusions

This paper presents a one-step study of textual variation (ITV) across three scenarios: random textual changes (omissions), smooth semantic alterations (paraphrases), and bias injection (rephrasing with polarity). These scenarios correspond to three levels of textual alteration.

Studying complex changes in ITV is crucial for developing automatic methods for text reconstruction in contexts that go beyond purely random modifications while preserving text structure. These findings may apply to situations where textual variation is mediated by bias or rephrasing.

Our results show that basic central tendency statistics—particularly the median vector—are effective estimators of the original text vector. However, when paraphrasing occurs, certain agents introduce minimal drift and may serve as better estimators. This observation aligns with the classical philological proposals of Joseph Bédier.

6.1 Limitations

Using LLMs to simulate textual variation oversimplifies real-world scenarios. As discussed, LLMs do not act as individual agents but rather aggregate collective perspectives. Nevertheless, the growing role of artificial agents as (re)-writers clearly indicates that this study can be directly applied to understanding one aspect of material reality.

Acknowledgment

The authors would like to thank the anonymous reviewers.

Funding

This work has been possible thanks to the support of the Mexican government through SECIHTI, and the Instituto Politécnico Nacional research grants SIP 2259, SIP 20250015; IPN-COFAA and EDI.

Availability of data and materials

The data can be found in `https://github.com/Pherjev/ITV` or cited in the article.

Code and prompt availability

The code and prompts can be found in `https://github.com/Pherjev/ITV`

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Microsoft's Copilot in order to improve writing in English, as it is not our native language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*.

Anthropic. 2025. Claude 3.5 haiku. Accessed: July 20, 2025.

Alberto Blecua. 1983. *Manual de Crítica Textual*. Castalia, Madrid.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. 2024. Agent-Based Modelling Meets Generative AI in Social Network Simulations. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 155–170. Springer.

Walter M Fitch. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, 20(4):406–416.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S^3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Mees Gelein. 2021. *Simulating Mistakes. Using Agent Based Models to simulate and study the effects of scribal errors in classical text transmission*. Ph.D. thesis, Leiden University.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

James K He, Felix PS Wallis, Andrés Gvirtz, and Steve Rathje. 2024. Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*.

John Healy and Leland McInnes. 2024. Uniform Manifold Approximation and Projection. *Nature Reviews Methods Primers*, 4(1):82.

Armin Hoenen. 2015. Lachmannian Archetype Reconstruction for Ancient Manuscript Corpora. In *Proceedings of the 2015 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1209–1214.

Armin Hoenen. 2018. From Manuscripts to Archetypes through Iterative Clustering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Joop J Hox. 2017. Computational Social Science Methodology, Anyone? *Methodology*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Moshe Koppel, Moty Michaely, and Alex Tal. 2016. Reconstructing Ancient Literary Texts from Noisy Manuscripts. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 40–46.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large Language Models as Superpositions of Cultural Perspectives. *arXiv preprint arXiv:2307.07870*.

Maik Larooij and Petter Törnberg. 2025. Do Large Language Models Solve the Problems of Agent-Based Modeling? A Critical Review of Generative Social Simulations. *arXiv preprint arXiv:2504.03274*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, et al. 2024. Computational Experiments Meet Large Language Model Based Agents: A Survey and Perspective. *arXiv preprint arXiv:2402.00262*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated Social Science: Language Models as Scientist and Subjects. Technical report, National Bureau of Economic Research.

Guilherme D Marmerola, Marina A Oikawa, Zanoni Dias, Siome Goldenstein, and Anderson Rocha. 2016. On the Reconstruction of Text Phylogeny Trees: Evaluation and Analysis of Textual Relationships. *PloS one*, 11(12):e0167822.

Alberto Bernabé Pajares and Felipe Hernández. 2010. *Manual de crítica textual y edición de textos griegos*, volume 32. Ediciones Akal.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. Representation Bias in Political Sample Simulations with Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267.

Teemu Roos, Tuomas Heikkilä, and Petri Myllymäki. 2006. A Compression-Based Method for Stemmatic Analysis. *Frontiers in Artificial Intelligence and Applications*, 141:805.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Darren Zammit. 2024. *Computational Stemmatology: Reconstructing Text Phylogenies through Computer Assisted Methods*. Ph.D. thesis, University of Groningen.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. 2025. SocioVerse: A World Model for Social Simulation Powered by LLM Agents and A Pool of 10 Million Real-World Users. *arXiv preprint arXiv:2504.10157*.