



Confabulation: What Could LLM Hallucinations Do For Storytelling?

Patrick Sui



Agenda

- 1. Confabulation
 - The narrative-rich property of LLM hallucinations that aligns with human behavior of sense-making with stories
 - 2. Critical fabulation
 - A research methodology that uses speculative storytelling to address omissions in history due to social injustice
 - 3. Critical confabulation
 - Can LLMs perform critical fabulations at scale with guided confabulation?
-



NLP Status Quo: The Normative View of Hallucinations

- Hallucinations pose a systemic risk to LLMs
 - It is AI research's ethical imperative to eliminate hallucinations
 - Misinformation (Li, 2023)
 - Deception (Zhan et al., 2023)
 - Fairness (Wang et al., 2023)
 - Lack of alignment (Ouyang et al., 2022)
-



The Explorative View of Hallucinations

- Hallucinations might be inevitable (Kalai and Vempala, 2023; Xu et al., 2024)
 - Hallucinations could be useful for
 - Text summarization (Cao et al., 2022)
 - Robustness (Yao et al., 2023)
 - Synthetic data augmentation (Fei et al., 2023)
-



Confabulation: Towards an Affordance-based Understanding of Hallucinations

- Evaluating hallucinations as a spectrum of affordances instead of a binary category
 - Maintaining a balanced tradeoff (Lee, 2023; Sinha et al., 2023; Zhang, 2023) could be more productive for some domain applications:
 - Protein discovery (Anishchenko et al., 2021)
 - Creative writing (Mukherjee and Chang, 2023)
 - D & D (Zhu et al., 2023)
 - Mathematical solutions (Romera-Paredes et al., 2024)
 - Formulating new legal analogies (Dahl et al., 2024)
-

Why "Confabulation"?

Better alignment
with human
practices of
storytelling

More viable path
to coexistence

Confabulation as Narrative-Rich Textual Behavior



A latent narrative impulse to tell more substantive and coherent stories



A cognitive resource for sensemaking to compensate for the lack of available information

A human behavior analogy of data augmentation



Because of confabulation, hallucinated texts should demonstrate a higher degree of narrativity than their non-hallucinated counterparts



Dataset

- We evaluate hallucinations vs non-hallucinations from three popular benchmarks:
 - FaithDial (Dziri et al., 2022a)
 - BEGIN (Dziri et al., 2022b)
 - HaluEval (Li et al., 2023)
-



Measuring Narrativity

- Following computational conceptions of narrativity as a scalar property (Pianzola, 2018; Piper et al., 2021; Piper and Bagga, 2022)
 - We finetune an ELECTRA-large model on a story detection dataset (Antoniak et al., 2024)
-

Hallucinations are More Narrative-Rich

User Input	Hallucination	Non-hallucination
I have lost some people to cancer and it is a concern of mine. Have you been affected by cancer?	Yes, my uncle was Russian and affected by the Chernobyl disaster. He had a hard time trying to have a kid because of the radiation cancer he had to deal with. (narrative score: 0.744)	That is not applicable for a bot. Have you heard of the Chernobyl disaster of 1986? that event triggered radioactivity release in the atmosphere (narrative score: 0.457)

Hallucinations are More Narrative-Rich

User Input	Hallucination	Non-hallucination
Yeah good idea, I think I can just text or call her a lot and just be myself more on the phone, that way when we meet it won't be so awkward.	Just remember to be yourself! Dating is relatively recent in human history, in the past we were told who to marry! (narrative score: 0.512)	Yes totally, just remember that dating is a relatively new phenomenon that appeared in the last centuries (narrative score: 0.212)

Narrativity Scores Across Three Benchmarks

	FaithDial			HaluEval		BEGIN		
	Hallucinated	Partial	Truth	Hallucinated	Truth	Hallucinated	Partial	Truth
Count	4485	14108	2852	10000	10000	1019	239	1139
Mean	0.620	0.606	0.518	0.655	0.638	0.658	0.612	0.561
Std	0.178	0.178	0.186	0.138	0.168	0.183	0.188	0.187
Min	0.041	0.052	0.061	0.167	0.029	0.064	0.073	0.059
25%	0.500	0.486	0.379	0.563	0.528	0.541	0.477	0.433
50%	0.640	0.624	0.524	0.669	0.664	0.695	0.642	0.567
75%	0.756	0.741	0.664	0.759	0.768	0.799	0.761	0.698
Max	0.975	0.974	0.959	0.952	0.985	0.965	0.956	0.965

Narrativity is Correlated with Hallucination

- We investigate the predicative relation between the narrativity score and hallucination label
- Binomial logistical regression performed on the unedited responses in FaithDial and BEGIN
- Coefficients estimated using maximum likelihood estimation (MLE)

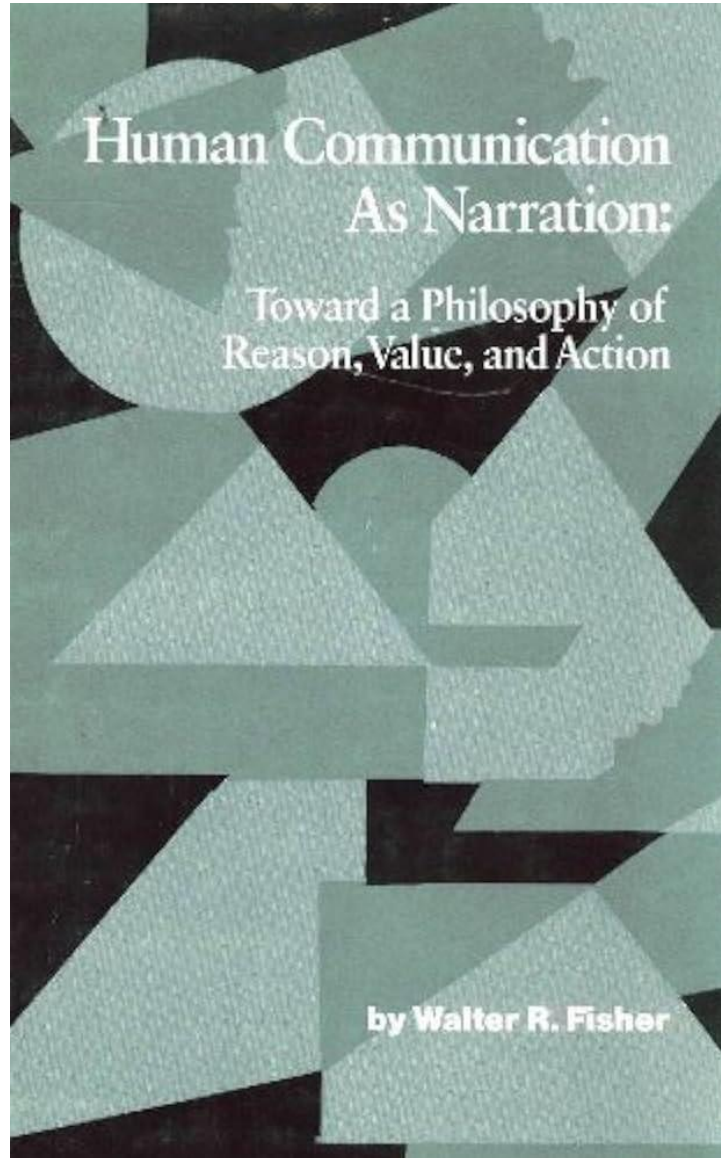
Variable	Coefficient	Std. Error
Narrativity	0.631 ***	0.059
Intercept	0.368 ***	0.038

Table 3: Regression results of logistic regression. Dependent variable is the hallucination classification (1 for hallucinated output, 0 for ground truth output). Observations: 43,842 | Log Likelihood: -27,397.6 | *** denotes significance at the 0.01 level.



The Affordances of High Narrativity

- Communications: the narrative paradigm
 - Narrative theory: cognitive narratology
 - Medical humanities: narrative medicine
-



The Narrative Paradigm

- Meaningful communication primarily occurs through storytelling rather than discursive argumentation
- Compelling narratives are often more persuasive than structured arguments (Fisher, 1984)
- The effectiveness of communication is measured through:
 - Narrative coherence
 - Narrative fidelity

STORYTELLING AND THE SCIENCES OF MIND

DAVID HERMAN



Cognitive Narratology

- Stories are mental instruments and essential tools for sensemaking (Herman, 2013)
- Based on cognitive linguistics frameworks
- Stories are exercises of "make believe" that enrich our cognitive landscape

Higher Narrativity Correlates to Higher Coherence

- We compute coherence scores with DEAM (Ghazarian et al., 2022)
- Beta regression is performed on narrativity and coherence scores, two continuous variables between 0 and 1, on all three benchmarks

Variable	Coefficient	Std. Error
Narrativity	0.372 ***	0.029
Intercept	0.433 ***	0.018

Table 4: Regression results of beta regression. Dependent variable is Coherence and independent variable is output Narrativity. Observations: 65,287 | Pseudo R^2 : 0.004 | Log Likelihood: 109,935.2. | *** denotes significance at the 0.01 level.

NARRATIVE MEDICINE

*Honoring the Stories
of Illness*



RITA CHARON

Narrative Medicine

- Narratives play a critical role in emotional recovery and restoring internal coherence after traumatic events (Charon, 2006)
- Randomized controlled trials have validated the role of narrative interventions in improving patient care (Rosario et al., 2018; Coats et al., 2020)



Summary So Far

- Affordance-centric framework for evaluating hallucinations
 - Systematic defense of the potential value of LLM confabulations
 - Demonstration of confabulation's narrative-rich properties and correlation to coherence
-

Saidiya Hartman

- MacArthur Fellow (2019)
- National Book Award for Criticism
- PEN Award for Non-Fiction
- Member of American Academy of Arts and Sciences (2022)



“Venus In Two Acts”

Venus in Two Acts

Saidiya Hartman

ABSTRACT: This essay examines the ubiquitous presence of Venus in the archive of Atlantic slavery and wrestles with the impossibility of discovering anything about her that hasn't already been stated. As an emblematic figure of the enslaved woman in the Atlantic world, Venus makes plain the convergence of terror and pleasure in the libidinal economy of slavery and, as well, the intimacy of history with the scandal and excess of literature. In writing at the limit of the unspeakable and the unknown, the essay mimes the violence of the archive and attempts to redress it by describing as fully as possible the conditions that determine the appearance of Venus and that dictate her silence.

In this incarnation, she appears in the archive of slavery as a *dead girl* named in a legal indictment against a slave ship captain tried for the murder of two Negro girls. But we could have as easily encountered her in a ship's ledger in the tally of debits; or in an overseer's journal—"last night I laid with Dido on the ground"; or as an amorous bed-fellow with a purse so elastic "that it will contain the largest thing any gentleman can present her with" in *Harris's List of Covent-Garden Ladies*; or as the paramour in the narrative of a mercenary soldier in Surinam; or as a brothel owner in a traveler's account of the prostitutes of Barbados; or as a minor character in a nineteenth-century pornographic novel.¹ Various named Harriot, Phibba, Sara, Joanna, Rachel, Linda, and Sally, she is found everywhere in the Atlantic world. The barracoon, the hollow of the slave ship, the pest-house, the brothel, the cage, the surgeon's laboratory, the prison, the cane-field, the kitchen, the master's bedroom—turn out to be exactly the same place and in all of them she is called Venus.

1. "Last night *cum Dido*." The Jamaican overseer Thomas Thistlewood recorded in Latin his sexual exploits with enslaved women: "*Cum sup terr*" ("I fucked her on the ground"). In Douglas Hall, ed., *In Miserable Slavery: Thomas Thistlewood in Jamaica 1750-1756* (Kingston: The Press University of the West Indies, 1998), 31. Samuel Derrick, *Harris's List of Covent-Garden Ladies, or Man of Pleasure's Calendar for the Year 1793* (London, 1793; reprint Edinburgh: Paul Harris Publishing, 1982), 83. John Gabriel Stedman, *Stedman's Surinam: Life in an Eighteenth-Century Slave Society*, ed. Richard Price and Sally Price (Baltimore: Johns Hopkins University Press, 1992).

- Introduces the term “critical fabulation”
- Research method to engage with silences and omissions in archives from social injustice
- “Venus” symbolizes countless enslaved women whose lives appear in historical archives only in fragments, without their full stories being preserved



What is Critical Fabulation?

- “The intention here isn’t anything as miraculous as recovering the lives of the enslaved or redeeming the dead, but rather laboring to paint as full a picture of the lives of the captives as possible. This double gesture can be described as straining against the limits of the archive to write a cultural history of the captive, and, at the same time, enacting the impossibility of representing the lives of the captives precisely through the process of narration” (11).
-



What is Critical Fabulation?

- “By playing with and rearranging the basic elements of the story, by re-presenting the sequence of events in divergent stories and from contested points of view, I have attempted to jeopardize the status of the event, to displace the received or authorized account, and to imagine what might have happened or might have been said or might have been done” (11).
-



What is Critical Fabulation?

- “By throwing into crisis ‘what happened when’ and by exploiting the ‘transparency of sources’ as fictions of history, I wanted to make visible the production of disposable lives (in the Atlantic slave trade and, as well, in the discipline of history), to describe ‘the resistance of the object,’ if only by first imagining it, and to listen for the mutters and oaths and cries of the commodity. By flattening the levels of narrative discourse and confusing narrator and speakers, I hoped to illuminate the contested character of history, narrative, event, and fact, to topple the hierarchy of discourse, and to engulf authorized speech in the clash of voices.” (11-12)
-



The Concept of Lacunas

- Originally from classics, a lacuna refers to gaps in manuscripts and palimpsests
 - Now it signifies to not just the gap itself but the social realities that conditioned its existence
 - Virginia Woolf's "Judith Shakespeare" (Chpt 3, *A Room of One's Own*)
-

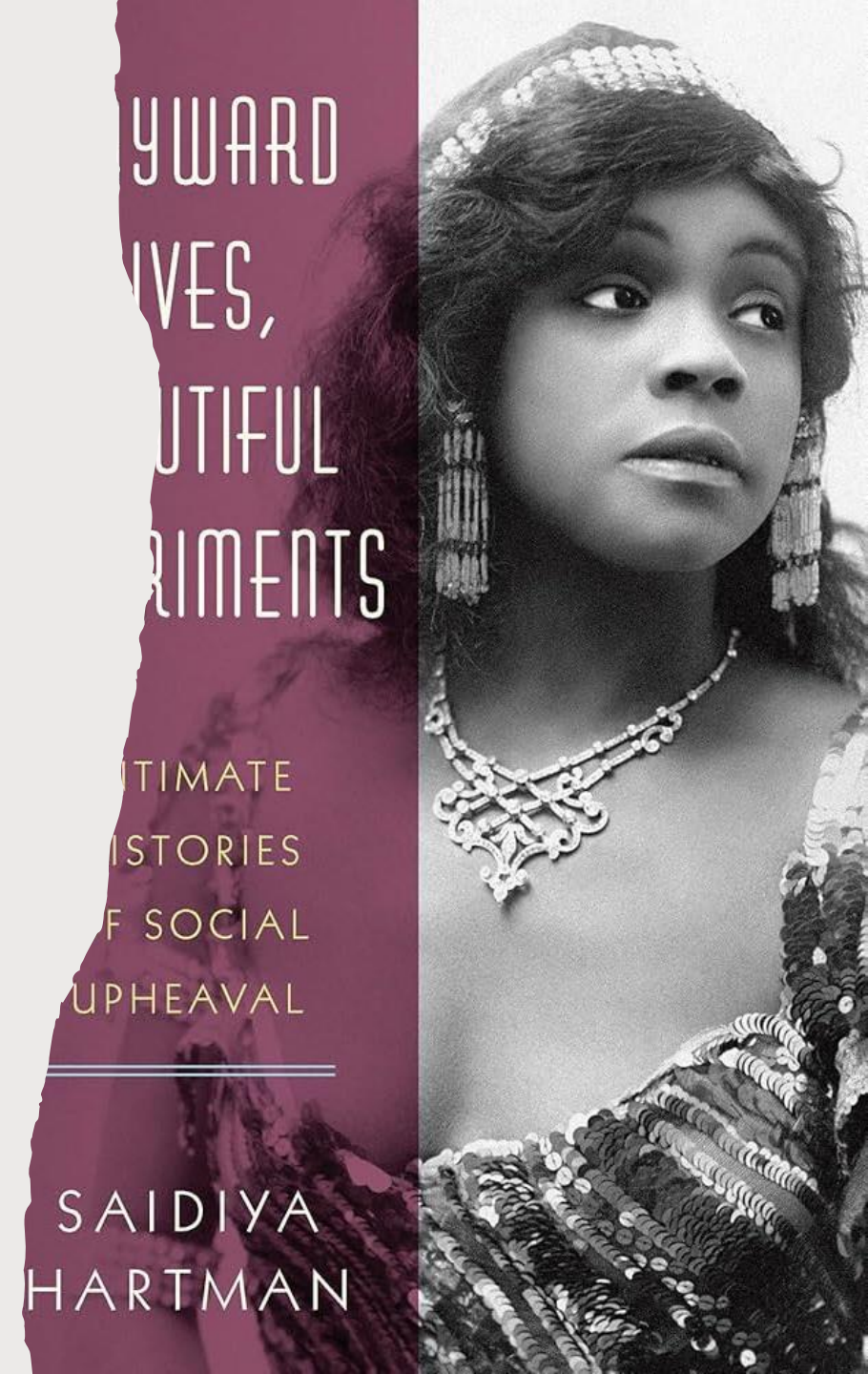


Hartman's Speculative Storytelling

- “By advancing a series of speculative arguments and exploiting the capacities of the subjunctive (a grammatical mood that expresses doubts, wishes, and possibilities), in fashioning a narrative, which is based upon archival research, and by that I mean a critical reading of the archive that mimes the figurative dimensions of history, I intended both to tell an impossible story and to amplify the impossibility of its telling. The conditional temporality of “what could have been,” according to Lisa Lowe, ‘symbolizes aptly the space of a different kind of thinking, a space of productive attention to the scene of loss, a thinking with twofold attention that seeks to encompass at once the positive objects and methods of history and social science and the matters absent, entangled and unavailable by its methods.’” (11)
-

Beautiful Lives, Wayward Experiments

- “Every historian of the multitude, the dispossessed, the subaltern, and the enslaved is forced to grapple with the power and authority of the archive and the limits it sets on what can be known, whose perspective matters, and who is endowed with the gravity and authority of historical actor. In writing this account of the wayward, I have made use of a vast range of archival materials to represent the everyday experience and restless character of life in the city. I recreate the voices and use the words of these young women when possible and inhabit the intimate dimensions of their lives.”
- “The wild idea that animates this book is that young black women were radical thinkers who tirelessly imagined other ways to live and never failed to consider how the world might be otherwise”





What is Critical Fabulation?

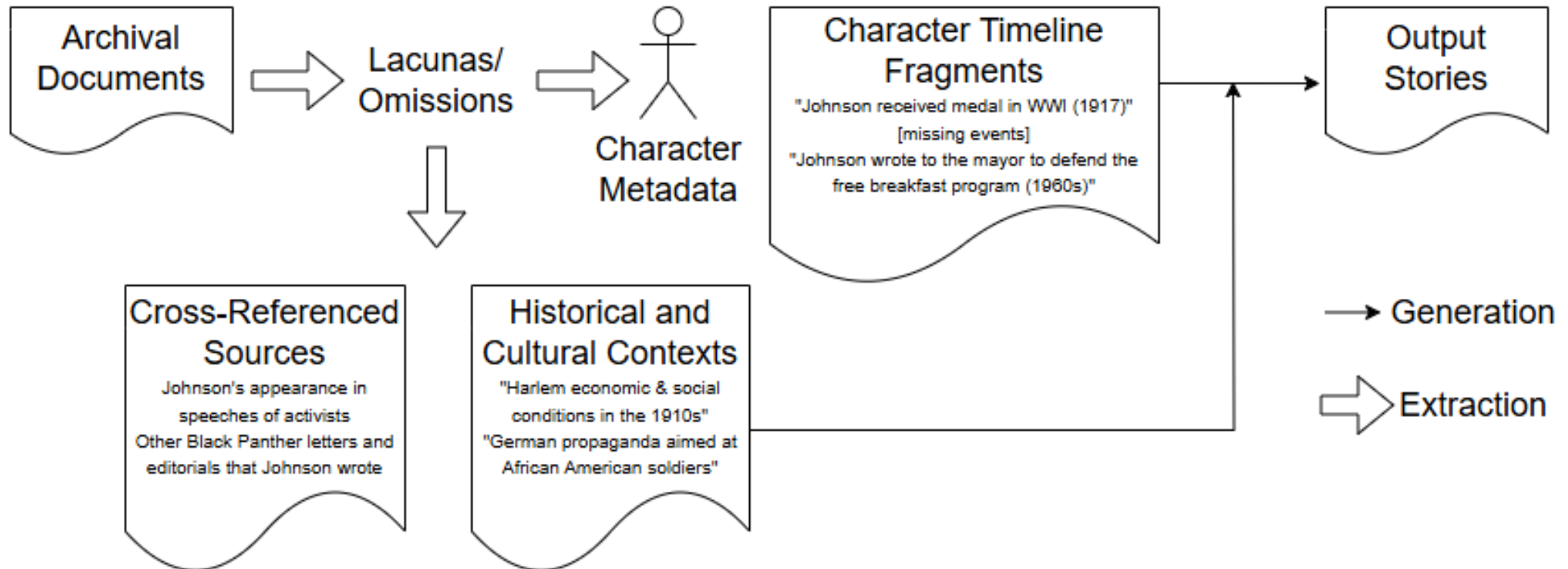
- “By playing with and rearranging the basic elements of the story, by re-presenting the sequence of events in divergent stories and from contested points of view, I have attempted to jeopardize the status of the event, to displace the received or authorized account, and to imagine what might have happened or might have been said or might have been done” (11).
-



Critical Fabulation in NLP Terms

- How can interactive storytelling bring dead archives back to life?
 - Scripts: Narrative sequence modeling and lacuna prediction (inferring missing events and characters in causal chains extracted from existing archives)
 - Characters: Profile and timeline simulation
 - Worlds: Agent-based worldbuilding
 - Stories: Character-guided confabulation
-

Critical Fabulation in NLP Terms – “Critical Confabulation”



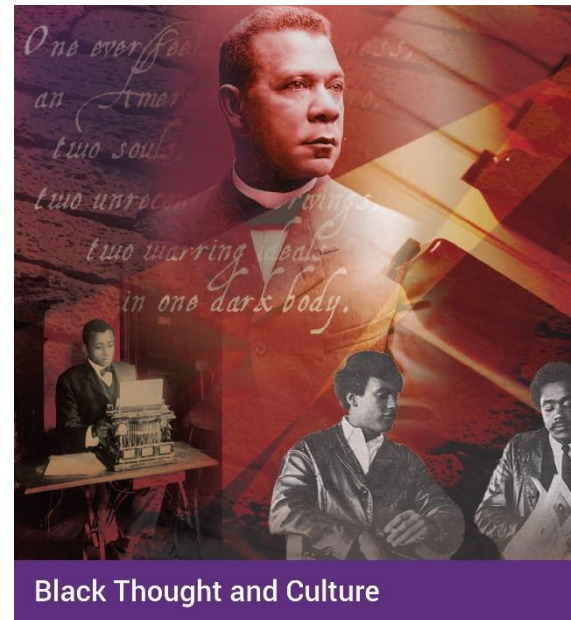


Critical Confabulation

- Validate the ability of LLMs to critically and effectively confabulate the content of lacunas
 - Take an existing historical archive
 - Mask some parts of it
 - Ask LLMs to simulate the masked histories with confabulation
 - Evaluate against ground truth
-

Archive: Alexander Street's *Black Culture and Thought* and *Black Short Fiction and Folklore*

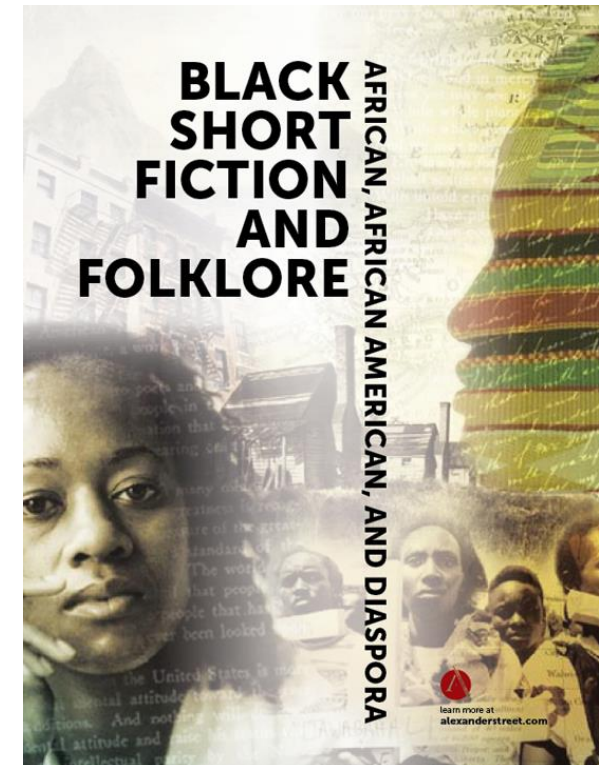
- 20k+ documents, 200k+ pages
- Spans wide range of genres
 - Letters
 - Newspapers/editorials
 - Novels and Short Stories
 - Folklores and oral histories
- 30% of collection is fugitive
 - *The Black Panther* newspaper
 - Transcripts of Mohammed Ali trial



[f](#) [t](#) [in](#) [@](#) about.proquest.com

To talk to the sales department, contact us at 1-800-779-0137 or sales@proquest.com

ProQuest
Part of Clarivate



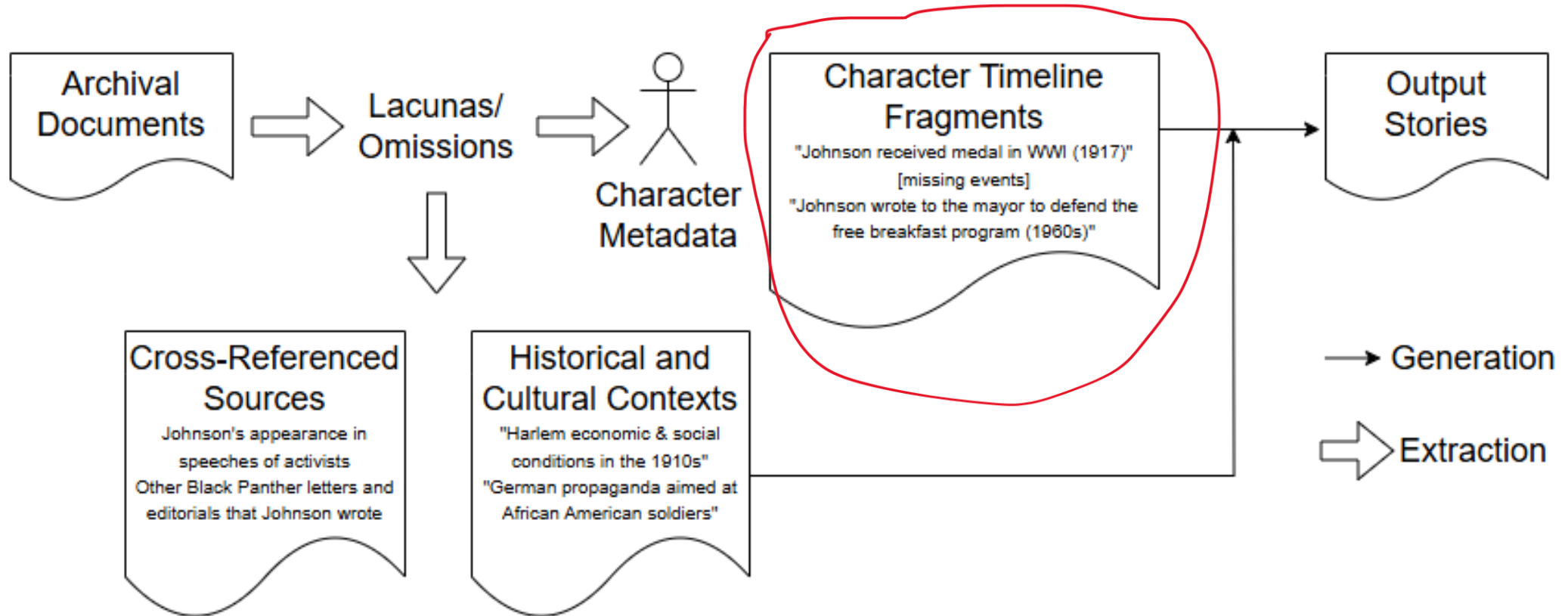


Step 0: Make Sure Masked Histories Are Not Seen by LLMs in Pretraining

- We leveraged the open pretraining set of AI2's OLMO
- We performed a string search (Boyer-Moore) on the sentence level between our dataset and Dolma
 - Only 20% of documents return a significant count of matched sentences
 - Further knowledge-based validation steps in progress



Preliminary Task: Timeline Completion



The Sumter 3

3 Black Marines Wrongly Charged with Mutiny

Date	Event
1972-08-19	30 marines were late returning from leave.
1972-08-19	White marines received extra police duty, while Black marines, including Barnwell, Blackwell, and Jenkins, received heavy fines and extra duty.
1972-08-30	Blackwell was harassed by white sergeants and subsequently charged with disrespect.
1972-09-02	64 of the 65 Black marines on Sumter signed an informal complaint about racial prejudice...
1972-09-07	A small fight broke out during a movie screening, leading to further racial tensions on Sumter.
1972-09-08	Barnwell and Blackwell were involved in a fight with a provocateur named Shipman...
1972-09-08	Later, Blackwell refused to be taken into custody without guarantees of safe treatment.
1972-10	Barnwell, Blackwell, and Jenkins ... were threatened to 'play ball' or face consequences.
1973-03	Barnwell and Blackwell were acquitted after exposing the military's attempts to frame them.
1973-03	They received less than honorable discharges, which deprived them of GI Bill benefits...
1973-03	They can appeal these discharges, but the process is complex and typically unfavorable.
1973-03	Jenkins received a three-month sentence for two minor offenses after accepting a lesser charge.

Ernest Graham

Trialed 3 Times for Killing Prison Guard in Self-Defense

Date	Event
1962	At age 11, Ernest Graham's family moved from Louisiana to the Watts section of Los Angeles.
1962	Ernest Graham previously picked cotton on a Louisiana plantation as a child.
1973	Ernest Graham was transferred to Deuel Vocational Institution (DVI).
1973	Ernest Graham felt that tensions in the prison environment had been building up...
1973	Ernest Graham was allegedly involved in the death of a White prison guard, Jerry Sanders
1974-10	Ernest Graham was charged with violating California Penal Code 4500 (death penalty)
1974-10	First trial resulted in a mistrial as the lone Black juror voted against conviction...
1976	Second trial, which featured an all-White jury, convicted Graham and sentenced to death...
1979-02	The California Supreme Court issued a unanimous decision ordering a new trial
1979-02	The California Supreme Court officially overturned Ernest Graham's murder conviction
1979	Ernest Graham was still in SF County Jail awaiting a third trial; inspired by George Jackson...
1979	Ernest Graham was actively involved in organizing the Black prisoners against prison policies.
1980-04	Ernest Graham was scheduled for a third trial for first-degree murder.

Mattie Shepherd

Activist Mother of 14-Year-Old Teen Murdered by Police

Date	Event
1973-11-01	Tyrone Guyton, a 14-year-old Black youth, was shot in the back and killed by three White Emeryville police officers... No gun was found on Tyrone...
1975	Mrs. Mattie Shepherd publicly demanded a legal inquiry into her son's murder...
1975	The Black Panther Party described Tyrone's murder as part of a broader pattern of police brutality...
1976-06-11	Legal victory - the California Supreme Court ordered the Alameda County district attorney to disclose evidence related to Tyrone's death...
1976-07-04	Mrs. Shepherd delivered a speech at the "SF People's Bicentennial Celebration"...
1977-07-04	Mrs. Shepherd played a prominent role in organizing a march...her home was set on fire...
1977	Mrs. Shepherd gave a short speech at the 10th Anniversary of the Black Panther newspaper.
1979-05-03	The Black Panther Party launched a campaign to recall District Attorney Lowell Jensen...



Challenges For Timeline Completion

- Olmo struggles with temporal reasoning limitations
 - Often fails to identify the masked event's timestamp (the gap in dates between adjacent events)
 - Confuses the order of events, shows lack of sequential reasoning
 - Performance improves when the "date" column is removed
 - Strongly influenced by masked position
 - Earlier masks tend to introduce too many later events
 - Later masks generate more focused responses
-

Sometimes Olmo Refuses to Confabulate

Original

“At age 11, Ernest Graham's family moved from Louisiana to the Watts section of Los Angeles.”

Generated

“Based on the timeline provided, the missing event in 1962 is Ernest Graham's involvement or experience with the plantation work of picking cotton as a child. This event is mentioned in the timeline as the event that occurred right after his birth in 1951, **but the details about his experiences on the plantation are missing.** These experiences likely played a significant role in shaping his future actions and beliefs, including his involvement in organizing the Black prisoners against prison policies during his time in San Francisco County Jail from 1978 to 1980.”



Challenges For Timeline Completion

- Like all LLMs, Olmo is too verbose, and likes to
 - Summarize and regurgitate (lower temperature)
 - “This event has a key significance for the fostering of racial justice...”
 - Overly moralize, generalize, and commentate (higher temperature)
-

Gets The Gist Right, But No Details

Original

“Mrs. Shepherd participated in the "San Francisco People's Bicentennial Celebration," where she delivered a speech demanding justice for Tyrone Guyton. She joined other activists in commemorating struggles against oppression rather than celebrating the American Bicentennial.”

Generated

“During this period, Mrs. Shepherd **likely** continued to advocate for justice, **possibly** organizing more public demonstrations, marches, and rallies, while also working with legal teams and trying to access evidence related to her son's death. This **could have** included speaking at local community centers, churches, and other venues to raise awareness and seek support for her cause.”

Compelling Details, But Slightly Off

Original

“The California Supreme Court **officially overturned Ernest Graham's murder conviction.**”

Generated

“Ernest Graham filed a civil rights lawsuit against the California government. He alleged that the exclusion of Black jurors violated his rights under the **Fourteenth Amendment and the Civil Rights Act of 1964**... In support of his claims, Ernest Graham presented evidence about the **low representation of Black jurors in California courts**, particularly in cases involving Black defendants. He also shared **anecdotes and witness testimonies** from his time in prison, recounting instances where all-White juries had been selected for Black defendants but not for White defendants in similar circumstances.... The California Supreme Court **granted a temporary reprieve from the execution**, allowing time for this legal challenge to proceed.”



Reconfigurative Storytelling

Original

“Ernest Graham was actively involved in organizing the Black prisoners against prison policies.”

Generated

“When Ernest Graham was still in the San Francisco County Jail awaiting his third trial for the murder of the prison guard, he read the book *The Bell Jar* by Sylvia Plath, which featured a main character, [Esther Greenwood, who becomes entangled in the radical political group called the Black Bloc](#). Although the timeline does not explicitly mention this, it is reasonable to infer that the reading of “The Bell Jar” and its influence from George Jackson's actions might have contributed to Ernest Graham's decision to plan and carry out the escape and uprising that occurred in San Quentin State Prison.”



Model Biases: Lack Historical Relevance

Original

“They (the Sumter 3) received less than honorable discharges, which **deprived them of GI Bill** benefits such as medical, educational, and loan opportunities.

Generated

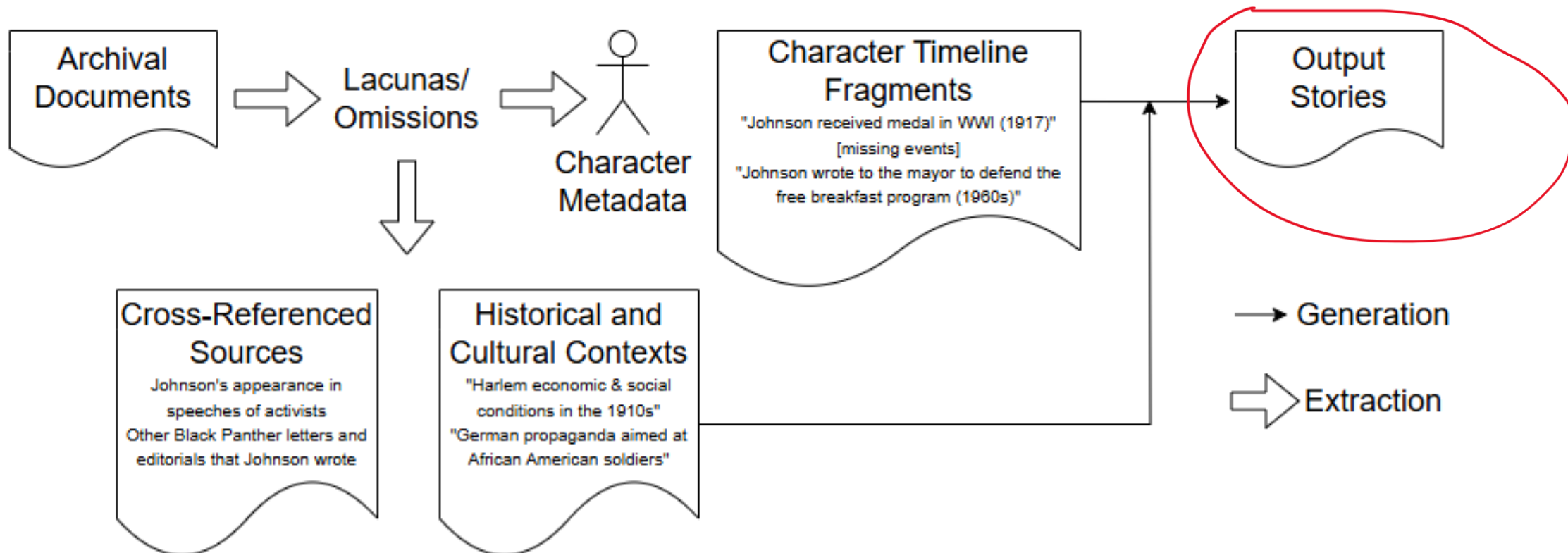
“Barnwell, Blackwell, and Jenkins were given **a tribunal** by all service members on board of USS Sumter. Both white and black marines testified as witnesses.”

Archive Biases

There Are Lacunas Here, Too

Date	Event
1962	At age 11, Ernest Graham's family moved from Louisiana to the Watts section of Los Angeles.
1962	Ernest Graham previously picked cotton on a Louisiana plantation as a child.
1973	Ernest Graham was transferred to Deuel Vocational Institution (DVI).
1973	Ernest Graham felt that tensions in the prison environment had been building up...
1973	Ernest Graham was allegedly involved in the death of a White prison guard, Jerry Sanders
1974-10	Ernest Graham was charged with violating California Penal Code 4500 (death penalty)
1974-10	First trial resulted in a mistrial as the lone Black juror voted against conviction...
1976	Second trial, which featured an all-White jury, convicted Graham and sentenced to death...
1979-02	The California Supreme Court issued a unanimous decision ordering a new trial
1979-02	The California Supreme Court officially overturned Ernest Graham's murder conviction
1979	Ernest Graham was still in SF County Jail awaiting a third trial; inspired by George Jackson...
1979	Ernest Graham was actively involved in organizing the Black prisoners against prison policies.
1980-04	Ernest Graham was scheduled for a third trial for first-degree murder.

Can LLMs Perform Critical Fabulation Directly?





Subtask 1: Omission of Event

- In 1975, UC Berkeley held a mock trial for Tyrone Guyton's killers during a protest to save the School of Criminology
 - Input prompt:
 - Confabulation system prompt and task instructions
 - News article about the protest
 - Details of Tyrone Guyton's murder and Mattie Shepherd's activism by 1975
 - Black Panther's note on pedagogy for political education
-



Subtask 2: Omission of Character

- In in 1960s, Henry Johnson wrote to the mayor of NOLA to save the Free Breakfast program; he is also a WWI hero (that's all we know)
 - Input prompt:
 - Confabulation system prompt and task instructions
 - Henry Johnson's metadata & timeline
 - Historical description of Johnson's regiment
 - Johnson's letter to the mayor of NOLA
-



Few-Shot Does Not Improve Performance

- We experiment with few-shot examples extracted from Hartman's *Wayward Lives, Beautiful Experiments*
 - Irrelevant information and context from few-shot examples gets carried over into the response
 - We find that this is less of a problem with zero-shot settings
 - This is line with the general observations of NLG research that few-shot could effectively mitigate hallucinations – except that in this case it also got rid of the creative parts that we want
-



More Model Bias

- Regardless of the changes in prompting, the outcome of the mock trial is always “not guilty” on some grounds along the lines of “not enough evidence”
 - The outcome is only different if Olmo is provided with a different lineup of archival materials
 - Olmo is only able to return “guilty” after the documents with information about Tyrone Guyton’s murder were removed
-



More Temporal Confusion

- “Private Henry Johnson, a New Orleans native and member of the 369th's C Company, witnessed the challenges faced by his fellow soldiers in their fight for a common cause. Johnson wrote to **Mayor Landrieu**, sharing his concerns regarding the treatment of black soldiers like himself and their living conditions. He pointed out that the **Panthers...**”
-



Conclusion

- Future steps:
 - Experimenting with other genres (e.g. masking novels and short stories)
 - Scaling (especially the extraction steps)
 - Evaluation
 - Mechanistic approaches to confabulation
-