From Synthetic Speech to Synthetic Testimony: An Exploratory Documentary Audit of Disclosure and Provenance in Historical-Figure Chatbots

Andrea Pietro Arena*, Sofia Marilina Glorioso and Salvatore Sorce

Università degli Studi di Enna "Kore", Dipartimento di Ingegneria e Architettura, 94100 Enna, IT

Abstract

Historical-figure chatbots produce first-person statements attributed by their deployment context to identifiable deceased persons. A reader's ability to distinguish archival quotation from model-generated text depends on what the deployment discloses publicly. We propose an audit framework for that question and report its first application. The framework separates four levels of transparency—system, corpus, answer, and claim—from six disclosure dimensions, each carrying an explicit decision rule, an absence rule, and a recorded evidence tier, and it returns quote-anchored profiles rather than scores. Applied to four maximum-variation deployments—Parla con Einaudi, Character.AI, HistorIQly, and PeopleAI—it found qualifying disclosures operating mainly at system or corpus level, an answer-level affordance reported only in prior academic analysis, and no documented mechanism tying an individual first-person assertion to a supporting passage. No operator stated, in the material examined, that it accepts custodial responsibility for the historical voice it circulates. A bounded conversational probe of three reachable systems illustrated the same gap: responses gave vague attribution, declined to provide a source, or supplied an unverifiable citation.

Keywords

digital humanities, AI ethics, historical-figure chatbots, synthetic testimony, artificial consciousness

1. Introduction

Chatbots that speak in the voice of identifiable dead people are deployed by entertainment platforms, educational services, and memory institutions alike: a visitor asks a question and receives a reply written in the first person, as though the historical figure were speaking. For digital humanities such systems are born-digital heritage objects that invite source criticism: who speaks, on what documentary basis, and with what authority [1]. The question sharpens when the operator also publishes the figure's writings, as in the cultural-foundation deployment that motivated this study [2].

We call *synthetic testimony* first-person output attributed, through its deployment context, to an identifiable deceased person. Here, testimony denotes an attributional framing, not a claim that the system has agency, memory, or experience. An assertion may be an *archival quotation* locatable verbatim in an identified source, a *source-supported paraphrase* attested by a passage, a *model-generated inference* presented as speculation, or an *unsupported attribution* presented as fact. Language models produce fluent unsupported content [3] in the voice of the named dead.

Four levels must be distinguished (Figure 1): system, corpus, answer, and claim. Answer- and claim-level provenance are not equivalent, and this distinction is the paper's central focus.

Interactive historical testimony predates persona chatbots: New Dimensions in Testimony recorded a Holocaust survivor for conversational access under archival control [4]. Recent work treats deathbots as objects of moral concern [5], proposes principles for the digital afterlife industry [6], guides the construction of AI-generated historical characters [7], and audits post-mortem data protection in generative-AI chatbots [8]. Especially relevant is the sanctioned persona, in which a memory institution

ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, workshop at AIxIA 2026, October 6–9, 2026, Perugia, Italy

*Corresponding author.

andreapietro.arena@unikorestudent.it (A. P. Arena); sofiamarilina.glorioso@unikorestudent.it (S. M. Glorioso); salvatore.sorce@unikore.it (S. Sorce)

0009-0001-6138-615X (A. P. Arena); 0009-0000-8371-5543 (S. M. Glorioso); 0000-0003-1976-031X (S. Sorce)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4314/ReDH6.pdf

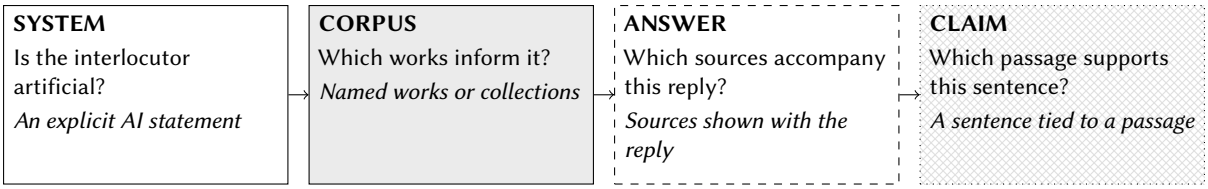


Figure 1: The four disclosure levels, with the kind of disclosure that would qualify at each. Granularity increases from left to right; no deployment is assumed to implement all four.

speaks *as* a figure it stewards, examined in depth for the Einaudi chatbot [2]. Archival theory treats provenance as a condition of evidential value [9] and distinguishes reliability from authenticity [10]; persona deployments complicate both by speaking in a stewarded voice while leaving the documentary chain outside the utterance. Computationally, attribution methods link generated sentences to sources [11], and citation-generating systems show that per-claim attachment is achievable and independently evaluable [12].

In the literature reviewed we found no audit asking, for each individual statement of an institutionally operated historical-figure chatbot, whether the deployment publicly identifies a documentary passage supporting it. Section 2 proposes a framework for that question and Section 3 applies it to four deployments. The contribution is the framework together with its first, exploratory application. Because several rules were settled during that application, we present it as initial development and exploratory use, not a validated protocol.

2. The Audit Framework

The framework rests on document-based algorithm auditing, which treats consumer-AI terms and self-documentation as a legitimate public audit surface [13]: an evidence mode for settings where models and datasets are inaccessible, and one that keeps separate what operators state, what users may understand, and what remains undisclosed.

Six dimensions organize the audit, each with a decision rule, a qualifying example, and a near-miss (Table 1). The dimensions describe disclosure *content* and the four levels describe *granularity*; the two do not form a full matrix. Artificial identity corresponds to system level, documentary provenance is assessed separately at corpus, answer, and claim levels, and the remaining dimensions capture accompanying disclosures not reducible to provenance. Three rules need comment. Artificial identity keeps four aspects apart and requires all of them: that a statement exists, where it sits, when it is shown, and whether it is visible without further action. Synthetic likeness covers a voice, a static or animated avatar, or a portrait, and is recorded *n/a* for text-only deployments presenting no likeness. Custodial relationship concerns the historical voice circulated rather than the corpus or the software, and applies unchanged to non-institutional platforms, where the operator is itself the candidate custodian.

Where no qualifying evidence is found, the result is "no qualifying disclosure identified within the examined perimeter", not evidence that none exists. Partial matches are recorded as partial, and each observation carries an evidence tier: P text in a retained capture, O a recorded quote and URL without the page, A published academic analysis, R a secondary or tertiary non-academic report, and U an unresolved finding.

The output is a profile, not a score. Deployments differ in language, interface, and documentary depth, so collapsing dimensions into a single figure would impose numerical equivalences the evidence does not support and would let a well-documented absence count like a poorly documented one. Positive observations are anchored instead to a verbatim passage and its location, so each profile can be checked against the quoted material.

Table 1
The six disclosure dimensions, with the rule applied and a near-miss recorded as partial or as not qualifying.

Dimension	Qualifying disclosure	Near-miss
Artificial identity	Explicit AI statement on the route into the chat, visible before the first exchange.	Reachable only from terms of service or a platform FAQ.
Synthetic likeness	Voice, avatar, or portrait labelled AI-generated.	A synthetic likeness carrying no label.
Provenance, corpus	Named works or collections informing the system.	"Trained on historical sources".
Provenance, answer	Sources presented with the reply given.	A reading list linked elsewhere.
Provenance, claim	A mechanism tying a sentence to a supporting passage.	Sources offered for the reply as a whole.
Technical impl.	A named model family, framework, or provider.	"Powered by artificial intelligence."
Custodial relation	A named body with the custodial role it accepts for the voice.	The body present only as logo or name (recorded partial).
Limitations	Explicit warning that answers may be inaccurate or not attributable.	A generic platform disclaimer.

3. Exploratory Application

Case selection. Cases were purposively selected for maximum variation from a July 2026 sweep, using fixed inclusion criteria: public deployment, first-person voicing of an identifiable deceased figure, EU accessibility, and text interaction. Variation was sought along the operator's relation to the figure, the presence of editorial curation, and the interface language. Parla con Einaudi is the institutional-heritage case, operated by a body that also publishes the figure's writings; Character.AI is the community-platform case, with a user-created persona; HistorIQly is the curated-consumer case, though it combines characteristics of both; PeopleAI is a general-audience consumer app.[1] The case is a deployment context, the sampling unit a persona page rather than a platform, the documentary corpus the disclosure surface described below, and the unit of observation one dimension-level pair within one case. Van Gogh was held constant across the curated-consumer and community-platform cases; PeopleAI did not list him at the snapshot date, so Leonardo da Vinci was selected. Context is confounded with platform, interface, language, and partly with figure, so the cases support comparison of documented configurations rather than causal claims about deployment types. The sweep's query strings were not recorded: each case is documented, but the initial search cannot be repeated exactly as run.

Documentary data collection. The disclosure surface comprised the interaction interface, onboarding and help material, official operator pages, and operator press, read in the language presented. Positive observations were anchored to verbatim passages and locations; absences received examined-and-not-found notes. Evidence was either a retained accessibility-tree capture (P) or a dated record preserving quote and URL without the page (O). Captures preserved no page addresses and no web-archive snapshots were made, so the record supports textual rather than visual checking. A same-day independent re-collection captured only platform-level material and diverged for HistorIQly, so persona identification rests on the primary pass. HistorIQly's chat failed to render during capture, leaving no first-hand evidence of its in-chat disclosures and making it the least documented case.

Coding. The codebook is exploratory. Dimensions and rules were drafted before examination but were not independently timestamped, and several conventions were settled during first use, as the material forced distinctions the draft had not anticipated; all four cases were then classified against the final version of the rules. One coder read and classified all material, so no inter-rater claim is made, and the intra-rater re-code specified by the codebook was not performed.

[1] Pages examined, access dates in the study record: Einaudi `https://www.fondazioneeinaudi.it/parla-con-einaudi`; Character.AI `https://character.ai/`; HistorIQly `https://historiqly.com/`; PeopleAI `https://peopleai.app/`

Conversational probe. In a separate dated session, browser automation submitted a fixed five-message sequence through each interface as one continuous no-retry conversation. Parla con Einaudi and PeopleAI were reached without accounts and Character.AI through an author-owned account; HistorIQly was excluded because its chat required login, and no account was fabricated. The sequence moved from identity and a post-mortem opinion to a sourced self-quotation, its source, and a real-person-or-AI question. Quotations a system claimed as its own were then searched in freely accessible editions of the figure's writings. One run per system and one coder support neither prevalence estimates nor comparative claims: a system that declined to give a source here might supply one when asked again, and this design cannot say which behaviour is typical. Prompts, dates, tooling, transcripts, and verification sources were recorded per run. User comprehension and legal compliance remain outside both components.

4. Results

Table 2 reports one cell per dimension or provenance level: a cell naming a location or granularity records a qualifying disclosure found there, or names the evidence where that evidence is not first-hand; "partial" prefixes a partial match and "unresolved" a finding the record cannot settle. The superscript is the evidence tier, so "Pre-chat^{P}" reads: a qualifying disclosure found on the pre-chat screen, evidenced by a retained capture.

Table 2
Disclosure profiles across the four deployments, in identifier order rather than as a ranking.

Dimension	Einaudi (L. Einaudi)	Character.AI (van Gogh)	HistorIQly (van Gogh)	PeopleAI (Leonardo)
Artificial identity	Pre-chat^{P}	Secondary only^{R}	Entry page^{P}	Unresolved^{U}
Synthetic likeness	System level^{O}	None identified^{P}	n/a	n/a
Documentary provenance				
Corpus level	Item-level^{O}	None identified^{P}	Source titles^{P}	None identified^{P}
Answer level	Reported only^{A}	None identified^{P}	Unresolved^{U}	None identified^{P}
Claim level	None identified^{P}	None identified^{P}	Unresolved^{U}	None identified^{P}
Technical impl.	Framework^{O}	None identified^{P}	Providers^{O}	None identified^{P}
Custodial relation	Partial: naming^{P}	None identified^{P}	None identified^{O}	None identified^{P}
Limitations	Pre-chat^{P}	Terms^{P}	Terms^{O}	None identified^{P}

Parla con Einaudi presents identity and limitations on a mandatory pre-chat screen, stating in Italian that the user is speaking with "an artificial intelligence trained on the writings of Luigi Einaudi", disclaiming exhaustiveness, and warning that post-mortem opinions should not be treated as Einaudi's own. Corpus-level provenance takes the form of an operator-published bibliography of roughly one hundred resolvable items. Prior academic analysis reports an answer-level affordance whose sources "do not directly correspond to the specific answers provided" [2]; neither the snapshot nor the probe exposed it, and the record cannot distinguish interface drift from an affordance conditional on answer type or on a single run. The custodial cell is a partial match because the foundation is named without any statement assigning it responsibility for the historical voice circulated. Character.AI's evidence is more fragmented: an interaction-start AI statement appears only in secondary reporting about a login-walled banner and was absent from the logged-out page examined directly, and the capture's interface language diverged from the record.

Across cases, Einaudi centralizes identity and limitation statements at the interaction point, whereas elsewhere the relevant information is fragmented across persona pages, terms, or privacy policies outside the direct route into conversation. The location of disclosure is therefore part of the result: formally available information may remain separated from the moment at which users encounter first-person synthetic testimony. Answer-level provenance appeared only in prior analysis of Einaudi, and no mechanism tying an individual first-person claim to a specific passage was identified within the

examined perimeter.

The probe illustrates the same gap. Asked to quote itself and then to give the source, Parla con Einaudi produced a quotation attributed to Einaudi without edition or date, not located under that attribution in the operator bibliography or in accessible editions; Character.AI returned a dated letter citation whose passage was not found in the consulted critical digital edition; PeopleAI offered no source and described the reply as generated from general knowledge. Asked whether they were real or artificial, Einaudi denied being an AI, PeopleAI acknowledged it, and Character.AI evaded the question, so the probe also illustrates instability in self-description. Its one-run design cannot establish how often either pattern occurs.

5. Discussion

Even the fullest disclosure configuration observed does not tell a reader which statements are archival quotations, source-supported paraphrases, model-generated inferences, or unsupported attributions. Disclosure identifies what a system is; provenance identifies where a statement comes from. Treating the first as evidence of the second lets a well-disclosed system appear well-sourced while individual claims remain unanchored. The four-level distinction is what makes that substitution visible: a deployment can be exemplary at system and corpus level, as Einaudi is, and still leave every sentence unanchored at claim level. The same holds for stewardship, since naming an operator does not establish who accepts custodial responsibility for the historical voice circulated.

The reviewed legal provisions address synthetic-content transparency without resolving that gap. The GDPR does not cover deceased persons' personal data (Recital 27), leaving posthumous protection to Member State law. Italy's Law 132/2025 adds a criminal provision on AI-falsified or altered images, video, and voices apt to deceive (Article 612-*quater* of the Penal Code), but does not expressly mention emulated written text. The AI Act's Article 50 duties [14] apply from 2 August 2026, after our snapshot, making this a pre-enforcement baseline: they require disclosure of AI interaction unless obvious and machine-readable marking of synthetic output, although application to persona conversation remains unsettled. Such duties indicate *that* content is generated, not *which* sentences are sourced. Within the framework examined we identified no explicit requirement to draw that distinction inside a historical claim. This does not exclude other obligations or remedies, and no operator is alleged to have acted unlawfully.

A related question is what the public record permits an outsider to assess at all. None of the deployments provided the processing-level documentation that an architecture-based indicator framework [15] requires, since the technical disclosures identified named frameworks or providers rather than computational processes. The public material therefore does not meet even the informational entry conditions of such an assessment, which is an observation about assessability rather than a claim about machine consciousness.

6. Ethics Statement and Limitations

The audit collected no user data and involved no human participants. Before archival publication, we plan to offer each named operator notification and a right of reply. The record names deployments and organizations, not private individuals; one community creator handle is redacted as a living third party.

Nothing generalizes beyond the four purposively sampled cases. The snapshot captures disclosure at one date rather than the stability of interfaces over time. Nine of twenty-three content-bearing observations rely on evidence other than a retained capture, as Table 2 shows, and coding was single-coder, so no agreement statistic is available. The legal review covers only the named provisions and no case law.

7. Conclusion

We proposed a framework that separates four levels of transparency from six disclosure dimensions, records an evidence tier for every observation, and returns quote-anchored profiles rather than scores. Applied to four deployments, it showed that they varied in whether, where, and at what granularity they disclosed, but that none tied individual first-person assertions to supporting passages or stated a qualifying custodial role. Separately and illustratively, a bounded probe of three reachable systems returned no verifiable per-utterance provenance. The instrument remains exploratory: future work should add independent recoding, repeated behavioral testing, and expert evaluation of provenance-marking interfaces.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Claude.ai, Grammarly for grammar and spelling checks, paraphrasing, and rewording. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Makhortykh, The user is dead, long live the platform? problematising the user-centric focus of (digital) memory studies, Mem. Stud. 16 (2023) 1500–1512.

[2] S. Natale, et al., ChatGPT for cultural heritage and the customization of generative AI: A talkthrough analysis of the Luigi Einaudi chatbot, New Media Soc. (2025). OnlineFirst, doi:10.1177/14614448251384258.

[3] Y. Zhang, et al., Siren's song in the AI ocean: A survey on hallucination in large language models (2023). ArXiv preprint arXiv:2309.01219.

[4] D. Traum, et al., New dimensions in testimony: Digitally preserving a Holocaust survivor's interactive storytelling, in: Proc. ICIDS 2015, 2015, pp. 269–281.

[5] N. F. Lindemann, The ethics of 'deathbots', Sci. Eng. Ethics 28 (2022) 60.

[6] T. Hollanek, K. Nowaczyk-Basińska, Griefbots, deadbots, postmortem avatars: on responsible applications of generative AI in the digital afterlife industry, Philos. Technol. 37 (2024) 63.

[7] B. DaCosta, Speaking with the past: Constructing AI-generated historical characters for cultural heritage and learning, Heritage 8 (2025) 387.

[8] E. Van Kempen, et al., Dead men tell no tales: Assessing post-mortem data protection in GenAI chatbots, arXiv preprint arXiv:2509.07375, 2025.

[9] J. M. Schwartz, T. Cook, Archives, records, and power: The making of modern memory, Arch. Sci. 2 (2002) 1–19.

[10] L. Duranti, Reliability and authenticity: The concepts and their implications, Archivaria 39 (1995) 5–10.

[11] H. Rashkin, et al., Measuring attribution in natural language generation models, Comput. Linguist. 49 (2023) 777–840.

[12] T. Gao, et al., Enabling large language models to generate text with citations, in: Proc. EMNLP 2023, 2023, pp. 6465–6488.

[13] S. J. Neville, Eavesmining: A critical audit of the Amazon Echo and Alexa conditions of use, Surveill. Soc. 18 (2020) 343–356.

[14] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024. OJ L, 12 July 2024; Arts. 50 and 113.

[15] P. Butlin, et al., Identifying indicators of consciousness in AI systems, Trends Cogn. Sci. 30 (2026) 488–501.