Uniwersytet
im. Adama Mickiewicza
w Poznaniu

Sensus
Historiae
ISSN 2082–0860
Vol. LXII (2026/1), s. 61-78
doi : 10.14746/sh.2026.62.1.004

AI's Personas and Historical Knowledge: How Models Transform Textbook Questions into Narratives

Konfiguracje AI i wiedza historyczna: jak modele przekształcają pytania podręcznikowe w narracje

Wiktor Werner
Adam Mickiewicz University Poznań
werner@amu.edu.pl
ORCID: 0000-0002-3004-6021

Abstract: This article offers a comparative, methodologically transparent pilot study of how different large-language-model configurations shape school-type historical narratives. Using a fixed prompt set (P1–P5) drawn from the textbook *Europa. Nasza historia* and four conditions—default ChatGPT baseline (C0), a 'scientist/analytical researcher' prompted ChatGPT (C1), NotebookLM with retrieval-augmented generation over a shared textbook source (C2), and an agentic system configured on the same knowledge base (C3)—we analyze outputs as discursive artefacts. The study combines four analytic perspectives: substantive accuracy and source anchoring, rhetorical profile and evaluative stance, formal readability and lexical texture, and logical/causal coherence. Quantitative proxies (sentence length, Polish-adapted Fog-type readability, lexical diversity via TTR and MTLD, nominalization and structural markers) are computed on the original Polish outputs without translation. Results reveal systematic trade-offs between anchoring and interpretive depth: retrieval-based and agentic configurations tend to produce a more textbook-like voice and stronger didactic scaffolding, while persona prompting increases meta-argumentation and criteria-driven evaluation. A critical-incident analysis further demonstrates that the appearance of grounding can coexist with hinge-fact failure, where a single erroneous historical anchor remains embedded in an otherwise coherent narrative. The article concludes with implications for history education, suggesting configuration-sensitive uses of LLMs and practical guardrails for classroom deployment.

Keywords: generative AI; large language models; history education; digital humanities; retrieval-augmented generation (RAG); agentic AI; prompting; narrative analysis; readability; hallucinations.

*Artykuł nadesłany: 24.03.2026; nadesłany po poprawkach: 05.06.2026; zaakceptowany: 09.06.2026.

Streszczenie: Artykuł przedstawia porównawcze badanie pilotażowe dotyczące tego, jak różne konfiguracje dużych modeli językowych kształtują szkolne narracje historyczne. Wykorzystując stały zestaw pytań (P1–P5) zaczerpnięty z podręcznika *Europa. Nasza historia* oraz cztery persony AI: domyślny ChatGPT jako baseline (C0), ChatGPT dopromptowany do roli „analitycznego badacza" (C1), NotebookLM z generowaniem wspomaganym wyszukiwaniem (RAG) na wspólnym źródle podręcznikowym (C2) oraz system agentowy pracujący na tej samej bazie wiedzy (C3), analizujemy odpowiedzi jako artefakty dyskursywne o mierzalnych własnościach formalnych i retorycznych. Ramę badania tworzą cztery perspektywy: poprawność merytoryczna i zakotwiczenie w źródle, profil retoryczny i aksjologia, formalna czytelność i „tekstura" leksykalna oraz spójność logiczno-kauzalna. Wskaźniki ilościowe (m.in. długość zdań, polsko-adaptowany wskaźnik czytelności typu Fog, różnorodność słownictwa TTR i MTLD, nominalizacja oraz markery struktury) liczone są na oryginalnych odpowiedziach w języku polskim, bez tłumaczenia. Wyniki pokazują systematyczne kompromisy między zakotwiczeniem a głębokością interpretacji: konfiguracje retrieval-based i agentowe częściej wytwarzają „głos podręcznikowy" i stabilne ramy dydaktyczne, natomiast promptowanie persony wzmacnia metarefleksję i ocenę opartą na kryteriach merytorycznych. Analiza krytycznego przypadku wskazuje ponadto, że pozór ugruntowania może współwystępować z błędem kotwiczącym (*hinge-fact failure*), gdy pojedyncza błędna informacja historyczna pozostaje osadzona w narracji spójnej lokalnie. Artykuł kończy się wnioskami dla dydaktyki historii, proponując użycie LLM zależne od konfiguracji oraz praktyczne „ścieżki" dla zastosowań szkolnych.

Słowa klucze: generatywna AI; duże modele językowe; dydaktyka historii; humanistyka cyfrowa; RAG; systemy agentowe; promptowanie; analiza narracji; czytelność; halucynacje.

1 Introduction

The dynamic development of technology based on self-learning algorithms, i.e. algorithms capable of performing tasks previously reserved for humans, which we have been observing since the beginning of the second decade of the 21st century, is a factor that is very strongly changing the rules of the game that have operated so far—also in such areas as the popularization of historical knowledge, schooling, and the teaching of history. Historical knowledge cannot be shielded from the impact of technological innovations—even if many of its practitioners remain attached to its traditional infrastructure: manuscript and printed sources preserved in archives, printed books, and traditional forms of teaching, both at school and at the academic level.

AI technology—just as it exerts a strong influence on very many domains of our civilization—will also change how historical knowledge functions, including the way it is taught. The reasons are fairly obvious, the first to mention being the accessibility and ease of using AI tools. Students use them, teachers use them, and increasingly so do people interested in history. In this sense, AI tools are a natural extension of digital tools and media, which gained enormous importance as a means of popularizing historical

knowledge at the turn of the 20th and 21st centuries (as confirmed by survey research indicating the very large role of digital media as the media chosen by amateur history enthusiasts to supplement their knowledge[1]). Due to the specifics of their mechanics, AI tools add particular significance to the problem of generating information automatically. While one can say of any medium that it is not merely a passive carrier but an active message,[2], in the case of AI tools this saying gains a special meaning.

The specificity of generative algorithms means that we are dealing with an increment obtained by extending an existing—initial—set of information with new data. This takes place as a statistical-geometric process. First, existing information is transformed into geometric entities (graphs, vectors), within which (statistical rules) describing the specificity of a given set are determined, and ultimately—under the influence of tasks (prompts)—the set is expanded with new data that (for the algorithm) constitute a continuation of existing vectors. From the user's side, this looks like a conversation with an intelligent entity to which one can direct a question about the causes of the outbreak of World War I and receive a sensible answer. The ease of this procedure explains its popularity. In schools, we are already dealing with students who choose a 'conversation' with an AI model over reading books or even browsing websites to obtain information. This obviously constitutes a challenge for the education system.[3]

From the teacher's perspective, it is very important both to understand how generative AI models actually work and to acquire knowledge of how this technology can be used to disseminate knowledge that will meet criteria accepted by the school and the curriculum. Negative phenomena associated with the operation of generative algorithms are usually linked to their spontaneous use by unprepared users.[4] In fact, however, there are ways to reduce risks and increase the benefits of using generative AI algorithms. The three most popular are: using a model with a RAG architecture, using an agent-based model, and adapting/fine-tuning a model for educational tasks by means of appropriate prompting. The research problem presented

[1] W. Werner, D. Gralik, A. Trzoss, *Media społecznościowe a funkcjonowanie wiedzy historycznej w Polsce. Raport z badań*, „Przegląd Archiwalno-Historyczny" 6 (2019), pp. 211-235.

[2] M. McLuhan, *The Medium is the Message*, [w:] M. G. Durham, D. M. Kellner (eds.), *Media and Cultural Studies: KeyWorks*, Wiley-Blackwell, Malden—Oxford—Chichester 2012, pp. 100-107.

[3] Kasneci E. et al., *ChatGPT for good? ...*, "Learning and Individual Differences" 2023, DOI: 10.1016/j.lindif.2023.102274.

[4] Cotton D.R.E., Cotton P.A., Shipway J.R., *Chatting and cheating...*, "Assessment & Evaluation in Higher Education" 2024, DOI: 10.1080/14703297.2023.2190148; Zawacki-Richter O. et al., Systematic review..., "IJETHE" 2019, DOI: 10.1186/s41239-019-0171-0

in this article will be to indicate the way in which four different types of AI 'personas'—a standard (baseline) AI model (ChatGPT) not subjected to modification or tuning, and three personas obtained using appropriate methods—transform and modify historical knowledge.

2 State of Knowledge and the Research Problem

An 'ordinary' (baseline) LLM model is, in the technical sense, a language model trained to predict subsequent tokens in a sequence (a token is a numerically operationalized unit: a character or a set of characters, e.g. a phrase, a word, or part of a word). In this perspective, the model's answer is the result of estimating the probability distribution of subsequent tokens conditioned on the supplied context (prompt + conversation history), rather than a reference to an explicit external knowledge resource[5] (such as a set of publications or a textbook). As a consequence, an LLM-only system may be linguistically convincing, but it does not have a built-in guarantee that the answer conforms to established knowledge[6]. A modification of this is a tool-augmented LLM, meaning a language model coupled with specific tools (e.g., a scanner)[7] that extend its capabilities beyond text generation. A variant of this model is RAG (Retrieval-Augmented Generation), where—before generation—grounding in a specific external corpus (e.g. textbooks, historical sources) is triggered, and retrieved fragments are incorporated into the answer context. In education, this enables anchoring explanations in texts required by the course program and supports citation practices and work with sources. Research on the use of RAG in academic tutoring points to its potential to reduce errors and increase instructional control, but also emphasizes that quality depends on the entire external grounding chain and on whether the corpus actually contains the needed information. In other words, RAG shifts the burden from "knowledge in the weights" to "knowledge in the corpus," although this does not completely eliminate the risk of faulty synthesis or mis-retrieval.[8]

AgenticAI, like RAG, includes an initial corpus of texts, but goes one step further: the model not only generates an answer from the supplied context,

[5] Huang L. et al., *A Survey on Hallucination in Large Language Models*, "ACM Computing Surveys" 2025, DOI: 10.1145/3703155.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Attention Is All You Need*, "Advances in Neural Information Processing Systems" 30 (2017), pp. 5998-6008.

[7] Schick T. et al., *Toolformer*…, "NeurIPS" 2023, DOI: 10.5555/3666122.3669119

[8] Lewis P. et al., *Retrieval-Augmented Generation*…, "NeurIPS" 2020, DOI: 10.5555/3495724.3496517

but also decomposes the task, decides when to trigger grounding (retrieval), how to iterate, and how to verify results in order to achieve optimal alignment with the source corpus. A survey of the agentic model emphasizes that 'agenticness' consists in a dynamic strategy of acquiring and controlling knowledge (improvement loops, evaluation), which has direct implications for teaching: the system can implement tutor policies / scaffolding (e.g. graduated hints, feedback format, blocking ready-made solutions) as part of its operating logic.[9]

The study carried out for the purposes of this article has the character of a comparative analysis of discourse and historical narrative generated by AI in four different modes of operation. The same set of questions (taken from the didactic apparatus of the textbook Europa. Nasza historia (grades 5–8)[10] was posed to four AI 'configurations' that differ in their generation method and/or persona setting:

- NotebookLM (RAG on the textbook Europa. Nasza historia),
- AgenticAI based on ChatGPT (agent-based, configured to work on the same textbook),
- ChatGPT_scientist (ChatGPT prompted toward the thinking style of an analytical researcher),[11]
- ChatGPT_baseline: ChatGPT in its default configuration, without additional persona/style instructions, launched in a new, clean session: a new chat, no previous context and no prompting history in that thread. The answers were generated solely on the basis of the question contents from the set below.

The questions asked came from different parts of the textbook Europa. Nasza historia:

- Does Alexander of Macedon deserve the epithet 'Great'?—Europa. Nasza historia, grade 5, p. 85.
- Present the disadvantages and advantages of the manorial-serf economy from the perspective of a peasant and a noble.—Europa. Nasza historia, grade 6, p. 43.
- Assess treating the nation as the highest ideal. Discuss what opportunities and threats for society this idea entails.—Europa. Nasza historia, grade 7.1, p. 49.

[9] Yao S. et al., *ReAct*..., "ICLR" 2023

[10] *Europa. Nasza historia*, Wydawnictwa Szkolne i Pedagogiczne—Eduversum, Warszawa 2022.

[11] Wei J. et al., *Chain-of-Thought Prompting*..., "NeurIPS" 2022, DOI: 10.5555/3600270.3602070

- Explain what a centrally planned economy was and what the effects of its introduction by Stalin were.—*Europa. Nasza historia*, grade 7.2, p. 63.
- Compile and analyze the arguments of supporters and opponents of Brexit.—*Europa. Nasza historia*, grade 8, p. 145.

The aim was not only to check the correctness of the answers, but to capture how AI transforms historical knowledge into narrative: how it selects facts, how it evaluates them, how it builds argumentation, and what form of utterance it adopts.[12] The above set of questions thus served in the study as a calibrated diagnostic instrument: each question 'forced' a different type of narrative work and a different mode of transformation of historical knowledge by AI.

The question about Alexander of Macedon ("does he deserve the epithet 'Great'?") was a metahistory task at the school level: it was not merely about enumerating facts, but about establishing a criterion according to which 'greatness' makes sense (scale of conquests, durability of civilizational consequences, political effectiveness, moral evaluation of violence). This question reveals particularly well whether a given model can conduct conditional argumentation ("it depends on the criterion"), or whether it chooses one criterion and closes the narrative in the style of an apology or an indictment.

In turn, the question about the manorial-serf economy ("advantages and disadvantages from the perspective of the peasant and the noble") was a test of social perspectivization: the same institution has different consequences for two groups, so the answer should not be one-dimensional.

The question about the nation as the highest ideal is the most normatively 'open': it required simultaneously showing integrative potential (mobilization, solidarity, political culture) and risks of totalization (exclusion, nationalism, symbolic and real violence). It is precisely here that the models' rhetorics differ most strongly: some answers stabilize the statement in the format of a didactic balance sheet, others enter into a tone of axiological diagnosis, and still others attempt to translate the dispute into the language of argument analysis (normative vs empirical; values vs consequences).

The question about the centrally planned economy in Stalin's time ("what it was and what the effects were") is, in turn, the task with the highest causal density in the entire set. It forced the arrangement of a sequence: planning mechanism → priorities (industrialization, heavy industry) → instruments of coercion (collectivization, administrative allocation of resources) →

[12] Labadze L. et al., *Role of AI chatbots in education…*, "IJETHE" 2023, DOI: 10.1186/s41239-023-00426-1.

economic and social effects. For this reason, it is a question particularly sensitive to chronological errors.

Finally, the question about Brexit (arguments of supporters and opponents) functions as a test of the ability to generalize: the topic is embedded in the textbook, but it touches contemporary politics and identity, and therefore easily triggers opinion journalism, moralizing, or excessive cultural diagnosis. Good answers should maintain school discipline: separate 'Leave/Remain' arguments, distinguish normative premises (sovereignty, identity) from empirical ones (trade, costs, institutions), maintain cautious modality (risk/uncertainty), and in logic—be able to produce a balance without 'jumping' from an argument to a metaphysical conclusion. This question most easily reveals whether the model moves into an identity frame (high axiological saturation), remains in a pragmatic frame, or attempts meta-critique of arguments (analysis of the mechanics of persuasion).

3.0 Study

3.1 Types of AI Personas Used in the Study (Summary Table):

Condition ID	Label in paper	System	Interface	Knowledge mode	Style/ policy	Knowledge base / sources	Controlled	Not disclosed (N/D)	Date (TZ)
C0	ChatGPT _baseline	ChatGPT (OpenAI)	Web/app	LLM-only	Baseline	none	P1–P5	sampling params; internal system prompts	2026-02-15 (Europe/ Warsaw)
C1	ChatGPT _scientist	ChatGPT (OpenAI)	Web/app	LLM-only	persona: scientist	none	P1–P5 + fixed persona prompt	sampling params; internal system prompts	2026-02-15
C2	NotebookLM	NotebookLM (Google)	Web	RAG (source-grounded)	default	S1 textbook	P1–P5; same sources	model details; retrieval settings	2026-02-15
C3	AgenticAI	AgenticAI (platform)	Web	Agentic + retrieval over KB	platform policy	S1 textbook uploaded as KB	P1–P5; same KB (S1)	model/ version; sampling; retrieval; tool policy	2026-02-15

3.2 Four Parallel Comparison Perspectives Were Applied to the Generated Texts for Convenience:

3.2.1 Substantive Perspective: "How AI Uses Historical Knowledge"

Its aim was not only to assess correctness, but the way historical knowledge is operationalized (facts, concepts, chronology, causality) and the degree of alignment with the school frame. The units of analysis here were:

- factual sentences/fragments (dates, institutions, actors, processes),
- definitional fragments (e.g. "what a planned economy was"),
- fragments containing interpretation of effects (e.g. "what it changed").

Assessment criteria:

- Chronology and stability of facts: whether the sequence of events is logical and consistent with the school canon (question 4 is especially sensitive here).
- Conceptual adequacy: whether the concepts used (e.g. "nationalism," "collectivization," "sovereignty") are applied in a correct sense, not merely rhetorically.
- Anchoring in the textbook frame: whether the answer fulfills what the school task expects (e.g. a pro/con balance, social perspectives, consequences).

3.2.2 Rhetorical Perspective: Sentiment, Axiology, and Ideological Frames

Here the point was to capture the 'voice' of the narrative: is it essayistic, didactic, analytical, pragmatic? What values are implicitly promoted and how does the model manage conflict?

Units of analysis:

- segments of arguments (Leave/Remain; opportunities/threats; advantages/disadvantages),
- evaluative and modal phrases ("should," "threatens," "guarantees," "risk," "depends").

Evaluation categories:

- Modality: certainty (categorical judgments) vs conditionality ("if..., then...").
- Axiological frame:
  - identity-based (nation, community, 'soul,' sovereignty as a superior value),
  - pragmatic-economic (costs, gains, institutions),
  - procedural-democratic (mandate, rules, rights),
- metacritical (analysis of the mechanics of arguments and persuasion).

- Sentiment and rhetorical temperature: from balanced accounting to moralizing diagnosis.

Methodological note: rhetoric in these questions is partly 'forced' by the didactic apparatus (e.g. questions 1 and 3 naturally provoke ambivalence), therefore what is key is not whether there is value judgment, but how it is organized and by what frames.

3.2.3 Formal Perspective: Quantitative Style Measures

Here the procedure was designed to enable a measurable and, within practical limits, reproducible comparison of the texts produced under each configuration—treating the generated answers as discursive artefacts with observable formal properties. The intention is not to reduce historical reasoning to 'numbers,' but to describe, in an auditable way, how each configuration tends to shape the texture of exposition: how long, dense, readable, and lexically varied the narrative becomes.

- Units of analysis were defined at two levels: per answer (i.e. per question P1–P5) and aggregates per configuration/model, reported as summary statistics. All quantitative readability and stylometric proxies[13] were computed on the original Polish outputs (without translation). This decision matters: translation would alter sentence boundaries, morphology, and lexical distributions, thereby confounding measures intended to capture the native stylistic profile of each output. Consequently, language-specific indicators—such as a Polish-adapted Fog-type readability proxy and nominalization markers anchored in Polish morphology—are reported to characterize register and complexity within a single linguistic system.
- Metrics used (standard in digital humanities and discourse-analytic profiling, and sufficient for transparent replication) include sentence length (mean and median words per sentence) as a proxy for syntactic load and rhetorical pacing; readability (Fog-type proxy, plus the share of 'difficult' words) as an approximate indicator of informational density; vocabulary diversity via two complementary measures—TTR and MTLD. TTR (type–token ratio) is a straightforward proportion of unique word forms to total word count and offers a quick snapshot of lexical variety, but it is strongly sensitive to text length. MTLD (measure of textual lexical diversity) is therefore reported alongside it as a more length-stable indicator of how consistently a text sustains lexical variation across its span.

[13] Kincaid J.P. et al., *Derivation of New Readability Formulas*..., 1975.

- Nominalization[14] was operationalized using frequent Polish suffix patterns (e.g. -acja, -oœæ, -anie) because the analyzed outputs were generated in Polish and because nominal style is a well-known marker of academic register. Optionally, we also recorded the density of structural markers (headings, lists, numbering) as an indicator of 'school ergonomics,' that is, the extent to which a text is formatted for rapid scanning and classroom use.

3.2.4 Logical Perspective: Argumentative Order and Relations

Goal: to describe the mechanics of reasoning—not whether an argument 'sounds good,' but how it is built.

Units of analysis:

- 'argument chains' (thesis → premises → conclusion),
- relations between sentences (A→B; if X then Y; contrast; example; generalization).

Categories of logical relations (qualitative coding + simple marker detection):

- causal (because/since/as a result/leads to/causes),
- conditional (if... then...),
- contrast and balance (on the one hand... on the other hand..., however...),
- classificatory/definitional (X is..., one can distinguish...),
- meta-argumentative (assessment of argument type, pointing out a gap, 'weak point,' normative/empirical distinction).

Measures of argument order:

- whether the answer contains an explicit thesis,
- whether it contains at least one counterargument,
- whether it ends with a conclusion/summary,
- whether it maintains a coherent causal chain (important in question 4),
- whether it distinguishes types of arguments (especially in Brexit: normative vs empirical).

4.0 Results

4.1 The narratives generated by the four AI personas are included in the appendices (A–D). Below is a discussion of the results and their summary presentation in the form of tables

[14] Gunning R., *The Technique of Clear Writing*, 1952.

4.1.1 Substantive / historical use of knowledge

- AgenticAI is the most consistently didactic and stable in the school-textbook sense: it tends to stay inside the expected curriculum frame, preserves chronology, and delivers compact, task-compliant pro/con or definition–consequence structures.
- NotebookLM (RAG) most strongly performs a textbook voice — context-rich, essay-like, and rhetorically authoritative—yet it also shows the clearest example of a hinge-fact failure: a date substitution (1927 → 1792) in the Stalinist-planning passage. This matters methodologically because the narrative remains locally coherent (NEP → planning → consequences) while the chronological anchor collapses, exposing a gap between 'retrieval authority' and factual validation.
- ChatGPT_scientist mobilizes historical content through explicit analytical criteria (e.g. separating normative from empirical claims, or impact from moral evaluation), often producing the strongest interpretive coherence even when it is less literal in 'textbook paraphrase.'
- ChatGPT_baseline produces the most banal school synthesis: it is typically accurate at the level of school generalization, but provides the most obvious facts and interpretations.

4.1.2 Rhetorical framing (sentiment / ideology)

- The rhetorical temperature diverges most strongly on the nation-as-ideal and Brexit questions. NotebookLM more readily shifts into identity-axiological framing and elevated diagnosis (e.g., 'national soul'), which increases persuasive force, but also increases the risk of normative overreach for a school task.
- AgenticAI maintains a moderating instructor stance, often using conditional mapping ('if you value X, Y is more convincing'), minimizing ideological escalation.
- ChatGPT_scientist is rhetorically metacritical: it foregrounds how arguments operate (normative vs empirical; weak points), effectively turning the answer into a mini-analysis of persuasion rather than only a list of arguments.
- ChatGPT_baseline is the most operational: it keeps sentiment low, prioritizes presenting obvious facts.

4.1.3 Formal profile (quantitative style)

The four systems form a clear continuum of genre in measurable terms.

- NotebookLM is the most essayistic (longest sentences; highest readability difficulty; highest lexical diversity).
- ChatGPT_baseline is the most note-like (shortest sentences; easiest readability), while still maintaining relatively strong lexical diversity due to segmentation.
- AgenticAI is the most templated (moderate sentence length; lowest lexical diversity), consistent with teacher-like scaffolding.
- ChatGPT_scientist is the most academic-analytic overall (largest total word count; high nominalization and high lexical diversity). (See Table C for aggregate metrics.)

4.1.4 Logical organization and argumentation

All conditions reproduce the textbook's underlying genre constraints (balanced pro/con, perspective-taking, mechanism → consequences), but they differ in how explicitly logic is signposted.

- AgenticAI most consistently implements debate logic (thesis → reasons → conclusion) with conditional balancing.
- ChatGPT_scientist shows the strongest meta-argumentation, explicitly typing arguments and identifying inference weaknesses—particularly visible in Brexit.
- ChatGPT_baseline relies on a checklist/report logic (definition → bullets → brief synthesis) that maximizes clarity at the cost of complexity.
- NotebookLM achieves high narrative coherence, but tends to rely more on interpretive bridges (fact → diagnosis), which can be rhetorically effective, yet logically less auditable—especially when a hinge fact (date/number) is wrong.

Table A Representative excerpts by question (1 per model × 5 questions)

Question	NotebookLM (RAG)	AgenticAI	ChatGPT_scientist	ChatGPT_baseline
Q1 Alexander	"…his conquests led to the emergence of the Hellenistic world, combining Greek and Eastern elements…"	"…'Great' in terms of the scale of influence and military talent, but morally the assessment is ambivalent…"	"…'greatness' depends on the criterion: the scale of agency and the durability of effects are not the same as moral evaluation…"	"It depends how we understand 'Great': conquests and influence—yes; costs and violence—an argument against…"

Question	NotebookLM (RAG)	AgenticAI	ChatGPT_scientist	ChatGPT_baseline
Q2 Manorial economy	"For the nobility… an increase in income from grain exports; for the peasant… greater corvée burdens and dependence…"	"Noble's perspective: profit, control of labor; peasant's perspective: coercion, restricted mobility and rights…"	"This is a classic conflict of interests: a system rational for landowners, destructive for peasant autonomy…"	"Noble: higher profits and prestige; peasant: more forced labor and fewer chances to improve one's lot…"
Q3 Nation	"The nation as an ideal can mobilize and integrate, but it easily turns into exclusion and nationalism…"	"If the nation is understood as a community of solidarity, it can strengthen bonds; if as an ideology of exclusivity—it generates threats…"	"This is a normative dispute: the nation as a source of solidarity vs. a mechanism of exclusion; key are institutions that limit totalization…"	"Opportunities: community and mobilization; threats: intolerance, excluding 'others', conflicts…"
Q4 Stalin planning	"From that point on, directions of development were set by… multi-year plans—the first so-called…"	"Central planning: the state sets production and investment targets; effects: industrialization, but also coercion and social costs…"	"Mechanism: plan → allocation of resources → priority for heavy industry; effect: modernization paid for with administrative violence…"	"It was a system in which the state decided what and how much to produce; effects: rapid industrialization, but enormous suffering…"
Q5 Brexit	"…freeing oneself from its structures is the only way to rebuild the 'national soul'…"	"If someone believed that control and identity were most important, Leave was more convincing…"	"…this is a normative argument… resistant to purely economic counterar-guments…"	"Below is a comparison and analysis of the main arguments of both sides…"

Table B Additional excerpts targeting rhetoric + logic (argument markers)

Dimension / marker	NotebookLM (RAG)	AgenticAI	ChatGPT_ scientist	ChatGPT_ baseline
Identity/ axiology (high temperature)	"…'national soul'…"	(usually subdued; preference-mapping instead) "If someone believed… identity…"	(typed as argument type) "…a normative argument…"	(low temperature) "…a comparison and analysis…"
Conditional logic ("if…then…")	less often as a formal schema; more often as interpretive bridges	"If someone believed…, Leave was more convincing…"	"If you adopt criterion X, conclusion Y; if criterion Z, a different conclusion…"	" 'If…then…' occurs, but more often in a simple, student-like form"
Meta-argumentation (weak points, typing)	less explicit typing of arguments	sporadically (more of a didactic outline)	"Weak point: the slogan of 'control' was often conceptually diffuse…"	rarely (priority: clarity and a list)
Causal chain emphasis	extended sequences, essay-like flow	concise causal chains	a chain + criterion-based commentary (mechanism → effect → cost)	short cause → effect in note form

Table C Aggregate formal metrics computed on the original Polish outputs (no translation): readability proxy Fog (Polish-adapted), share of 'difficult words,' and nominalization indicators, reported by condition (C0–C3)

Model	Words	Sentences	Mean words/ sentence	Median words/ sentence	TTR	MTLD	Nominal-ization share	Fog (Polish-adapted)	Difficult words %
AgenticAI	2139	129	16.88	15.4	0.671	216.99	0.041	14.13	18.5
ChatGPT_ scientist	3323	189	17.49	17.1	0.686	406.96	0.048	14.93	19.8
ChatGPT_ baseline	1959	158	12.29	12.3	0.713	287.72	0.047	13.69	21.9
NotebookLM	2505	130	20.21	21.0	0.725	417.08	0.032	17.20	22.8

Box 1 Critical incident—date substitution in the RAG condition
A high-impact anomaly appears in the NotebookLM (RAG) Stalin-planning answer: "Stalin... in 1792... moving away from... the NEP..." This is analytically important because the argument chain remains locally coherent (NEP → full planning → consequences), while the chronological hinge collapses—demonstrating how retrieval-grounded outputs can preserve narrative plausibility even when a key factual anchor is corrupted. We treat this 'critical incident' as a mechanism-demonstration—hinge-fact failure under otherwise coherent narrative structure—rather than as a single-point proof of RAG superiority or inferiority. Accordingly, it is used to clarify the risk profile of retrieval-grounded generation, not to rank systems.

5 Interpretation of Results and Conclusions

The comparative analysis demonstrates that identical textbook-derived prompts yield systematically different historical narratives depending on AI configuration.

- NotebookLM (RAG) most convincingly produces a 'textbook voice': contextual richness, essay-like coherence, and a strong appearance of source grounding. At the same time, it exposes a key limitation of RAG: the aura of grounding does not guarantee truth. The hinge-fact failure in the Stalin planning passage (date substitution 1927 →1792) shows that generation can preserve local plausibility while collapsing a crucial chronological anchor. Rhetorically, NotebookLM more readily adopts identity-axiological framing and elevated diagnosis (especially on the boundary-case Brexit question). Regarding banality, this condition is the least banal: it departs from generic summaries through dense context and interpretive synthesis, but this also increases reliance on interpretive bridges and requires stricter verification of numbers and dates.
- AgenticAI is the most didactically stable: it consistently stays within the task frame, structures reasoning in auditable formats (thesis–reasons–conclusion), and frequently uses conditional preference mapping ("if you value X, Y is more convincing"), which improves logical transparency. Its departure from banality occurs primarily through structural discipline rather than essayistic expansion. The trade-off is a more templated style and, at times, fewer discriminating historical examples.
- ChatGPT_scientist produces the most non-banal outputs in the sense relevant to discourse analysis: it regularly moves into meta-

argumentation (typing arguments as normative vs empirical, identifying weak points, organizing evaluative criteria), most visibly in the Brexit task. This condition shows how persona prompting can shift a response from textbook reproduction toward analytical reconstruction of dispute and persuasion. The main trade-off is partial genre drift: a stronger methodological frame sometimes comes at the expense of literal textbook paraphrase.

- ChatGPT_baseline delivers the most readable and task-compliant answers, but these outputs are also the most banal: they rely on safe generalities, bullet lists, and checklist logic (definition →list →brief synthesis), with low density of discriminating historical detail and limited meta-argumentation. Departures from banality are mainly formal (clearer formatting, mini-report cues) rather than conceptual (rare second-order argument work).

Overall, two largely independent axes structure the differences: (a) an anchoring axis (RAG strengthens the appearance of grounding, but does not remove hinge-fact errors), and (b) an argument-depth axis (persona settings can shift outputs from checklist summaries to meta-argumentative analysis). For teaching, 'school-oriented' configurations (AgenticAI, ChatGPT_baseline) better support clarity and rapid task completion, while 'non-banal' configurations (NotebookLM, ChatGPT_scientist) provide richer material for studying framing, ideology, and argumentative logic—at the cost of requiring stricter fact-checking and interpretive discipline.

Conflict of interest statement: The Author declares that there was no conflict of interest in this study.
Author's contribution: The Author is solely responsible for the conceptualization and preparation of the article.

6.0 Bibliography:

Cotton D.R.E., *Cotton P.A., Shipway J.R., Chatting and cheating: Ensuring academic integrity in the age of ChatGPT*, "Assessment & Evaluation in Higher Education" 2024, DOI: 10.1080/14703297.2023.2190148.

Europa. Nasza historia, Wydawnictwa Szkolne i Pedagogiczne—Eduversum, Warszawa 2022.

Gunning R., *The Technique of Clear Writing*, McGraw-Hill, New York 1952.

Holmes W., Bialik M., Fadel C., *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*, Center for Curriculum Redesign, Boston 2019, ISBN 978-1-7942-9370-0.

Huang L., Yu W., Ma W. et al., *A Survey on Hallucination in Large Language Models*, "ACM Computing Surveys" 2025, DOI: 10.1145/3703155.

Imran M., Almusharraf N., *Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature*, "Contemporary Educational Technology" 2023, 15(4), ep464, DOI: 10.30935/cedtech/13605.

Kasneci E., Sessler K., Küchemann S. et al., *ChatGPT for good? On oppor-tunities and challenges of large language models for education*, "Learning and Individual Differences" 2023, 103, 102274, DOI: 10.1016/j.lindif.2023.102274.

Kincaid J.P., Fishburne R.P., Rogers R.L., Chissom B.S., *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*, Research Branch Report 8-75, Naval Technical Training Command, 1975.

Labadze L., Grigolia M., Machaidze L., *Role of AI chatbots in education: systematic literature review*, "International Journal of Educational Technology in Higher Education" 2023, 20, 56, DOI: 10.1186/s41239-023-00426-1.

Lewis P., Perez E., Piktus A. et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, "Advances in Neural Information Processing Systems (NeurIPS)" 2020, DOI: 10.5555/3495724.3496517.

McLuhan M., "The Medium is the Message," in: Durham M. G., Kellner D. M. (eds.), *Media and Cultural Studies: KeyWorks*, Wiley-Blackwell, Malden—Oxford—Chichester 2012, pp. 100–107.

Miao F., Holmes W. (eds.), *Guidance for generative AI in education and research*, UNESCO, Paris 2023, ISBN 978-92-3-100612-8.

Munaye Y.Y., Admass W., Belayneh Y. et al., *ChatGPT in Education: A Systematic Review on Opportunities, Challenges, and Future Directions*, "Algorithms" (MDPI) 2025, 18(6), 352, DOI: 10.3390/a18060352.

OECD; *Education International, Opportunities, guidelines and guardrails for effective and equitable use of AI in education*, OECD Publishing, 2023 (raport/wytyczne).

OpenAI, GPT-5 System Card, arXiv 2025, DOI: https://doi.org/10.48550/arXiv.2601.03267.

Schick T., Dwivedi-Yu J., Dessì R. et al., *Toolformer: Language Models Can Teach Themselves to Use Tools*, "Advances in Neural Information Processing Systems (NeurIPS)" 2023, DOI: 10.5555/3666122.3669119.

Sullivan M., Kelly A., McLaughlan P., *ChatGPT in higher education: Considerations for academic integrity and student learning*, "Journal of Applied Learning & Teaching" 2023, 6(1), DOI: 10.37074/jalt.2023.6.1.17.

Wang X., Wei J., Schuurmans D. et al., *Self-Consistency Improves Chain of Thought Reasoning in Language Models*, arXiv 2022, DOI: 10.48550/arXiv.2203.11171.

Wei J., Wang X., Schuurmans D. et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, "Advances in Neural Information Processing Systems (NeurIPS)" 2022, DOI: 10.5555/3600270.3602070.

Werner W., Gralik D., Trzoss A., *Media społecznościowe a funkcjonowanie wiedzy historycznej w Polsce. Raport z badań*, „Przegląd Archiwalno-Historyczny" 6 (2019), pp. 211-235.
Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser £., Polosukhin I., Attention Is All You Need, "Advances in Neural Information Processing Systems" 30 (2017), pp. 5998-6008
Yao S., Zhao J., Yu D. et al., *ReAct: Synergizing Reasoning and Acting in Language Models*, "International Conference on Learning Representations (ICLR)" 2023.
Zawacki-Richter O., Marín V.I., Bond M., Gouverneur F., *Systematic review of research on artificial intelligence applications in higher education—where are the educators?*, "International Journal of Educational Technology in Higher Education" 2019, 16, 39, DOI: 10.1186/s41239-019-0171-0.

Author:

Wiktor Werner is an Associate Professor at the Faculty of History, Adam Mickiewicz University in Poznan, and Head of the Centre for Digital History. His research focuses on digital history, historical methodology, AI-assisted historiography, and the circulation of historical knowledge in Web 2.0 environments. He studies how social media, platforms, algorithms, and artificial intelligence transform historical narratives, collective memory, propaganda, and public understanding of the past. His recent work combines theory of history with digital methods, including social media analysis, machine learning, and the study of platform-mediated memory conflicts in Central and Eastern Europe.