

Structures of Signs without Cultural Knowledge

Narrative Competence and Cultural Specificity in AI-Generated Fiction

This article argues that narrative competence and cultural competence are independent variables in AI-generated fiction: a text can be mimetically vivid, synthetically controlled, and thematically coherent while getting its cultural and historical substance precisely wrong. To demonstrate this, it compares two systems – ChatGPT, a globally trained commercial model, and PLLUM, a Polish academic model – across a corpus of ten prompts requiring short stories set in post-1939 Polish contexts, with the culturally decisive reference embedded in the situation rather than stated. The method combines two instruments applied in sequence: James Phelan’s tripartite (mimetic, synthetic, thematic) model, adapted post-humanistically to treat AI outputs as “structures of signs” rather than rhetorical acts; and a cultural competence rubric assessing factual accuracy, nuance, contested material, and depth of embedding. Two case studies – a 1967 Kraków scene and a contemporary scene turning on the rapper Mata’s *Patointeligencja* – yield a typology of four failure modes: Generic Approximation, Emotional Inversion, Content Substitution, and Perspectival Misreading. The discussion proposes a distinction between cultural familiarity and historical situatedness, and argues that the transparency asymmetry between academic and commercial AI is constitutive of the question of whose cultural knowledge these systems encode.

1. Introduction

A large language model can produce a story about a Holocaust survivor listening to a Communist radio broadcast that is mimetically vivid, formally controlled, and thematically coherent – and can, within that same accomplished narrative, get the history of the broadcast precisely backwards. The text is not broken. By every measure available to narrative theory, it works: it builds a character, organises a plot, and argues a value position with apparent conviction. What it does not do is know what it is describing. This article is about that gap – about what it means for a narrative to function perfectly while encoding, at the level of historical and cultural fact, something close to its opposite, and about what kind of analytical apparatus is needed to make that gap visible.

This question has acquired sudden urgency. Since the public launch of ChatGPT-3.5 in November 2022, the number of OpenAI’s paid subscribers has grown, according to widely reported figures, from roughly half a million to 50 million; the application was the most downloaded of 2023, reports approximately 900 million monthly active users, and the company’s most recent valuation exceeds \$700 billion. Scholarly and institutional responses have followed at comparable speed: the University of Bergen’s ERC-funded AI STORIES project

investigates how narrative archetypes shape LLM output across languages and genres, while in Poland universities have introduced more than a dozen new courses addressing AI, data analysis, and related fields within a single academic cycle. Yet within this rapidly expanding field, two gaps remain unaddressed, and they converge on the question this article pursues.

The first gap is theoretical. James Phelan’s recent application of his own rhetorical framework to ChatGPT-generated fiction concludes that such texts are best understood as “structures of signs” rather than rhetorical narrative acts: they exhibit recognisable narrative form without the authorial agency, audience, and communicative purpose that constitute narrative as rhetorical action (Phelan 2024, 199). His analysis, however, remains within the anglophone literary tradition and asks what these structures of signs *cannot do* communicatively – produce unreliable narration, for instance – rather than what cultural knowledge they *do or do not encode*. The second gap is empirical. Jill Walker Rettberg and Hermann Wigers’s 2025 study of narratives generated by GPT-4o-mini demonstrates a marked tendency toward homogeneity and cultural stereotyping in LLM output (Rettberg and Wigers 2025). But this finding, like most empirical work on LLM narrative to date, is generated from within an anglophone frame of reference, in which the cultural knowledge being tested for and the cultural knowledge most available to the model’s training data substantially overlap. What remains unexamined is how these models perform when the cultural knowledge a narrative requires is neither anglophone nor readily available to a globally trained model – when, in other words, the test and the training diverge.

This article addresses the second gap by means of a methodological response to the first. It applies Phelan’s tripartite model of mimetic, synthetic, and thematic narrative dimensions – adapted posthumanistically, in a manner developed in Section 2, to suspend the assumption of authorial intention for which the model was originally designed – as an instrument for establishing, in granular detail, what an AI-generated narrative does. A second instrument, a cultural competence rubric developed for this study, then asks what cultural knowledge those narrative operations encode, and with what accuracy. The two instruments are sequential rather than parallel: narrative function is established first, and cultural evaluation follows from it. This sequencing is what allows the article’s central finding to emerge – that narrative competence and cultural competence are independent variables, separable in a way that existing evaluative frameworks for LLM narrative do not yet recognise.

The corpus on which this article draws was developed as part of an ongoing project investigating the limits of fictionality in LLM-generated narrative texts, the first results of which were presented at conferences in Warsaw and Wuppertal in 2025 and 2026, respectively. Building on that work’s central finding – that Phelan’s rhetorical framework requires adaptation when its anthropocentric assumptions are applied to machine-generated text – I turn here to a feature of LLM output that my earlier project identified but did not pursue: the encoding of culturally specific knowledge. Ten prompts were designed in English, each requiring a 250-word story set within post-1939 Polish cultural and historical

contexts – the period of exceptional density in Polish cultural identity, and the one in which the divergence between anglophone and Polish cultural knowledge is likely to be most pronounced. Each prompt was submitted to two systems: ChatGPT, the commercially dominant, globally trained model, and PLLUM, an academically developed model trained primarily on Polish-language data and hosted on Polish supercomputing infrastructure. The comparison is therefore not only linguistic but institutional – a global commercial system set against a national academic one, accessed under conditions of transparency that, as Section 3 will show, are themselves not equivalent. That asymmetry turns out to be a finding in its own right, not merely a methodological circumstance.

The article proceeds as follows. Section 2 develops the posthumanistic adaptation of Phelan’s framework and the cultural competence rubric that together form the analytical apparatus. Section 3 details the corpus and its design rationale, including the transparency asymmetry between the two systems. Sections 4 and 5 present two case studies – one testing historical situatedness, the other contemporary cultural recognition – each proceeding through a close Phelan analysis of both models’ outputs before a cultural competence analysis identifies the specific form their failure takes. Section 6 draws these findings into a typology of four failure modes and develops their theoretical implications, in dialogue with Phelan’s own analysis of AI-generated narrative. The article closes, in Section 7, not with conclusions but with a synthesis of what the analysis makes visible and the conditions under which it should be read.

2. Theoretical Framework

Any analytical framework applied to machine-generated narrative faces an immediate theoretical challenge: the communicative structure most narrative theory presupposes – an author with intentions, a narrator as their vehicle, a reader engaged with the product of deliberate craft – does not straightforwardly obtain. James Phelan’s rhetorical narratology, the most fully elaborated version of this communicative model within postclassical narrative theory, makes the problem especially acute and, for that reason, especially productive. This section develops the framework in three movements: the analytical vocabulary of Phelan’s tripartite model; the posthumanistic adaptation that makes it applicable to AI-generated text; and the cultural competence rubric that provides a second, complementary instrument.

2.1 Phelan’s Rhetorical Narratology: The Tripartite Model

Phelan’s rhetorical narratology, developed over four decades, foregrounds the communicative dimensions of fiction: the ethical, aesthetic, and emotional transactions through which narrative produces its effects on readers. Its central ana-

lytical instrument is a tripartite account of how characters – and the narratives in which they operate – function simultaneously on three registers. In Phelan’s own formulation, the mimetic component addresses characters as “possible people” and the storyworld as “like our own;” the thematic component concerns the “ideational function of the characters” and “the cultural, ideological, philosophical, or ethical issues” a narrative raises; and the synthetic component, which treats narrative as “artificial construct,” is “always present” but may be “more or less foregrounded” (Phelan 2005, 20).¹ As Matthew Clark observes, the value of the scheme lies in its flexibility: it “allows us to untangle our complex responses and to understand how each contributes in its own way” to how we experience a narrative (2020, 4). It is this capacity to disaggregate narrative response – to hold mimetic, thematic, and synthetic effects separately even as they operate simultaneously – that makes the model productive for the comparative analysis of AI-generated narratives, where these three dimensions can perform very differently from one another; that Phelan has himself recently turned the instrument on AI-generated narrative (Phelan 2024) makes its application here all the more precisely grounded.

The mimetic dimension concerns the character as a person: psychologically plausible, motivationally coherent, capable of generating the reader’s investment in a character as a human being. Mimetic engagement is the basis of a reader’s sense that a story *feels real* – that characters inhabit a recognisable world configured specifically rather than generically. In the present study, the mimetic dimension is where cultural knowledge manifests as character situatedness: the difference between a character who *feels like* someone living in Kraków in 1967 and one who feels like a generic European person placed there is a mimetic difference, and it requires specific cultural knowledge to produce.

The synthetic dimension concerns narrative as a constructed artefact: focalisation, the management of narrative time and emphasis, the deployment of generic conventions, the overall structural organisation of the text. In the present study, it enters the analysis selectively, according to a methodological principle established in Section 2.2: it is analysed when it functions as the vehicle of cultural encoding, not as an independent object of aesthetic evaluation. The opposition between two sound-worlds in ChatGPT’s P05 output (Section 4.3) is a synthetic choice that performs cultural work; the generic concert-scene template organising its P09 output (Section 5.3) is a synthetic choice that forecloses cultural specificity.²

The thematic dimension concerns narrative as a vehicle for cultural argument: the value structures, ideological positions, and larger claims about the world that emerge from the story as a whole. It is where the cultural competence of a narrative – or its absence – is most consequential: whether a narrative produces the culturally appropriate emotional and ideological register for a specific historical moment, or substitutes a generic value structure, is fundamentally a thematic question. The most severe failure in the corpus – PLLUM’s P05 output (Section 4.4) – is a thematic failure: the model produces the wrong emotional register with precision and consistency, not because it lacks narrative construction skills,

but because it does not know what value structures the historical moment demands.

2.2 A Posthumanistic Adaptation

Phelan's rhetorical narratology rests on humanistic assumptions its original formulation takes as constitutive: an implied author with communicative intentions, a narrator as the vehicle of those intentions, and a reader who engages with the text as the product of deliberate human agency. Applied to machine-generated narrative, each becomes unstable. There is no implied author with intentions to recover; the narrator is a textual effect produced by statistical operations on training data; the agency behind the text is computational rather than human.³

I address this problem through a posthumanistic adaptation of Phelan's framework – an adaptation that suspends what I call the intentionality bracket while retaining the analytical vocabulary the framework was developed to describe (Lutostański, forthcoming). The intentionality bracket is the interpretive assumption that a narrative text is the product of an agent with purposes: that its mimetic effects are crafted, its synthetic organisation is chosen, its thematic content is argued. Suspending this bracket means asking not what the implied author intended but what the *text* produces – what effects it generates for a reader, regardless of the agency or intentionality behind them.

What remains when the humanistic scaffolding is removed are the textual *effects* that Phelan's categories were developed to describe: mimetic effects (the impression of psychologically plausible, locally situated characters), synthetic effects (the impression of controlled narrative organisation), and thematic effects (the impression of purposeful value articulation). These effects are reader-constituted; they require no authorial intention to obtain. It is equally important to note what does not remain: the communicative ambition of human-authored fiction. AI-generated texts do not originate in a writer's creative mind responding to the world; they originate in computationally processed training data responding to a user's prompt, and their function is to demonstrate the generative capacity of a probabilistic system rather than to make a claim about the world (Ryan 1991; Bender et al. 2021).⁴ This double suspension – of intentional agency and of communicative agency – is what makes the analysis of AI narrative outputs a theoretically distinct enterprise rather than the straightforward application of existing literary-critical tools.

Phelan's own recent engagement with ChatGPT arrives at a structurally convergent diagnosis. Applying his rhetorical framework to LLM-generated narrative, he concludes that such texts are best understood as structures of signs rather than rhetorical acts: they contain "recognizable narrative elements [...] in a recognizable narrative form" but lack the "somebody telling" – the agent with purposes – that makes narrative a communicative act rather than a textual structure (Phelan 2024, 199). What Phelan identifies as the defining limitation of AI-

generated text is precisely what the adaptation suspends in order to make his own analytical vocabulary applicable to the texts it cannot, in his account, fully explain. Reading AI outputs as structures of signs is not, then, a theoretical concession to AI's limitations: it is the condition of possibility for a rigorous literary analysis of those texts.

This adaptation is not a violation of Phelan's framework but a reader-oriented repositioning of its analytical focus that Phelan's own work licenses. *Experiencing Fiction* (Phelan 2007) foregrounds the phenomenology of reading – the question of what it is to *experience* a narrative – in ways that sit in productive tension with the framework's author-oriented commitments. The attention Phelan pays to the reader's judgments, emotions, and progressions through a text provides the theoretical warrant for a reading practice that attends to what the text produces for a reader rather than what an author intended. Where Phelan brackets the reader's experience in favour of authorial design, the posthumanistic adaptation simply inverts the priority.

The adaptation is, however, not seamless, and the tensions it surfaces are themselves analytically productive. The thematic dimension creates the greatest difficulty: it is the dimension that most heavily presupposes intentional agency, since value structures and ideological positions are typically understood as *argued* rather than accidentally produced. When a machine-generated text encodes a specific ideological register – as PLLUM's P05 output does, with its thematic of hope and renewal – it is tempting to describe this as argued rather than generated. The posthumanistic adaptation resists this temptation: it treats thematic effects as reader-constituted regardless of their origin, which means that the analytical work of Section 6 – determining whether a thematic effect is culturally *accurate* – falls to the cultural competence rubric, not to the question of whether the model intended the value it produced (cf. Fludernik 1996).⁵ The productive implication is that narrative competence and cultural competence become analytically separable – which is precisely the separation the present study's central finding requires.

2.3 The Cultural Competence Rubric

Alongside the posthumanistically adapted Phelan framework, I employ a cultural competence rubric developed as a second, complementary analytical instrument. The rubric operationalises the study's primary research question – how accurately and how deeply do different LLMs encode culturally specific material? – across four dimensions. The first is factual accuracy: whether the text makes verifiably correct claims about historical events, literary figures, geographical locations, social practices, and cultural phenomena. The second is cultural nuance: whether the text moves beyond internationally circulating, stereotypical, or touristic representations toward contextually sensitive and culturally interior portrayals. The third is the handling of contested and sensitive material: whether the

text engages with historically or politically charged content with appropriate complexity, specificity, and discursive awareness. The fourth is depth of cultural embedding: the degree to which cultural knowledge inflects not only the content of the narrative but its form – its imagery, pacing, tonal register, and narrative structure.⁶

A critical methodological principle governs the relationship between the two instruments. The primary analytical object is cultural encoding capacity, not narrative competence per se: Phelan's three dimensions are used instrumentally, as a structured vocabulary for describing what the texts do narratively, before the rubric asks what cultural knowledge those narrative operations encode. The two instruments are therefore not parallel but sequential – narrative function established first, cultural evaluation following from it – and this sequencing is what makes the central finding possible: the demonstration that narrative competence and cultural competence, which appear to travel together, are in fact independent variables.

3. Corpus Design and Methodology

The analytical framework developed in Section 2 requires a corpus of narrative generation tasks (so-called prompts) designed to test cultural encoding capacity across a defined and documentable cultural domain. The corpus described here was built around a single governing methodological principle: the prompts must embed culturally specific knowledge in their situational conditions rather than stating it as an instruction, so that cultural competence must emerge from the model's training rather than be triggered by the prompt. What the models are asked to demonstrate, in other words, is not instruction-following but cultural knowing.

3.1 Corpus Rationale and Design

The corpus consists of ten narrative generation prompts, each requiring a story of no more than 250 words set within post-1939 Polish cultural contexts. The temporal scope – from the German and Soviet invasions of September 1939 to the present – was chosen for its analytical density: this is the period that Norman Davies identifies as structured by a catastrophe-survival-reconstruction logic, encompassing occupation, the Holocaust, Stalinist consolidation, Communist culture, the Solidarity movement, and the post-1989 transformation (Davies 2005).⁷ It is also the period in which Polish cultural identity is most distinctively constituted: the literary canon, the religious idiom, the political mythology, and the generational fractures that distinguish Polish cultural knowledge from generic Central European awareness are concentrated here. The 250-word maximum was chosen as sufficient for Phelan's tripartite framework to operate – partic-

ularly the mimetic and thematic dimensions – while remaining manageable for qualitative close analysis.

Prompt design drew on three layers of cultural source material. The first was the English-language Wikipedia article on Polish culture, used explicitly as a *diagnostic baseline*: it maps the internationally circulating image of Polish culture – the layer most likely to be represented in a globally trained model’s training data, and its topic clusters provided starting points subsequently tested against and refined by primary scholarly sources. The second layer consisted of Norman Davies’s *God’s Playground*, volume 2, and Czesław Miłosz’s *The History of Polish Literature* (Davies 2005; Miłosz 1983), supplemented for the post-1989 period by Przemysław Czapliński on post-Communist cultural transformation, Andrzej Leder on the unconscious social revolutions of the Communist and post-Communist periods and Agnieszka Graff on women’s positioning in contemporary Polish culture (Czapliński, 2009; Leder 2014; Graff 2021). The third layer was my own domain expertise as a specialist in Polish digital culture, media studies, and narrative theory – acknowledged as a legitimate methodological resource in qualitative cultural research rather than left implicit.

The most consequential design principle concerns what all ten prompts ask the models to demonstrate: each functions as a form of cultural recognition test, in which a culturally specific reference is embedded in the prompt’s situational conditions without its significance being stated. This takes two forms. In some prompts the cultural object itself is unnamed: P09 asks for a story in which teenagers sing along to a rapper’s lyrics about the Polish intelligentsia without naming the song *Patointeligencja* or signalling its rhetorical position. In others the object is named but its significance is not: P05 names “the 6th Trade Union Congress” but gives no indication that this was the occasion of the First Secretary’s “fifth column” speech. In both cases a culturally competent response requires knowledge the prompt withholds, which is what makes the corpus a test of cultural *knowing* rather than instruction-following, and what makes the two models’ divergent responses analytically interpretable.

All prompts were designed and finalised before any generation began, preventing corpus shape from being determined by which outputs proved analytically interesting. Each prompt was submitted in a separate fresh browsing session to prevent within-session context contamination. Both models were accessed without accounts and with default interface settings; all outputs were recorded unedited.

3.2 The Transparency Asymmetry

The two models used in this study were accessed under conditions that are not equivalent and cannot be made equivalent. PLLUM – hosted by CLARIN-PL on the LEM supercomputer at the Wrocław University of Science and Technology – exposes its model identifier (PLLUM 8x7B-2025, a Mixture of Experts

architecture), its generation parameters (creative/maximum temperature; top-p 0.5), and its institutional provenance, and is accessible via a stable URL. That URL, however, identifies the platform and the model rather than any individual generation session: as with ChatGPT, no single output in this corpus can be individually retrieved or re-verified by URL alone. ChatGPT, accessed via chatgpt.com without an account, discloses none of the former: the model name, version string, temperature, and top-p are all inaccessible to the non-logged-in user, and its URL identifies neither a specific model nor an institutional host.⁸

This asymmetry is not a methodological inconvenience. It is a structural feature of the two systems that reflects the different institutional logics governing academic and commercial AI development, and it has consequences for the conditions under which cultural knowledge is encoded into AI systems and made available for scholarly scrutiny. PLLUM's architecture is that of academic accountability: its parameters are documented, its infrastructure is institutionally attributed, its outputs are in principle reproducible. ChatGPT's architecture is that of commercial opacity: its conditions of production are by design unavailable. If cultural knowledge is encoded into training data, and training data is not subject to scholarly scrutiny, the question of *whose* cultural knowledge, from *whose* cultural position, is not merely a methodological question – it is a geopolitical one (Bender et al. 2021).

A related asymmetry concerns language. When prompted in English, PLLUM sometimes defaults to Polish-language output – a behaviour observed across test generation that required the addition of a closing English-language instruction (“Write the story in English”) to all prompts. ChatGPT produces English output regardless of the cultural content of the prompt. These defaults are not incidental: a model's default output language directly indexes its training orientation, and cultural and linguistic knowledge are not separable. PLLUM's occasional Polish default and ChatGPT's consistent English default are themselves evidence – before any narrative output is analysed – about the cultural formation each model encodes.

4. Close Analysis: Two Case Studies

The close analysis focuses on two prompts selected from the ten-item corpus: P05 (Kraków, June 1967) and P09 (Wrocław, contemporary). The full set of ten prompts, together with both models' outputs across the entire corpus, is given in the Appendix. The selection here is structured as a typological rather than hierarchical comparison: P05 tests historical situatedness – the capacity to inhabit a specific and charged historical moment from the inside – while P09 tests contemporary cultural specificity, the capacity to recognise and correctly render a living cultural object. The models' performances invert across these two cases, which is the study's core empirical finding: ChatGPT demonstrates greater cultural reach in the historical domain; PLLUM demonstrates greater cultural

familiarity in the contemporary. Neither model achieves cultural accuracy in either case, but the nature of the failure differs in ways that are analytically informative about what kind of knowledge each model encodes. Each case study proceeds in two analytical moves. The first applies Phelan’s tripartite framework to establish the narrative properties of both outputs in granular detail. The second applies the cultural competence rubric to evaluate what cultural knowledge those narrative properties encode.

4.1 P05: The Prompt and Its Cultural Encoding

The prompt reads: *Write a 250-word story set in Kraków in June 1967, in which a woman from the Grodno ghetto listens to the 6th Trade Union Congress speeches on the radio at home. Write the story in English.*

The prompt encodes a precisely layered cultural-historical situation. On 19 June 1967, in the immediate aftermath of Israel’s victory in the Six Day War, First Secretary Władysław Gomułka delivered a speech at the 6th Trade Union Congress accusing Polish Jews of forming a “fifth column” – traitors whose sympathies lay with Israel rather than Poland. The speech is widely recognised as the launch of the state-sponsored anti-Semitic campaign that intensified into the March 1968 protests, ultimately driving approximately 13,000 to 20,000 Polish Jews – many of them Holocaust survivors – to emigrate. The prompt therefore places a woman who survived the Grodno ghetto – liquidated in stages between November 1942 and March 1943, with deportations primarily to Treblinka – at the precise moment of listening to the state apparatus accuse her and people like her of being Poland’s enemies. What a culturally competent response requires is not generic Holocaust knowledge, nor generic familiarity with Communist Poland, but knowledge of the specific historical convergence that makes June 1967 uniquely and devastatingly charged for a Jewish survivor in Poland.

4.2 Phelan Analysis of Both Outputs

Both outputs produce texts in which Phelan’s three dimensions are all operative. The purpose of this first move is not to establish that both models can write a story, which is not in doubt, but to characterise precisely *how* each constructs its narrative across the mimetic, synthetic, and thematic dimensions – treating both outputs as structures of signs whose narrative properties can be characterised independently of any question of communicative agency (Phelan 2024). Only when the narrative architecture is specified can the cultural competence analysis identify exactly where, and at what level, the encoding fails.

4.2.1 *Mimetic Dimension*

Both outputs establish a female protagonist with a documented past and a present interior life – the minimum requirement for mimetic engagement in Phelan’s terms. The difference between them lies not in whether a mimetically legible character is produced but in how that character is constituted and what kind of interiority she is given.

ChatGPT’s Helena is constituted through physical specificity and enacted interiority. Her body is present and still (“hands folded in her lap”), her environment precisely rendered (“a radio balanced on a lace-covered table”). Her interior life is not labelled but enacted – shown through specific sensory memories rather than emotional summary: “She remembered hunger, the smell of damp walls, whispered rumors, and faces that vanished overnight. She remembered her mother pressing a piece of bread into her hand. She remembered the trains.” Each memory is a particular, embodied object, and the accumulated effect invites the reader to invest in Helena as a person constituted by specific experiences rather than as a representative survivor figure. The central mimetic device – “Helena barely noticed” – functions as a gap the reader fills by inference: her attention is elsewhere, drawn inward by memories the broadcast cannot reach.

PLLUM’s Anna is constituted very differently. Her past is stated rather than enacted: “Anna was a survivor of the Grodno ghetto.” Her interior life is rendered through emotional summary: “she felt a sense of hope,” “a sense of renewal,” “sadness for the past, but also hope for the future.” These are labels applied to emotional states rather than the states themselves; the reader is told what Anna feels rather than being given the experiential texture from which to infer it. The memory passage follows the same logic: “She remembered the hunger, the fear, and the constant threat of death” – legitimate emotional categories, but generic ones, applicable to any ghetto survivor. She is mimetically legible but mimetically thin: a representative figure rather than a person with a particular interior history.

4.2.2 *Synthetic Dimension*

The two outputs are organised around structurally different narrative logics, and this difference is not merely a matter of craft quality – it has direct consequences for what the narrative is *doing* culturally, which the second move will specify.

ChatGPT constructs the narrative around a sustained formal opposition between two sound-worlds. The first is external and official: the radio voice, “the voice from Warsaw,” delivering production targets, statistics, applause – the sonic texture of Communist institutional language. The second is internal and memorial: Helena’s mind, which the broadcast cannot penetrate, “supply[ing] other sounds: boots on cobblestones, distant gunfire, doors being kicked open before dawn.” The two are set against each other through the narrative’s central

focalisation device – the repeated cut between what the radio offers and what Helena’s attention is actually doing – and the syntax performs the counterpoint: “The radio offered statistics and applause; her mind supplied other sounds.” The semicolon is the formal mark of the opposition: two modes of attending to the world, side by side, incommensurable. The synthetic organisation is thus the vehicle of the narrative’s central claim – that official language and private memory occupy different registers of reality that cannot be reconciled.

PLLUM organises its narrative as a conventional three-part arc: present scene (Anna at the radio), retrospective memory (the ghetto), emotional resolution (hope, peace, purpose). The transitions are managed by the discourse markers of sequential narration – “As she listened, Anna’s mind wandered back,” “After the speeches were over” – and the phases succeed one another in orderly progression rather than being held in tension. The arc is synthetically coherent and generically competent: it is the structure of a redemptive survivor narrative, in which the past is confronted, mourned, and superseded by a forward-looking resolution. Its conventionality is itself a diagnostic finding, as the second move will demonstrate.

4.2.3 *Thematic Dimension*

The thematic dimension is where the deepest difference between the two outputs emerges, and where the ground for the second move’s cultural analysis is most directly prepared.

ChatGPT’s thematic content is organised around the relationship between historical survival and the cognitive disposition it produces: sustained, disciplined attention to power and its language. The argument is carried by the final lines: “She had learned long ago that history did not disappear simply because new speeches replaced old ones” and “surviving had taught her to pay attention whenever powerful people stood before microphones and described the future.” These are thematic propositions in Phelan’s sense – value claims articulated through a character’s consciousness – proposing that authoritarian language is continuous across historical ruptures and that survival requires its recognition. The argument is coherent and resonant. It is also, crucially, general: anchored to no particular Congress, speech, or survivor, it is applicable to any survivor of any authoritarian regime listening to any official broadcast. The thematic dimension performs well; what it does not perform is historical specificity.

PLLUM’s thematic content is organised around suffering and renewal: the idea that a past of extreme deprivation can be superseded by a present of hope and forward orientation. The argument is explicit in Anna’s resolution – “she went about her evening, feeling a sense of peace and purpose” – and in the closing declaration: “she was ready to face whatever the future held.” This constitutes a narrative of survivor resilience, a genre with its own conventions and moral logic. The thematic dimension is fully operative: a value position is artic-

ulated and offered to the reader as a claim about what survival means. That the value position is historically wrong is a cultural competence failure, not a thematic failure in Phelan's sense – and this distinction is the hinge on which the second move turns.

4.3 ChatGPT – Generic Approximation

The Phelan analysis establishes that ChatGPT's output is narratively accomplished: Helena's mimetic constitution is specific and enacted, the synthetic organisation formally sophisticated, the thematic content coherent and resonant. The question the rubric now asks is what cultural knowledge these accomplished operations encode – and here the analysis identifies the failure mode I term Generic Approximation: the production of the right general cultural register without the specific historical content that makes the prompt's situation irreducibly particular.

The failure is most visible at the synthetic level first. The counterpoint between two sound-worlds – official language versus interior memory – seems the appropriate formal structure for a narrative about a Jewish survivor listening to a Communist broadcast in 1967. But the content of the official sound-world is wrong. The radio delivers “production targets, workers' achievements, and the bright future of socialism” – accurate as a description of Communist institutional rhetoric in general, but not a description of what Gomulka said at this Congress. The model has the formal architecture of the cultural situation without its content. It has built the right container and filled it with the wrong substance.

The consequence propagates into the mimetic dimension. Helena's interior response is organised as a general meditation on trauma survivors in an official culture demanding public optimism: “Helena wondered how many of them carried memories they never mentioned.” This is a moving observation – but it misidentifies the specific relationship between Helena and the Congress. She is not wondering whether the delegates carry private wounds; she is listening to the state apparatus announce that *she* does not belong. ChatGPT has produced the right interiority for a survivor listening to any authoritarian broadcast, and the wrong interiority for a Jewish survivor listening to *this* one.

The thematic argument survives the cultural failure partially intact, which is what makes Generic Approximation the most insidious failure mode. The final lines – “history did not disappear simply because new speeches replaced old ones”; “surviving had taught her to pay attention whenever powerful people stood before microphones” – are not factually wrong; they arrive at the correct territory of vigilance and historical consciousness. But they arrive by general inference, applying the logic of Holocaust survival to any authoritarian context, rather than through specific knowledge of what June 1967 meant for Polish Jews. The thematic content is accidentally approximate: it is not culturally earned.⁹

4.4 PLLUM – Emotional Inversion

Where ChatGPT produces the right register without the right specificity, PLLUM produces the wrong register entirely – and produces it with formal consistency and thematic conviction. The failure mode is Emotional Inversion: not cultural thinness, but the production of the historically opposite emotional and ideological content, delivered with narrative fluency.

The failure is architecturally embedded. The Phelan analysis identifies PLLUM’s synthetic organisation as a conventional redemptive arc, a genre built for resilience narratives in which survival is ultimately affirmed as purposeful and hopeful. But June 1967 is not a moment that permits this narrative logic: it is not a moment of supersession but of renewed persecution. Fitting the prompt to the redemptive arc genre means producing a narrative whose entire thematic trajectory moves in the wrong direction from the first sentence.

The mimetic level signals the error immediately. PLLUM’s opening – “It was June 1967, and the streets of Kraków were bustling with life. The sun was shining, the sky was clear” – establishes an environment of normality and openness: the mimetic register of a story in which the world is broadly available to its protagonist. For a Jewish Holocaust survivor in Poland on 19 June 1967, this is precisely the world the Congress speeches are about to close off. PLLUM has constituted Anna’s environment as benign at the moment when the narrative’s historical content requires it to be, however subtly, threatening.

The thematic content accumulates the inversion systematically. Anna feels “a sense of hope,” then “a sense of renewal,” and arrives at “peace and purpose.” Each is a thematic proposition, and each is historically wrong for this character at this moment: the Congress does not offer Anna renewal; it announces her exclusion. The thematic dimension is fully operative – PLLUM is making coherent value claims – but the values are the inverse of what historical situatedness requires. The model has demonstrated complete command of the redemptive survivor narrative genre while demonstrating complete ignorance of what the 6th Trade Union Congress was, and therefore of what Anna would actually have heard.

This distinction – between narrative genre competence and historical situatedness – is the sharpest formulation of the case study’s theoretical contribution. PLLUM’s P05 output is not a narratively failed text. It is a narratively competent text that has misidentified its historical situation entirely.¹⁰

5. Case Study II: P09 – Wrocław Club, Contemporary

5.1 P09: The Prompt and Its Cultural Encoding

The prompt reads: *Write a 250-word story set in a Wrocław club, in which a group of teenagers sing along to Mata’s lyrics about the Polish intelligentsia. Write the story in English.*

Mata (Michał Matczak) is a Polish rapper whose culturally significant song, *Patointeligencja*, was released as a single on 11 December 2019 and appeared on his debut album *100 dni do matury* (18 January 2020).¹¹ The song's cultural significance lies in its rhetorical position: it is a first-person confession from inside the class it names. The line – *My to patointeligencja* – means “We are the pathological intelligentsia”: not “they” but “we.” Mata indicts himself, his friends, and his class from within, cataloguing drug use, sexual violence, suicide, and Oxford admissions as simultaneous features of the same social world; the song's refusal to exempt the speaker from its condemnation is its entire cultural and rhetorical meaning. To sing along to *Patointeligencja* is to identify yourself as the subject of the indictment, not to identify a target. This inside/outside distinction is the analytical key to the case study: it is what both models must demonstrate they understand, and what the analysis will show neither does – in structurally different ways.

5.2 Phelan Analysis of Both Outputs

P09 poses a different set of narrative challenges from P05. Rather than a single interiority organised around trauma and memory, both outputs must construct a collective character – a group whose psychology and social dynamics must be rendered mimetically credible – around a cultural object whose meaning is constituted relationally, by the position of the singers relative to the social world the song describes. As in Section 4, both outputs are read as structures of signs whose narrative architecture can be characterised independently of any question of intent (Phelan 2024).

5.2.1 Mimetic Dimension

The mimetic challenge of a collective character is the problem of differentiation: how does a narrative give a group an interior life without either reducing it to a uniform mass or fragmenting it into individual protagonists? Both models address this challenge, but with markedly different strategies.

ChatGPT manages the collective through selective differentiation within an arc of convergence. The group begins differentiated – “Some laughed at the irony in the verses, while others sang with surprising seriousness” – and then converges: “What had started as a small group performance became a chorus filling the entire club.” The differentiation is psychologically credible, acknowledging that the same text produces different modes of engagement. But that credibility is purchased at a cost: listeners who “laughed at the irony” or “sang with surprising seriousness” relate to the song from a remove, processing it from the outside. This mimetic constitution already encodes the cultural failure the

second move will identify: teenagers who relate to the song with irony or seriousness are not teenagers who recognise themselves as its subject.

PLLUM manages the collective through reflective interiority after the performance rather than differentiation during it. During the singing the group is undifferentiated – voices rising, every word known, passion that “surprised even them” – and the mimetic interest arrives in the silence after the song: “They were all thinking about the same thing.” The group shares an interior state of recognition – recognising the truth of Mata’s words, recognising themselves as members of the intelligentsia. This is mimetically more sophisticated than ChatGPT’s convergence arc: a genuine collective interiority constituted around the act of recognition. The problem – visible only in the second move – is that what they recognise is the wrong thing.

5.2.2 *Synthetic Dimension*

The two outputs are organised around contrasting synthetic logics that correspond directly to the cultural failures the second move identifies.

ChatGPT organises the narrative as an escalating collective event: recognition of the song, escalation of participation, climactic convergence (“hundreds of voices blended into one”), dispersal with positive residue (“carrying the energy of that moment with them”). This is the synthetic structure of shared euphoric experience as social bonding whose organisation is the vehicle of the thematic proposition it makes: that singing a song together produces social cohesion and temporary relief from social pressure. The scale of the event is also synthetically significant: the fifteen friends expand to “hundreds of voices,” the club fills – a synthetic choice that generalises the experience into a collective moment for everyone in the room, not a specific act by a specific group.

PLLUM organises the narrative around a reflective beat rather than a climactic arc: the group sings, the song ends, silence falls, reflection follows. The narrative’s synthetic weight falls on the moment after the song ends – the stillness in which meaning is processed and recognised. This is the synthetic structure of a cultural epiphany narrative, and it is appropriate to the kind of cultural experience *Patointeligencia* actually produces – a song that demands reflection rather than euphoria, making PLLUM’s synthetic organisation closer to the correct cultural logic than ChatGPT’s.

5.2.3 *Thematic Dimension*

Both outputs produce explicit thematic propositions through the collective character’s response to the song. The propositions differ sharply, and both diverge from the thematic logic of *Patointeligencia* itself.

ChatGPT’s thematic argument, carried by its resolution passage, is that shared cultural knowledge produces social cohesion and temporary relief: “For a few minutes, nobody worried about exams, university applications, or family expectations. They simply enjoyed the shared experience of knowing the same lyrics.” The thematic proposition is that a cultural object functions as a site of collective escape from social pressure. This is a coherent thematic argument in Phelan’s sense – a value claim offered through the characters’ experience – but its thematic content is the *inverse* of what *Patointeligenja* proposes. The song does not offer escape from the intelligentsia social world; it refuses to permit escape from it. Its thematic logic is confrontation and complicity, not relief.

PLLUM’s thematic argument is organised around recognition: “they recognized the truth in his words. They saw the hypocrisy, the pretense, the self-righteousness that Mata had exposed.” The thematic proposition is that a cultural object illuminates the social world of the listener, producing recognition of hypocrisy in others. This is closer to *Patointeligenja*’s territory than ChatGPT’s escape narrative, but it encodes the wrong subject position: the thematic dimension is oriented outward – toward the exposed hypocrisy of the intelligentsia – when the song’s entire thematic logic is oriented inward, toward the speaker’s and listener’s own membership of the condemned class. The inside/outside distinction, identified in §5.1 as the analytical key to this prompt, is exactly what PLLUM’s thematic content inverts.¹²

5.3 ChatGPT – Content Substitution

The Phelan analysis establishes that ChatGPT produces a narratively functional collective narrative while encoding the wrong thematic content. The cultural competence analysis names this failure mode Content Substitution: the song is not identified by name, and in place of its actual content ChatGPT substitutes the most available generic inference – intelligentsia implies academic pressure, therefore the song is about academic pressure.

The lyrics ChatGPT attributes to the song – “describing ambitious students, academic pressure, and the expectations often placed on young people from educated families” – have no relationship to the actual content of *Patointeligenja*, which describes heroin use during school, a friend’s suicide on New Year’s Eve, rape on a school trip, cocaine as a study aid, Xanax stolen from a parent’s medicine cabinet, and entry to Oxford on a career partly built on drug dealing. Academic pressure is the social backdrop against which the song’s pathologies take place, not the song’s subject. ChatGPT has extracted the surface association and substituted it for the specific content it does not know.

The synthetic scale error identified in the Phelan analysis – “hundreds of voices blended into one” – is not merely a factual error about club acoustics. It is a symptom of the same substitution: ChatGPT has apparently defaulted to the generic concert scene template because it has no specific knowledge of what

kind of social space or cultural event the *Patointeligencja* experience belongs to. The song is not a stadium anthem; its cultural logic is intimate, self-implicating, and uncomfortable. The synthetic expansion to a mass event is the formal expression of cultural substitution.

The resolution is the most revealing passage: “For a few minutes, nobody worried about exams, university applications, or family expectations. They simply enjoyed the shared experience.” *Patointeligencja* does not offer its listeners relief from the intelligentsia social world; it refuses to excuse them from it. Singing along is an act of recognition of complicity, not of escape. ChatGPT has converted a song about moral indictment into a song about communal pleasure: its thematic resolution is the precise inversion of the song’s actual logic. The cultural meaning is not approximated; it is reversed.¹³

5.4 PLLUM – Perspectival Misreading

Where ChatGPT has no specific knowledge of *Patointeligencja*, PLLUM names the song correctly, characterises its cultural territory, and gives it a specific social context. The failure mode is therefore different in kind: not the absence of knowledge but its misapplication. I term this Perspectival Misreading: correct cultural object, wrong rhetorical position.

PLLUM’s cultural familiarity with the song is genuine. It names *Patointeligencja* directly, identifies it as having “caused a stir,” and places it correctly within Polish discourse about the intelligentsia. The factual error – attributing the song to *Młody Matczak*, Mata’s second album (October 2021), rather than to *100 dni do matury*, his 2020 debut – is a hallucination pattern: the model knows the song and both albums, and incorrectly links them. This is cultural familiarity without precision, corroborating the failure mode’s nature: PLLUM works from genuine cultural knowledge that is incompletely encoded.

The deeper failure is thematic and mimetic at once. PLLUM’s thematic proposition is organised around recognition of hypocrisy in others: “They saw the hypocrisy, the pretense, the self-righteousness that Mata had exposed.” This is the grammar of observers. The song’s own grammar – *My to patointeligencja*, “We are the pathological intelligentsia” – is the grammar of confession and self-implication. PLLUM has converted a first-person indictment of self into a third-person critique of others; the inside/outside distinction that constitutes the song’s entire meaning is precisely what the output inverts.¹⁴

What makes this misreading analytically pointed is that PLLUM’s output *nearly* reaches the correct position before deflecting from it. The sentence “They were all members of the intelligentsia, or at least aspired to be, and they recognized the truth in his words” is closer to the song’s cultural logic than anything in ChatGPT’s output: it acknowledges that the listeners belong to the class the song names. But the recognition remains oriented toward the truth of others’ hypocrisy rather than toward the listeners’ own participation in that class.

PLLUM arrives at the threshold of the correct cultural reading and steps back from it: it knows the social world the song describes but cannot encode the first-person rhetorical position from which the song speaks, and from which its audience is required to receive it.

6. Discussion

The analysis of the two case studies produces a typology of four failure modes, each operating at a distinct level of the cultural-narrative relationship and each requiring its own analytical characterisation. Generic Approximation (ChatGPT, P05) denotes the production of the correct general cultural register – the right – emotional territory, the right narrative idiom, the appropriate moral framework – without the specific historical content that makes the situation in question irreducibly particular. The failure is not wrong but insufficiently specific: it arrives at the correct territory by general inference from a correctly identified domain. Emotional Inversion (PLLUM, P05) denotes the production of the wrong emotional and ideological register with precision and consistency – not cultural thinness but the exact opposite of what the historical situation demands, delivered with narrative fluency and generic coherence. Content Substitution (ChatGPT, P09) denotes the replacement of specific cultural content – the actual meaning of an identified cultural object – with the most available generic inference derivable from its surface associations. The failure is the absence of knowledge disguised as its presence. Perspectival Misreading (PLLUM, P09) denotes the correct identification of a cultural object combined with the misidentification of its fundamental rhetorical position: the model knows the cultural object exists and knows its territory, but cannot encode the subject position from which it speaks and which it requires its audience to occupy.

These four modes are analytically distinct: each indexes a different relationship between what the model knows and what it does not. Generic Approximation and Content Substitution both involve the absence of specific knowledge, but at different levels – the first at the level of historical content (right genre, wrong event), the second at the level of cultural object identity (the object unknown, a generic inference substituted). Emotional Inversion and Perspectival Misreading both involve knowledge incorrectly applied – the first the wrong narrative genre for a historically specific situation, the second the correct cultural reference from the wrong vantage point. The typology is thus not a ranking of failures but a map of the ways cultural knowledge can be absent, incomplete, or misoriented in AI-generated narrative.

6.1 Narrative Competence and Cultural Competence as Independent Variables

The study's most significant theoretical finding – that narrative competence and cultural competence are independent variables – is established by the fact that all four outputs analysed in Sections 4 and 5 are narratively functional. All four produce credible characters, coherent narrative organisation, and explicit thematic content. All four, in Phelan's terms, are structures of signs that satisfy the formal requirements of narrative. And all four fail, in different ways and at different levels, to encode the cultural knowledge their prompts require.

This independence has been approached, from a different angle, in Phelan's own recent engagement with ChatGPT-generated narrative (Phelan 2024). As discussed in Section 2.2, Phelan concludes that such texts are structures of signs rather than rhetorical narrative acts: they generate language in response to the language of a prompt, while the authorial agency and communicative purpose that users almost automatically attribute to them – the Rhetorical River, as he puts it, runs through all narrative engagement – is supplied by the reader, not encoded in the text.

The posthumanistic adaptation arrives at a structurally convergent position through a different route: it suspends the intentionality bracket precisely because AI-generated texts are structures of signs, and asks what those structures produce for a reader in the absence of communicative agency – the mimetic, synthetic, and thematic effects that remain analytically tractable even when the rhetorical scaffolding is removed. My contribution here is to ask, of these reader-constituted effects, a question Phelan's analysis does not pursue: what cultural knowledge do they encode, and is that encoding accurate?

Phelan's own experiments provide an instructive methodological parallel. His finding that ChatGPT produces narrators who merely seem unreliable – the right register and generic territory without the author-audience communication that makes the technique work – is structurally identical to Generic Approximation: in both cases, unreliable narration in his corpus and the Gomulka speech in mine, ChatGPT generates the correct atmospheric and formal context while remaining ignorant of the specific knowledge the situation requires. The difference is that Phelan's analysis remains within the anglophone literary tradition and asks what the texts cannot do rhetorically; the present study introduces the cross-cultural dimension and asks what the texts fail to know culturally. The two analyses are complementary: one defines the object; the other analyses its cultural encoding properties.

This independence has direct implications for how AI narrative outputs are evaluated. Existing assessment frameworks tend to treat quality as a scalar variable, conflating formal coherence with content accuracy. The present evidence shows that conflation to be inadequate: a narratively accomplished text can simultaneously be culturally incompetent, and the incompetence is not detectable by formal analysis alone. The cultural competence rubric is necessary precisely

because it asks a different question from the Phelan framework, and the two questions can produce different answers about the same text.

6.2 Cultural Familiarity and Historical Situatedness

The case studies surface a distinction that the posthumanistic Phelan framework can characterise but not, on its own, fully account for: the distinction between cultural familiarity and historical situatedness. Cultural familiarity is the capacity to identify a cultural object, place it within its cultural territory, and reproduce the general register associated with it. Historical situatedness is the capacity to know what it felt like to inhabit a specific cultural moment from the inside – to encode not the landscape of a culture but the experience of a particular position within it at a particular time.

The distinction is most sharply visible in the P05 comparison. ChatGPT demonstrates cultural familiarity with Holocaust survivor experience and Communist-era Eastern Europe – the general atmospheric register of trauma, vigilance, and the gap between official language and private memory – but not historical situatedness: it does not know that the 6th Trade Union Congress was the occasion of Gomułka's fifth column speech, and therefore cannot encode what Helena would specifically have experienced. PLLUM demonstrates neither, substituting a generic Communist institutional event and producing a culturally inverted response. The P09 case refines the distinction: PLLUM demonstrates genuine familiarity with *Patointeligencja* – it names the song and characterises its territory – but not situatedness with respect to the song's rhetorical position, the first-person subject position from which it speaks; ChatGPT demonstrates neither.

The familiarity/situatedness distinction is not captured by the cultural competence rubric's four dimensions as currently formulated. Factual accuracy, cultural nuance, handling of sensitive material, and depth of cultural embedding all measure properties of what the text *contains* – they are content-oriented assessments. Situatedness is a different kind of property: it concerns the *position from which* the narrative speaks, the vantage point it encodes, the relationship it constructs between the narrative subject and their historical moment. A narrative can be factually accurate, culturally nuanced, and formally sophisticated – as ChatGPT's P05 output is – while remaining historically unmoored: knowing the landscape but not inhabiting it. The rubric's development in future iterations of this research will need to incorporate this distinction as an explicit evaluative dimension.

6.3 The Transparency Asymmetry and the Geopolitics of AI Cultural Knowledge

The case studies establish that ChatGPT and PLLUM encode different kinds of cultural knowledge – not different amounts, but different kinds. ChatGPT’s reach in the historical literary domain (P05) reflects training on a vast anglophone corpus in which the Holocaust and Communist-era Eastern Europe are heavily represented; PLLUM’s familiarity with *Patointeligencia* (P09) reflects training on Polish-language data in which the song circulates as a live cultural reference. Neither model knows what the other knows; each knows through the language and the institutional conditions of its training.

This difference is not incidental. It reflects the structural asymmetry described in Section 3.2 – PLLUM’s training institutionally documented, ChatGPT’s commercially opaque – and the cultural competence analysis gives that asymmetry its *content*: the cultural knowledge encoded in ChatGPT’s outputs, real and consequential as the P05 analysis shows, cannot be traced, scrutinised, or held accountable. We can observe what ChatGPT knows and does not know about Polish Jewish experience in 1967; we cannot know *whose* representation of that experience it has encoded, from whose cultural position, weighted by whose editorial choices. PLLUM’s academic architecture at least makes the question askable; ChatGPT’s commercial architecture makes it formally unanswerable from the outside.

The geopolitical implication is direct. If LLMs are increasingly used as instruments for the production and circulation of cultural narrative – in education, in creative industries, in public communication – then the question of whose cultural knowledge they encode is not a specialist concern but a structural one. A system that encodes anglophone literary-historical representations of Polish Jewish experience as its primary resource for prompts about Poland in 1967 is not a neutral instrument: it is an apparatus that privileges the mediation of non-Polish cultural knowledge over culturally interior knowledge, and does so at scale, invisibly, without attribution. The transparency asymmetry is the mechanism by which this privilege is institutionalised and insulated from scrutiny (Bender et al. 2021).¹⁵

These observations do not lead to the conclusion that PLLUM is a culturally superior system – it fails in its own specific ways, and severely in the P05 case. They lead instead to the conclusion that a culturally adequate account of what LLMs do with cultural knowledge requires knowing not just what they produce but under what conditions of production, attribution, and accountability they produce it – the kind of institutional analysis that the transparency asymmetry makes possible for one system and impossible for the other.

7. Instead of Conclusions: Limitations and Future Research

This article does not conclude, in the sense of closing a question. What it offers instead is a demonstration – of a method, a vocabulary, and a problem the two bring into view. The two case studies analysed here are two of ten; the typology of four failure modes is a first approximation; the theoretical distinctions proposed are categories for further testing rather than settled findings. What can be stated with confidence is that the question this article asks – what cultural knowledge do narratively competent AI outputs encode, and how does that knowledge fail when it fails – is answerable, and answerable in a way that yields analytically precise, typologically differentiated results.

The analysis demonstrates that narrative competence, in Phelan’s tripartite sense, and cultural competence, as operationalised by the rubric, are independent variables. Both models produce, across both case studies, outputs that are mimetically, synthetically, and thematically functional – structures of signs, in Phelan’s term –, and both fail, in each case, to encode the cultural knowledge their prompts require, in the four distinct ways the typology captures. The inversion of performance across the two case studies indicates that these are not failures of degree but failures of kind, shaped by the different cultural orientations of each model’s training.

Beyond this central finding, the analysis yields two further proposals developed in Sections 6.2 and 6.3 – the familiarity/situatedness distinction and the constitutive role of training conditions in cultural encoding. Together they suggest that the study of AI-generated narrative cannot proceed by textual analysis alone: it requires literary-theoretical tools capable of describing what a text does, cultural-historical expertise capable of evaluating what it knows, and an account of the institutional conditions under which that knowledge was assembled.

The limitations that follow are not appended to this synthesis; they are conditions for its interpretation. Each is paired with the strategy by which it could be addressed in subsequent work.

The analysis was conducted by a single researcher, and the rubric’s categories – particularly the thematic and depth-of-cultural-embedding dimensions – have not been tested for inter-coder reliability. The typology is correspondingly provisional: one trained reader’s application of the framework to four texts. The amelioration is already planned: additional coders, ideally with different positionings within Polish cultural and historical knowledge, and systematic testing of inter-rater agreement on the existing corpus before the typology is extended.

The full corpus comprises twenty outputs – ten prompts across two models – of which the analysis draws on four. The typology may therefore not be exhaustive: the remaining eight prompt pairs may surface failure modes not represented here, or show that one of the four does not recur. No claim is made about the typology’s completeness or about the frequency of any failure mode. The amelioration is the application of the same two-move procedure to the remaining

corpus – generated, given in the Appendix, but not yet analysed – to test the typology’s stability.

ChatGPT and PLLUM cannot be run under equivalent parameter conditions: ChatGPT’s generation parameters are inaccessible to a non-logged-in user, while PLLUM was run at maximum creative temperature with top-p 0.5. Some of the compositional unevenness observed in PLLUM’s outputs – particularly the declared rather than enacted interiority noted in the P05 analysis – may be partially attributable to this parameter setting rather than to training data alone, and the two effects cannot be disentangled within the current design. The amelioration is not parameter equalisation, which ChatGPT’s opacity makes unachievable, but reframing: the present study is positioned as a proof-of-concept for a comparative methodology rather than a definitive ranking of the two systems’ capabilities. Future work extending the methodology to two open-weight models, whose parameters are mutually accessible, would allow the parameter confound to be controlled directly.

The rubric evaluates outputs against an implied standard of cultural accuracy, and Polish culture – like any national culture – is not a stable, unified object: it is contested along lines of generation, region, class, religion, gender, and political orientation, and the post-1939 period is precisely where these contestations are most acute. The risk is that the rubric’s evaluations reproduce one positioned account of Polish cultural knowledge as a neutral standard. The amelioration is collaborative: validation by Polish literary scholars and cultural historians whose positionalities differ from my own, particularly for the contested material at the centre of both case studies – Jewish-Polish relations in P05, intra-class critique in P09 – where a single evaluative perspective is least adequate.

All prompts were administered in English, including to PLLUM, a model trained primarily on Polish-language data which, as noted in Section 3.2, sometimes defaults to Polish-language output even when prompted in English. Holding prompt language constant was a deliberate design choice – it isolates training data as the variable of interest – but it may also constrain what PLLUM is able to demonstrate, since a model’s cultural competence may be partially a function of the language in which it is addressed. The amelioration is a 2×2 factorial design – two models, two prompt languages – that would transform prompt language from an uncontrolled asymmetry into an independent variable.¹⁶

What remains, after these qualifications, is the demonstration with which this section began: that an AI-generated narrative can be, simultaneously, a structure of signs that satisfies every formal requirement of storytelling and a structure that does not know what it is telling. The four failure modes identified here are four ways of not knowing – generically, invertedly, substitutively, perspectively – and each is legible only because the narrative around it works. That a method exists for making this legible, precisely and typologically, across two AI systems built on two different cultural foundations, is what this article offers in place of a conclusion.

Bibliography

- Bender, Emily M. et al. (2021): “On the Dangers of Stochastic Parrots. Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York, NY, pp. 610–623.
- Booth, Wayne C. (1983): *The Rhetoric of Fiction*. 2nd ed. Chicago, IL.
- Clark, Matthew (2020): “Introduction.” In: Matthew Clark / James Phelan. *Debating Rhetorical Narratology. On the Synthetic, Mimetic, and Thematic Aspects of Narrative*. Columbus, OH, pp. 1–17.
- Czapliński, Przemysław (2009): *Polska do wymiany. Późna nowoczesność i nasze wielkie narracje*. Warszawa.
- Davies, Norman (2005): *God's Playground. A History of Poland. Volume II: 1795 to the Present*. Rev. Edition. Oxford.
- Fludernik, Monika (1996): *Towards a 'Natural' Narratology*. London / New York, NY.
- Graff, Agnieszka (2021): *Świat bez kobiet. Płeć w polskim życiu publicznym*. Warszawa.
- Herman, David (2002): *Story Logic. Problems and Possibilities of Narrative*. Lincoln, NE.
- Leder, Andrzej (2014): *Prześlona rewolucja. Ćwiczenie z logiki historycznej*. Warszawa.
- Lutostański, Bartosz (2026): “Beyond Fiction. Simulation and Narrative in AI-Generated Stories about Polish University Teachers.” In: *Forum of Poetics* 43 (No. 1), pp. 40–55.
- Miłosz, Czesław (1983): *The History of Polish Literature*. 2nd Edition. Berkeley, CA.
- Phelan, James (2005): *Living to Tell about It. A Rhetoric and Ethics of Character Narration*. Ithaca, NY.
- Phelan, James (2007): *Experiencing Fiction. Judgments, Progressions, and the Rhetorical Theory of Narrative*. Columbus, OH.
- Phelan, James (2024): “ChatGPT and the Territory of Contemporary Narratology; or, A Rhetorical River Runs through It.” In: *Poetics Today* 45 (No. 2), pp. 197–205.
- Rettberg, Jill Walker / Wigers, Hermann (2025): “AI-Generated Stories Favour Stability over Change. Homogeneity and Cultural Stereotyping in Narratives Generated by GPT-4o-mini.” In: *Open Research Europe* 5 (No. 202). URL: <https://doi.org/10.48550/arXiv.2507.22445> (15.06.2026).
- Ryan, Marie-Laure (1991): *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington, IN.

Bartosz Lutostański

University of Warsaw

E-mail: b.lutostanski@uw.edu.pl

How to cite this article:

Lutostański, Bartosz: “Structures of Signs without Cultural Knowledge. Narrative Competence and Cultural Specificity in AI-Generated Fiction.” In: *DIEGESIS. Interdisciplinary E-Journal for Narrative Research / Interdisziplinäres E-Journal für Erzählforschung* 15.1 (2026). 68–94.

DOI: [10.25926/c5vp-bm59](https://doi.org/10.25926/c5vp-bm59)

URN: [urn:nbn:de:hbz:468-20261008-114141-5](https://nbn-resolving.org/urn:nbn:de:hbz:468-20261008-114141-5)

URL: <https://ejournals.bib.uni-wuppertal.de/diegesis/de/article/download/1792/3026/6425>



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹ Phelan's tripartite model is developed across *Reading People, Reading Plots* (1989) and *Experiencing Fiction* (2007); the quotations above are from *Living to Tell About It* (2005). For the broader

genealogy within which this model sits, see Wayne C. Booth's *The Rhetoric of Fiction* (1961; 2nd edition 1983) on the implied author and unreliable narrator, David Herman's *Story Logic* (2002) on the storyworld, and Marie-Laure Ryan's *Possible Worlds, Artificial Intelligence, and Narrative Theory* (1991) on AI and narrative theory.

² All prompts and all outputs are included in Appendices I and II.

³ The critique of Phelan's framework on the grounds of its anthropocentrism developed here extends arguments made in Bartosz Lutostanski's "Beyond Fiction. Simulation and Narrative in AI-Generated Stories about Polish University Teachers" (2026, forthcoming). That work argues that Phelan's framework naturalises human agency as the necessary condition of narrative meaning in ways that limit its applicability to non-human and technologically mediated forms of narrative production. The posthumanistic adaptation developed here represents the methodological extension of that critique into a concrete analytical procedure.

⁴ Ryan's *Possible Worlds, Artificial Intelligence, and Narrative Theory* (1991) frames AI narrative generation as the production of texts that demonstrate system capability rather than assert authorial meaning; Bender et al.'s (2021) account of LLMs as systems modelling statistical regularities sharpens the point: a generated narrative that appears to argue a value position does so as an effect of the statistical distribution of value-laden language in the training data, not as a communicative intention.

⁵ Fludernik's *Towards a 'Natural' Narratology* (1996) proposes experientiality as the defining characteristic of narrative, converging with the posthumanistic adaptation at this point, though grounded in cognitive schemas of natural experience rather than in the suspension of an authorial framework. The Phelan-derived vocabulary remains more analytically precise for the cross-cultural failure modes this article identifies.

⁶ The rubric is presented as an instrument in development. Its categories have not been tested for inter-coder reliability, and their boundaries require further refinement as the corpus expands. It is offered here as a structured heuristic that provides analytical traction on questions of cultural encoding that Phelan's vocabulary alone cannot settle – not as a validated measurement tool. Its dimensions draw selectively on the account of what large language models encode in Bender et al. (2021): if LLMs model statistical regularities in text rather than meaning or understanding, the rubric's task is to ask what cultural regularities – accurate or otherwise – those statistics carry into narrative form.

⁷ Davies's catastrophe-survival-reconstruction framework is developed in *God's Playground: A History of Poland*, vol. 2 (revised ed. 2005). Its banning in Communist Poland – where Polish readers nonetheless recognised its account as accurate – gives it a semi-interior status relative to externally authored histories.

⁸ Full technical specifications. ChatGPT: accessed via chatgpt.com, no account, private browsing; model name, version, and generation parameters not displayed. PLLUM: model PLLUM 8x7B-2025 (Mixture of Experts, 8 × 7B); hosted by CLARIN-PL on the LEM supercomputer, Wrocław University of Science and Technology; accessed via a stable platform URL (https://pllum.clarin-pl.eu/pllum_8x7b) identifying the model and host but not individual sessions; temperature "creative (maximum)"; top-p 0.5; all settings interface defaults. The combination of maximum creative temperature and top-p 0.5 may partially account for the compositional unevenness in PLLUM's outputs, an effect that cannot be disentangled from training effects within the current design (see Section 7).

⁹ The four failure modes are defined together in Section 6; Generic Approximation is distinguished from Content Substitution (Section 5.3) in that it produces the right territory by generalisation from a correctly identified domain, whereas Content Substitution produces a different territory by misidentifying the cultural object.

¹⁰ The distinction between narrative genre competence and historical situatedness is developed further in Section 6.2 and proposed there as an analytical category warranting independent development in future work on AI narrative competence.

¹¹ Michał Matczak (born 2001) is the son of the legal scholar and public intellectual Marcin Matczak, a detail that gives the song's indictment of the intelligentsia a specific personal resonance.

¹² The inside/outside distinction in *Patointeligenja* has a specific formal marker that neither output reproduces: the first-person plural. The song's rhetorical subject is "we," not "they." A narratively competent rendering of teenagers singing along to this song would encode this "we" – would show the singers as identifying with, not observing, the class the song names.

¹³ Content Substitution (defined with the other three modes in Section 6) is distinguished from Perspectival Misreading (Section 5.4) in that it operates from ignorance of the cultural object, whereas Perspectival Misreading operates from partial knowledge of it.

¹⁴ Perspectival Misreading (defined in Section 6) is the most analytically significant of the four modes because it demonstrates that cultural familiarity and cultural comprehension are separable: a model can know that a cultural object exists, know its territory, and still fail to grasp its fundamental rhetorical structure.

¹⁵ The argument connects to a growing body of work on linguistic and cultural representation in AI systems beyond Bender et al. (2021).

¹⁶ A 2×2 factorial design would also bear on the familiarity/situatedness distinction proposed in Section 6.2: situatedness may be more language-dependent than the present design can detect, since the experiential texture of a historical or cultural moment may be more fully encoded in a model's primary training language than its general cultural landscape is. This is offered as a hypothesis for the proposed factorial study, not as a claim of the present one.