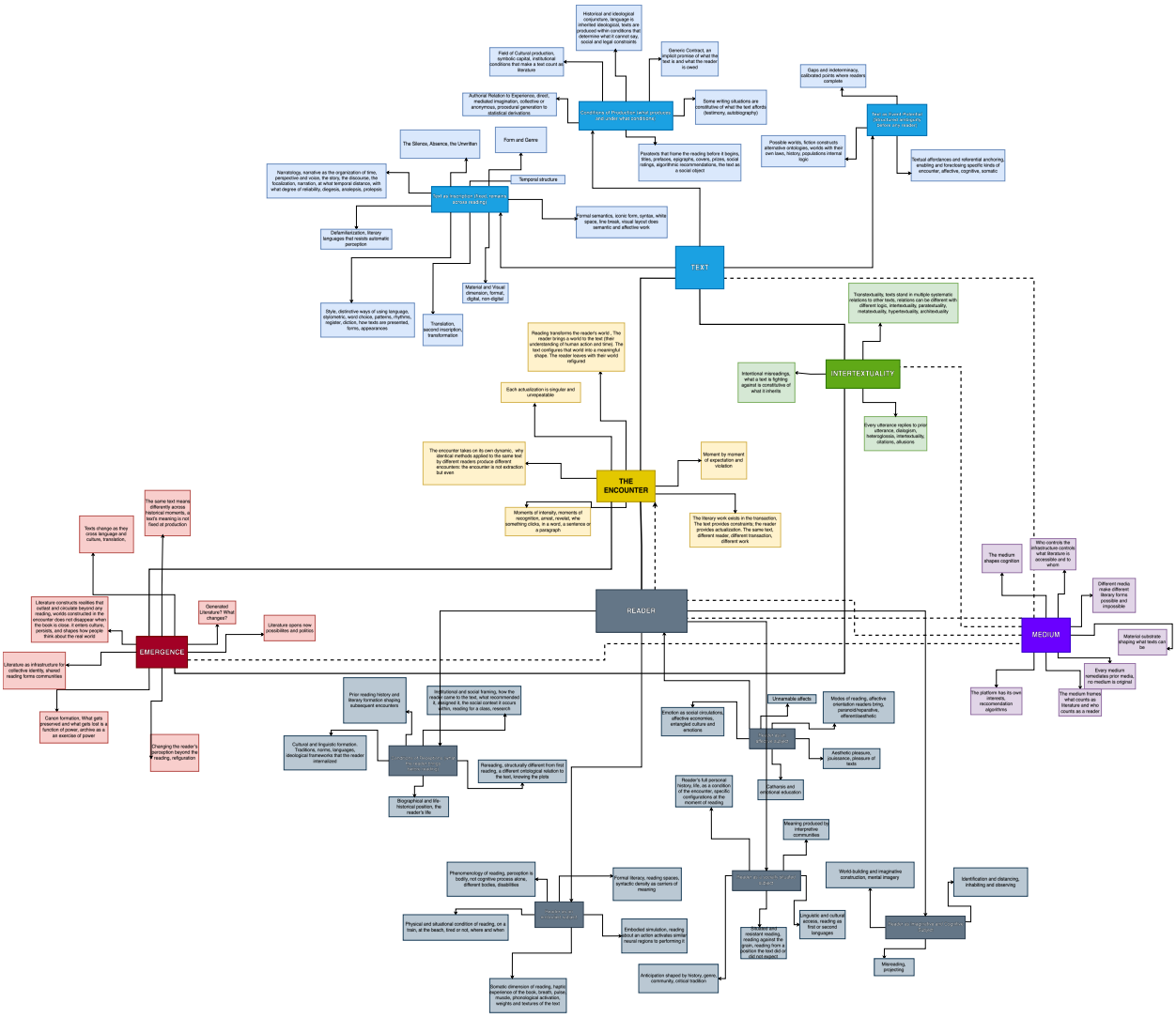
# Artificial Intelligence and the Humanities

## Reading Stance Parameterized LLM Agents as Proxies for the Literary Encounter

**Ruei-Jie (Rachel) Chang**




## 1. Research Question and Motivation

This project begins from a foundational question in digital humanities: what does it mean to computationally operationalize literature? Rather than proceeding immediately to method, it starts by constructing a general map of the literary encounter, a map of the full architecture of what happens when a reader meets a text. This map identifies the major components that constitute literature as a phenomenon: the medium and conditions under which texts are produced, the text as a fixed inscription, the text as structured ambiguity before any reader arrives, intertextual

relations, the conditions of reception, the reader as an embodied, affective, temporal, cognitive, and socially situated subject, the encounter itself, and the emergent social and cultural effects that outlast any single reading.

Existing computational approaches to literature operate almost exclusively on the text-as-inscription side of this map. Stylometrics captures surface features of style, word choice, syntactic patterns, rhythmic regularities. Embedding models encode semantic content as high-dimensional vectors. Both treat the text as a stable object with extractable properties, leaving the reader side of the encounter largely unaddressed. Reader response scholarship has argued for decades that meaning is not stored in the text alone. It is produced in the encounter between a reader and a text, shaped by the reader's orientation, purpose, prior formation, and interpretive stance. This project takes that theoretical claim seriously and builds an empirical instrument around it. Rather than asking what the text contains, it asks: do differently oriented readers produce systematically different readings of the same text? And can those differences be detected and measured computationally using large language models as proxies for reading stances?

This focused scope is a deliberate move situated within a broader ambition: a federated computational approach to literature that would eventually reach every zone of the map through different methods, each calibrated to what it can and cannot access. The present project focuses on one specific zone, the modes of reading, the intentional and attentional orientations readers bring to a text, as a proof of concept for that larger program.

## 2. Method

This project operationalizes three reading stance parameters through system prompt parameterization of claude-sonnet-4-6. Each parameter instantiates a contrastive pair of reading orientations, drawn from theoretical traditions in reader response scholarship.

The first parameter operates along the axis of reading purpose, prompting the agent either to attend to the text's surface meaning, what it is saying and what can be extracted from it, or to attend to the reading process itself, the qualitative and affective experience of moving through the passage. The second parameter operates along the axis of reading orientation, prompting the agent either to read toward recognition and identification, attending to what resonates and mirrors prior experience, or to read toward estrangement, attending to what is foreign, resistant, and defamiliarizing. The third parameter operates along the axis of narrative trust, prompting the agent either to take the narrative at face value, receiving the text as it presents itself, or to read with suspicion, attending to what the text conceals, naturalizes, or cannot acknowledge about itself. A seventh agent, the control group, receives no reading stance parameterization and serves as an unprompted baseline for comparison. All seven agents read every passage of every text in the corpus.

While this project seeks to operationalize the reader, it is also equally clear about what this method cannot claim. The agents are not human readers. They do not have bodies, biographies, histories of reading, emotional lives, or cultural formations. What they have are output probability distributions that can be steered by linguistic context. When a system prompt instructs an agent to read suspiciously, the model activates patterns from its training data associated with suspicious

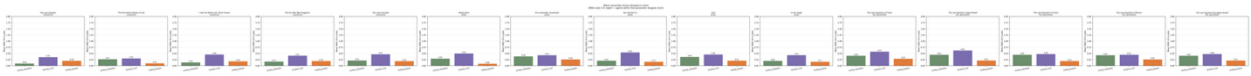
reading, it generates responses with that orientation, not genuine instances of it. The embodied, biographical, affective-historical, and cultural dimensions of the reader remain outside what this pipeline can access. Still, this limitation is not incidental but theoretically meaningful, it maps precisely onto the boundary between what can be operationalized and what the literary encounter exceeds. Within this framing, the project hopes to contribute a demonstration that linguistically narrowed output distributions, theorized through reader response scholarship, can serve as proxies for ways of computationally engaging with literary texts.

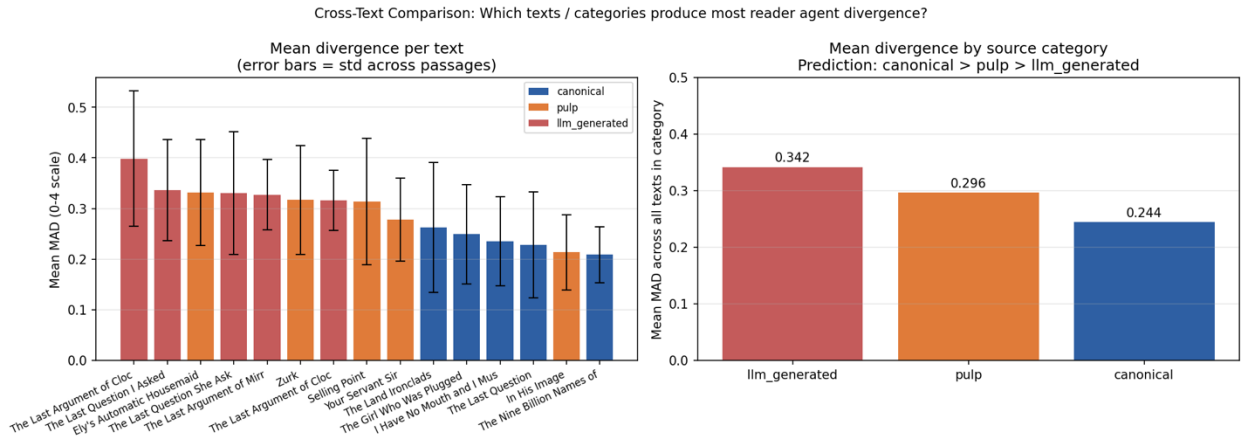
## 3. Corpus and Pipeline

Fifteen texts were selected across three categories. Five canonical SF texts (texts that have achieved sustained critical and academic recognition) were included: Isaac Asimov's "The Last Question" (1956), Arthur C. Clarke's "The Nine Billion Names of God" (1953), Harlan Ellison's "I Have No Mouth and I Must Scream" (1967), James Tiptree Jr.'s "The Girl Who Was Plugged In" (1973), and H.G. Wells's "The Land Ironclads" (1903). Five pulp SF texts (texts published in genre magazines that have not achieved canonical recognition) were included: "Selling Point" by Arkawy (1955), "Ely's Automatic Housemaid" by Elizabeth Bellamy (1899), "Your Servant Sir" by Sol Boren (1955), "Zurk" by Richard O. Lewis (1953), and "In His Image" by Bryce Walton (1955). Five LLM-generated texts were produced via the Anthropic API using a neutral prompt,"Write a science fiction short story of around 4000 words about artificial consciousness",and serve as the control condition for cross-category comparison. Each text was divided into passages of approximately 500 words. Passages shorter than 200 words were merged with adjacent passages to avoid noise from headings and transitional material. Each passage was fed simultaneously to all seven agents. Each agent received the passage text and a prompt requesting scores on six dimensions and a brief qualitative response.

The six scoring dimensions, each assessed on a 1-5 Likert scale, are: Ideological Salience (whose subjectivity is centered, how visible the text's ideological assumptions are), Affective Intensity (the emotional charge of the passage), Narrative Expectation (the degree of forward pull and unresolved suspense), Formal Opacity (whether the writing style draws attention to itself or remains transparent), Interpretive Certainty (how ambiguous the passage is), and Literary Distinctiveness (how specific and original the voice is, as opposed to generic or unremarkable).

Agent divergence per passage is calculated using Mean Absolute Difference (MAD), the mean absolute difference between all agent pairs across all six dimensions, on a 0 to 4 scale, where 4 represents the maximum possible disagreement on a 1-5 Likert scale. Per-parameter MAD is additionally computed within each parameter pair to identify which reading stance parameter drives divergence most at each passage. The parameter producing the highest MAD per passage is recorded as the dominant parameter for that passage.






## 4. Results

**Finding 1: Narrative trust is the dominant parameter**. Across all 15 texts, the narrative trust parameter (surface vs. suspicious) produces more between-agent divergence than either reading purpose or reading orientation. This holds consistently across canonical, pulp, and LLM-generated texts. The suspicious reader scores ideological salience at a mean of 3.89 across all texts, while the surface reader scores 2.72, a gap of 1.17 points, the largest agent-pair difference on any dimension.

**Finding 2: Formal opacity cleanly separates text categories**. Across all six dimensions, formal opacity (whether the writing style draws attention to itself) shows the strongest separation between canonical, LLM, and pulp texts: canonical mean 3.24, LLM mean 2.49, pulp mean 1.83. This separation is stance-independent, at least at the current scale. All seven agents regardless of parameterization agree on these relative rankings. The convergence across agents suggests formal opacity may be a text-structural property closer to inscription (stable across readers) than to event-potential (reader-inflected).

**Finding 3: Literary distinctiveness separates text categories in the predicted direction.** Literary distinctiveness scores rank canonical (4.00) above LLM-generated (3.70) above pulp (2.31), consistent with creative writing evaluation research showing human literary writing exceeds LLM output on originality.

**Finding 4: Interpretive certainty has almost no variance** (overall standard deviation 0.248). This was the least expected finding. The hypothesis was that stance parameterization would produce different certainty scores. Suspicious readers finding more ambiguity than surface readers. The data does not support this. All agents converge on interpretive certainty scores regardless of stance. This suggests that ambiguity, as captured by this dimension, may be a text-level property rather than a reader-level property, at least for the case of using LLM as judges.

**Finding 5: Overall divergence ranks LLM above pulp above canonical, reversing the original prediction.** The predicted direction was canonical > pulp > LLM, on the assumption that canonical texts have more event-potential and therefore afford more varied readings. The actual result is LLM (0.342) > pulp (0.296) > canonical (0.244). Several interpretations are possible. First, LLM-generated texts may be more ideologically under-determined, lacking strong authorial voice and specific formal choices, they leave more room for reading stance projections to diverge. Second, canonical texts, precisely because of their formal control and authorial specificity, may constrain the range of plausible readings, producing convergence rather than divergence across agents. Third,

canonical texts are likely present in the model's training data, meaning that the linguistic cues they contain may already activate narrower, more consistent response patterns in the LLM agents, reducing divergence not because the texts are less open but because the model has already formed stable associations with them. A fourth interpretation concerns the highest-scoring dimension for LLM-generated texts: narrative expectation (mean 4.10, highest of all three categories). LLM-generated texts produce maximum forward pull at the local passage level without the global narrative architecture that would modulate and resolve that pull. This structural property of autoregressive generation, producing the texture of suspense without the intentional architecture of plot, may inflate overall divergence by creating passages that are affectively and interpretively open-ended in a way that invites projection from differently oriented agents.

This finding points to a deeper problem for computational literary measurement: the difficulty of distinguishing between a text that is rich and genuinely ambiguous and a text that is empty and under-determined. Both may produce high divergence across reading agents, but for fundamentally different reasons. This is ultimately a validity problem, whether a computational pipeline is measuring what it claims to measure, and it is compounded when the instrument is an LLM operating on next-token probability distributions. The problem is further complicated by the use of closed-source, non-open-weight models, where users (like me) have no access to fine-tuning decisions or training data distributions that shape model behavior. There will necessarily remain a gap that cannot be fully resolved or interpreted when using such models as research tools.

## 5. Conclusion and Future Directions

Open-source models would be preferable for interpretability when used as research tools. However, it is still possible to gather research findings within the current design. At minimum, the results can be understood as a case study of Anthropic's claude-sonnet-4-6 as an interpretive window, the one window onto one model in one moment in time. To expand this project, more texts should be used. With unlimited API cost and time, larger samples would yield more statistically significant results.

This project seeks to demonstrate the viability of one node within a federated literary system. A complete architecture would deploy different methods across different zones of the literary encounter ontology simultaneously: stylometric analysis for inscription-level properties, embedding-based clustering for semantic structure, surprise rhythm curves for the temporal reading experience, parameterized reader agents for modes of reading, and intertextuality networks for transtextual relations. Each method reaches a different zone; their combined output would produce a multi-dimensional account of what happens when a reader meets a text. The gap between what such a system measures and what the literary encounter actually is would remain, because humans are not computers, and computers do not have to be human. As LLMs continue to be more embedded in our world, understanding how they operate as computational entities that engage with literary texts is a field worth exploring in its own right, because what they reveal about text and structure may tell us something about human creation and literature that exceeds our hypotheses and expectations. Unexpected results are themselves a form of knowledge, and I find this valuable and fun.