Position: A Semiotic-Hermeneutic Approach to Evaluating Meaning in LLM Summaries via the Inductive Conceptual Rating Metric

Natalie Perez
University of Hawai'i, USA

Sreyoshi Bhaduri
Private Corporation, USA

Aman Chadha
Google DeepMind, USA *

Abstract

Meaning in human language is relational and context-dependent, and it emerges through a dynamic system of signs rather than fixed relationships between words and concepts (Saussure, 1916). Relatedly, the fields of Semiotics and Hermeneutics emphasize that meaning emerges through contextually situated interpretive processes (Gadamer, 1975; Perez et al., 2026), a complexity that has historically challenged computational systems designed to process, represent, and evaluate human language (Shan, 2025). Building on these perspectives, this article advances an interdisciplinary framework for evaluating meaning in machine-generated language and introduces the Inductive Conceptual Rating (ICR) metric, a mixed-method approach, grounded in inductive content analysis and reflective thematic analysis that assesses semantic accuracy and meaning alignment in generative artificial intelligence (GenAI) outputs. The ICR metric is applied in an empirical study that compares thematic summaries generated by large language models (LLMs) with the human-generated output across five datasets (N = 50-800). Results show that although models achieve high linguistic similarity scores, they consistently unperformed relative to human outputs in capturing recurring, contextually grounded meanings. This work concludes with by discussing implications for meaning evaluation and future research recommendations.

1 Introduction

"On the morrow he will leave me, as my Hopes have flown before. Then the bird said Nevermore. Take thy beak from out my heart, and take thy form from off my door! Quoth the Raven Nevermore. And my soul from out that shadow that lies floating on the floor Shall be lifted, nevermore!" — *The Raven*, Edgar Allan Poe (1845)

Saussure (1916) described language as a system of signs composed of a *signifier* (word) and a *signified* (concept), whose meaning arises through relationships within the broader sign system (Saussure, 1916). This dynamic is illustrated in Edgar Allan Poe's work, *The Raven* (1845), where the repeated word "*Nevermore*" functions as a single signifier whose signified shifts across the poem, from lost love, to the search for comfort, to existential despair (Poe, 2013). The word's changing meaning (Nevermore) exemplifies *polysemy*, where one signifier carries multiple signifieds within a single text (Purba and Damanik, 2025).

Despite the inherently fluid and context-dependent nature of meaning, machine learning (ML) and large language model (LLM) outputs are often evaluated using automated metrics that treat words as fixed units of meaning (Bhargavi, 2025). However, this simplification can overlook the relational and dynamic aspects of meaning that texts, like *The Raven*, exemplify, raising concerns about the extent to which computational scores can serve as reliable proxies for true semantic meaning, particularly when models are tasked with interpreting reference texts, such as transcripts or survey data, whose contextual relationships, lived experiences, and evolving meanings may not be fully represented in the data on which the models were trained.

This challenge is central to computational linguistics and natural language processing (NLP). A common approach for modeling meaning is *word embeddings*, which encode words as vectors in high-dimensional semantic space (Apidianaki, 2023). Traditional static embeddings (e.g., BoW, TF-IDF, Word2Vec, GloVe) assign a single vector to each word, regardless of context, limiting their ability to represent polysemy (Bhargavi, 2025;

*Work done outside of role at Google.

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1036–1047

Grindrod, 2024). That said, contextualized embeddings (e.g., BERT, RoBERTa, DistilBERT, ELECTRA, ALBERT) improve on this by generating context-dependent representations, enhancing performance in tasks such as topic modeling, summarization, and sentiment analysis (Viegas et al., 2025; Alizadeh and Seilsepour, 2025; Gangundi and Sridhar, 2025; Paneru et al., 2025). Yet even these models approximate meaning through statistical patterns and co-occurrence rather than the cultural, historical, and experiential contexts that shape human interpretation (Arseniev-Koehler, 2024; Bhaduri et al., 2024; Saussure, 1916).

Consequently, evaluating the semantic fidelity, which we define as the accuracy of meaning representation, of LLM outputs requires going beyond surface-level, linguistic metrics, to also account for meaning within reference texts. To achieve this deeper examination, this study introduces the Inductive Conceptual Rating (ICR) metric, which uses a systematic methodology that combines interpretative thematic coding and content analysis to generate a metric for assessing the sematic fidelity of LLM outputs against a reference text. The ICR constructs a human baseline, based on human coder outputs, to determine an interpretative position on the meaning of a reference text and compare that meaning with LLM-generated outputs.

2 Purpose

The purpose of this study is to develop and demonstrate the Inductive Conceptual Rating (ICR), a semiotic (i.e., focused on how signs and symbols convey meaning) and hermeneutic (i.e., concerned with the theory and methodology of interpretation) informed metric that systematically evaluates the semantic fidelity of machine-generated language. In "The Raven" example, semiotic would convey the evolving relationship between the signifier (the repeated word 'Nevermore') and its various signified concepts, such as loss or finality; whereas hermeneutic would be the overarching interpretive process that integrates the narrator's psychological state and the bleak December setting to transform a bird's mechanical repetition into a profound existential prophecy. By integrating Reflective Thematic Analysis (RTA) and Inductive Content Analysis (ICA), ICR aims to capture the context-dependent and relational nature of meaning, based on human coders, to address the limitations of traditional automated evaluation metrics that treat words as fixed or context-independent.

3 Introducing the Inductive Conceptual Rating (ICR)

3.1 Epistemological Justification

From a semiotic perspective, quantitative and qualitative approaches reflect different ways of modeling the relationship between signifiers and signifieds. According to Borgstede and Scholz (2021), in variable-based quantitative modeling, relational structures are treated as functional mappings:

$$\forall i: \quad y_i = f(x_i)$$

Each output y_i is assumed to follow deterministically from input x_i (Borgstede and Scholz, 2021). In LLM evaluation, this reduces signifiers to tokens, n-grams, and embeddings; from this lens, signifiers are viewed as stable, context-independent units with fixed meanings. Variability across datasets reflects structural sensitivity or metric limitations rather than the fluid and relational nature of meaning.

Alternatively, Borgstede and Scholz (2021) illustrate case-based qualitative modeling treating meaning as emergent and context-dependent:

$$\exists i : XYZ_i$$

From this lens, both recurring patterns and their variations can be identified in specific cases, with variation itself offering epistemological insight by revealing contextual nuance and interpretive flexibility (Braun and Clarke, 2022; Creswell and Poth, 2016; Borgstede and Scholz, 2021). Relatedly, signifiers and signifieds are viewed as co-constructed relationships shaped by context, rather than as fixed elements mapped mechanically between a signifier and signified (Saussure, 1916; Creswell and Creswell, 2017).

Framed semiotically, quantitative metrics capture structural or surface-level regularities, while qualitative approaches illuminate how meaning is negotiated, relational, and context-dependent. It is important to note that the ICR metric does not claim to quantify the entirety of a text's meaning; rather, it quantifies interpretive understandings of a text through systematic procedures grounded in structured human judgment. This makes the ICR metric unique because it is inherently situated and contingent upon interpretation. Its interpretive flexibility makes it particularly valuable for examining

meaning, while also allowing variation in results depending on the perspectives, contexts, and interpretive frameworks brought to the analysis.

In this light, while the ICR metric produces a quantitative output in the form of a coverage score, that score is grounded in qualitative interpretive work and acknowledges that the baseline against which outputs are evaluated is itself constructed, situated, and plural. Whereas automated metrics often treat meaning as fixed and stable, ICR conceptualizes meaning as interpretive and variable, treating the baseline as a range of plausible interpretations and the score as an indication of how well a model output aligns within that interpretive range.

3.2 Scope and Applicability with NLG Tasks

The ICR metric is designed for natural language generation (NLG) tasks where semantic representativeness is the primary evaluative concern; that is, tasks where the central question is not simply whether an output is fluent or grammatically correct, but whether it accurately captures the conceptual content that conveys the meaning of a source text.

To this end, ICR is well-suited to:

- **Abstractive summarization**, where fidelity to key concepts is a primary criterion of quality.
- **Thematic synthesis**, including survey, interview, or document analysis, where models are required to identify and organize recurrent conceptual patterns across texts.
- **Open-ended generation evaluated against a reference corpus**, in which a human-constructed interpretive baseline serves as a meaningful standard for comparison.

ICR is less appropriate for:

- **Factual question answering**, where correctness is typically binary and not dependent on thematic interpretation.
- **Creative generation**, where interpretive plurality is constitutive rather than a limitation for measurement.

ICR is particularly valuable for examining summarization or thematic synthesis tasks because these contexts prioritize patterned meaning and interpretive meaning coverage. In this way, outputs that omit core themes or introduce unsupported ones fail this communicative objective, irrespective of surface-level fluency. By operationalizing concept coverage against a human-constructed interpretive baseline, ICR provides a measure of semantic meaining that is not readily captured by automated evaluation metrics.

3.3 Analytical Approach

ICR evaluates meaning in generative AI outputs through a two-stage interpretive process grounded in Reflective Thematic Analysis (RTA) within the tradition of Thematic Analysis (Braun and Clarke, 2006) and an Inductive Content Analysis (Zhang and Wildemuth, 2009).First, a blind RTA is conducted on reference texts prior to exposure to any LLM outputs, following Braun and Clarke's six-step analytical approach to theme development. In line with their emphasis on the active, interpretive role of the researcher, themes are constructed by human researchers through repeated engagement with the data, moving recursively between data extracts, codes, and potential themes (Braun and Clarke, 2006). This process foregrounds reflexivity, acknowledges the situated nature of interpretation, and treats meaning as produced through analytic practice rather than residing inherently within the text.

Second, drawing on Zhang and Wildemuth's (2009) inductive content analysis, the interpretively constructed themes are systematically compared against LLM-generated outputs using their structured, multi-step process of coding, categorisation, and abstraction. This 8-step procedure enables a rigorous progression from initial coding to the development of higher-order categories, ensuring that meaning is both systematically organised and grounded in the data (Zhang and Wildemuth, 2009). Within this stage, conceptual meanings are examined in terms of how they are expressed, transformed, attenuated, or omitted across outputs, with alignment assessed through patterns of convergence, partial correspondence, or divergence. By sequencing reflexive thematic construction prior to any model exposure, and then subjecting those meanings to a structured inductive comparison process, ICR integrates Braun and Clarke's interpretive, meaning-making orientation with Zhang and Wildemuth's systematic framework for category development. This combined approach enables a transparent and methodologically robust evaluation of semantic fidelity and conceptual congruence be-

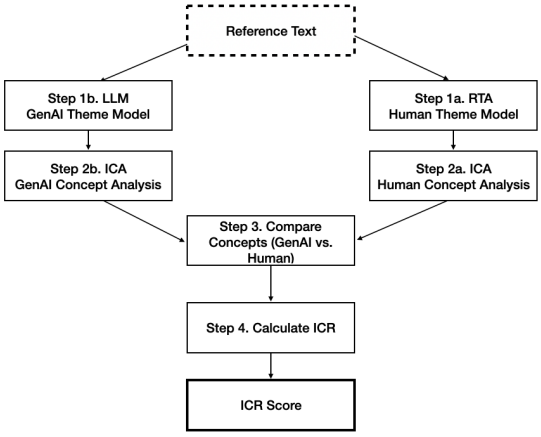

Figure 1: Flowchart illustrating the four-step Inductive Conceptual Rating (ICR) evaluation procedure, including Reflective Thematic Analysis (RTA), Inductive Content Analysis (ICA), comparative analysis, and final metric computation.

tween human- and model-generated texts.

3.4 Procedures

There are four steps required to achieve an ICR score, with each step described below (see Figure 1 for more details).

3.4.1 Step 1. Conduct a Reflective Thematic Analysis (RTA)

The first step (Step 1a) of the ICR process applies Reflective Thematic Analysis (RTA) to the reference or "golden" dataset to inductively identify patterns of meaning, recurring themes, and contextual relationships Thematic Analysis. Following Braun and Clarke's six-phase framework, the analysis proceeds through data familiarization, initial coding, theme generation, thematic review, theme refinement, and final insight development (Braun and Clarke, 2006). Across these phases, RTA foregrounds reflexivity, with the researcher actively engaged in meaning-making rather than passive identification of themes, and treats themes as interpretive constructs developed through iterative engagement with the dataset. This process can be further supported by the principle of meaning saturation, ensuring that thematic development captures both explicit and latent dimensions of meaning through sustained analytic immersion Thematic Analysis (Hennink and Kaiser, 2022). The outcome of this step are human-generated themes, which functions as a structured but interpretively derived representation of meaning within the reference corpus. Importantly, to preserve analytic integrity and prevent contamination from model-induced framing effects, the RTA is conducted in a fully blind manner prior to any exposure to LLM-generated outputs. This sequencing ensures that thematic construction remains grounded exclusively in the human dataset, minimizing interpretive bias and safeguarding the reflexive independence of the analysis. The resulting themes, therefore, serves as the epistemic baseline for subsequent comparative evaluation within the ICR framework.

To ensure rigor in the RTA process, multiple expert coders should independently analyze the dataset and compare their interpretations. While there is debate in the field about the role of inter-coder agreement in reflexive qualitative analysis, collaborative coding can still enhance interpretive depth when used to support, rather than constrain, meaning-making (Braun and Clarke, 2022; Hennink and Kaiser, 2022). Discrepancies in coding can be treated as analytically informative and resolved through reflexive discussion, iterative recoding, or inter-coder reliability measures such as Cohen's kappa Cohen's kappa, Krippendorff's alpha Krippendorff's alpha, or Fleiss' kappa Fleiss' kappa (Bedemariam et al., 2025).

Once the RTA is completed, researchers should conduct a parallel GenAI analysis (Step 1b) by providing the same reference text to one or more GenAI models. The LLM-GenAI Theme Model should generate themes inductively (no influencing codes or categories) or deductively (codes or categories prompted to a model) from the data. We acknowledge that prompts may vary depending on the research focus, design, data, and model-used, prompts should clearly guide the model to identify themes and associated concepts relevant to the analytical intent. To support structured comparison with the human-generated themes researchers are encouraged to design outputs in a standardized format, such as JSON. Structured outputs facilitate consistent organization of themes and concepts and simplify alignment with the human-derived themes during the comparative analysis conducted in Step 3.

3.4.2 Step 2. Conduct an Inductive Content Analysis (ICA)

After establishing the RTA interpretative baseline, researchers apply Inductive Content Analysis (ICA) to both the human-derived themes and the LLM-generated outputs. Following Zhang and Wildemuth (2009)'s eight-step qualitative content analy-

sis framework, the process proceeds through data preparation, defining units of analysis, category development, testing the coding scheme, coding the data, conducting consistency checks, analyzing results, and reporting findings. This structured sequence ensures that concept identification and meaning categorization are systematic, transparent, and comparable across datasets.

Applied first to the RTA-derived themes (Step 2a), ICA systematically identifies and organises the concepts embedded within the human-interpreted thematic structure. The same procedure is then applied independently to the LLM-generated outputs (Step 2b), without importing or constraining categories from the human analysis. This preserves analytical independence and allows model-generated meaning to emerge inductively, capturing which concepts are foregrounded, reconfigured, or omitted, as well as how semantic variation is handled. Conducting ICA in parallel across both datasets establishes a structurally aligned comparative framework, enabling systematic evaluation of conceptual convergence, partial overlap, and divergence between human-interpreted meaning and GenAI outputs.

The results of the ICA analysis are two reports, one derived from the RTA-based human themes and one from the LLM-generated outputs. The first report organizes meaning from the RTA into inductively generated categories that reflect human interpretation of concepts, relationships, and contextual nuance. The report applies the same ICA procedure to LLM outputs, producing a comparable set of categories that reflect how the model structures and represents meaning without reference to the human-derived framework. Together, these documents form a paired conceptual mapping of meaning. Because both are generated using the same inductive content analysis framework, they enable direct comparison of category structures and underlying concepts, allowing systematic identification of convergence, partial overlap, and divergence between human-interpreted meaning and GenAI-generated representations.

3.4.3 Step 3. Compare GenAI and RTA Outputs via ICA

This stage represents the comparative phase of the ICA process, where independently constructed meaning systems from the RTA-derived human themes and the LLM-generated outputs are brought into direct analytic alignment. After completing ICA for both datasets (Step 2a and Step 2b), the resulting category structures function as mapping systems that translate interpretive meaning into comparable conceptual units. ICA thus acts as a bridge between RTA-generated themes and LLM outputs by converting both into standardized inductive category frameworks. Each RTA theme is operationalised as a cluster of concepts, and LLM outputs are independently organised using the same coding protocol, creating a shared unit of comparison based on concept presence, structure, and relational grouping rather than surface-level language similarity.

The comparison then evaluates how these mapped concept structures align across datasets. Researchers assess whether concepts within an RTA-derived theme are preserved, redistributed, fragmented, or absent in the LLM category system. This makes it possible to trace not only whether concepts appear in both datasets, but also whether their thematic organisation is maintained or reconfigured in the model output. Practically, this enables analysis at three levels: (1) concept-level correspondence (presence/absence), (2) intra-theme structure (how concepts cluster within themes), and (3) inter-theme coherence (how relationships between themes are preserved or altered). For example, a human theme such as "Climate Change Consequences" may map onto a coherent LLM cluster, be distributed across multiple model categories, or be restructured into a different conceptual grouping. By aligning two independently derived inductive category systems, ICA transforms RTA-based interpretive meaning and LLM outputs into directly comparable conceptual architectures, making convergence, divergence, and transformation in meaning systematically traceable.

3.4.4 Step 4. Calculate the Inductive Conceptual Rating Score (ICR)

The ICR calculation emerges directly from the outputs of the ICA comparison stage, where both the Human Theme Model (derived from RTA) and the GenAI Theme Model have already been transformed into structured concept inventories.

After ICA, each dataset is represented as a standardized set of coded concepts organised into categories. These two concept sets form the basis for alignment: researchers first define a shared comparison space by taking the union of all unique concepts identified across both the human and model outputs. This ensures that both presence and ab-

sence can be evaluated consistently within a fixed conceptual universe.

Within this shared space, each concept is evaluated for its status in both datasets:

- If a concept appears in both the RTA-derived and LLM-derived ICA outputs, it is coded as a True Positive (TP).
- If a concept appears only in the RTA baseline but not in the LLM output, it is a False Negative (FN).
- If a concept appears only in the LLM output, it is a False Positive (FP).
- Concepts absent from both are treated as True Negatives (TN), relative to the defined conceptual universe.

This mapping step operationalises the ICA outputs into quantitative variables. In other words, ICA produces the *structured conceptual inventories*, and the ICR step translates these inventories into set-based alignment metrics by treating them as comparable categorical spaces. The accuracy formula is then applied to these aligned concept sets, producing a single score that reflects how faithfully the GenAI output reproduces the human-interpreted meaning structure.

In this way, the quantitative ICR score is not independent of the qualitative analysis but is directly derived from the ICA-generated concept mappings, ensuring continuity between interpretive thematic construction and formalised measurement of semantic alignment.

3.5 Positioning and Practical Implications

The ICR metric is designed to complement, rather than replace, automated evaluation methods, particularly in tasks such as summarization and thematic synthesis where semantic fidelity is central. Unlike automated metrics that often assume meaning is fixed and directly comparable, ICR treats meaning as dynamic, relational, and plural, requiring interpretive expertise grounded in hermeneutic and qualitative traditions. By combining human interpretation with a structured analytical framework, ICR makes visible how meaning is constructed, negotiated, and transformed across datasets, while also providing practical value for model evaluation, responsible deployment, and diagnostic insight. Importantly, the human baseline used in ICR is itself situated, reflecting the cultural, disciplinary, and experiential contexts of its coders. As a result, ICR scores measure alignment with a specific interpretive frame rather than with universal or context-independent meaning. This is also a strength, as it makes the evaluative baseline explicit and contestable, in contrast to automated metrics that often obscure their underlying assumptions. Researchers applying ICR in legal, clinical, or culturally specific contexts should therefore ensure that coders are appropriately selected to match the interpretive demands of the domain under analysis.

4 Case-Study: LLM Evaluation using the ICR Metric

This case-study used a mixed-methods triangulation design to assess the accuracy of LLM-generated thematic summaries (Creswell and Creswell, 2017). Two data types were analyzed: qualitative open-text comments and quantitative ratings. The quantitative analysis used post hoc survey data to measure textual similarity between reference texts and LLM-generated summaries, followed by descriptive statistics and correlation tests. The qualitative analysis applied Reflective Thematic Analysis (RTA) to establish human-interpreted outputs, then Inductive Content Analysis (ICA) to compare concepts and meanings across reference texts, RTA outputs, and LLM outputs. RTA was conducted before LLM analysis, and the Inductive Conceptual Rating (ICR) integrated surface-level (linguistic) and deeper-level (semantic) comparisons (Fitkov-Norris et al., 2023; Guetterman et al., 2018).

4.0.1 Dataset

Five datasets ranging from N=50 to N=800 of unstructrued text responses from a survey asking about perceptions of an organization were analyzed. Responses ranged from 1 to 300 words. Preprocessing removed personally identifiable information and low-quality responses (e.g., “none,” “N/A”). Both human analysts and LLMs produced three themes per dataset, each with a short name (<5 words) and a 3-4 sentence summary. LLMs were prompted to output this standardized format in JSON for direct comparison.

4.0.2 Establishing an Interpretative Baseline

An interpreative, human baseline was established by conducting a Reflective Thematic Analysis (RTA) on each dataset, using Braun and Clarke’s

six-step method (Braun and Clarke, 2006). Multiple raters ensured reliability, and concepts were coded as short phrases (<5 words) and categorized inductively using Inductive Content Analysis (ICA) (Zhang and Wildemuth, 2009). Iterative coding resolved discrepancies, and the presence or absence of each concept was documented for comparison across outputs.

4.0.3 LLM Selection and Implementation

Sonnet 3.5 and Nova Pro were selected for their differences in architecture, training, and representational capacity (Jacas et al., 2025; Bedemariam et al., 2025). Sonnet 3.5 emphasizes instruction-following and structured reasoning, while Nova Pro exhibits less abstraction and contextual integration (Bedemariam et al., 2025).

4.0.4 Prompt Engineering

Prompts were iteratively tested using zero-shot, few-shot, and chain-of-thought strategies, with optimization targeting output consistency and thematic specificity—namely, whether the model reliably generated three distinct, non-overlapping themes with conceptually grounded summaries in the required JSON format (He et al., 2024). Prompts that produced structurally inconsistent outputs, overly generic themes, or conceptually conflated categories were revised until stable performance was achieved. Final prompts were model-specific but structurally standardised, requiring three themes with 3–4 sentence summaries in JSON format (see Appendix A). Each model was then run ten times under standardised decoding parameters (Temperature = 0, top-k = 0.25, top-p = 0.99), producing 50 outputs per model. To ensure stability and reduce stochastic variation, the modal (most frequently occurring) output per model and dataset was selected as the final artifact for downstream evaluation.

4.0.5 Evaluation Metrics

Outputs were evaluated on two dimensions: linguistic similarity (cosine similarity, BERTScore) and semantic accuracy (ICR, measuring agreement in concept presence between RTA and GenAI outputs).

4.0.6 Results

The results show variability across models and datasets in similarity, accuracy, and reliability. Key findings include:

- **N = 50:** Linguistic metrics were identical across outputs, but semantic accuracy diverged. Human RTA achieved perfect ICR (1.00), while Sonnet 3.5 and Nova Pro scored 0.69, missing or distorting core meanings in the small dataset.
- **N = 100:** Sonnet 3.5 had high linguistic performance (Cosine = 0.89, F1 = 0.91) but low semantic alignment (ICR = 0.35). Nova Pro had slightly lower linguistic scores but higher ICR (0.48). Human RTA balanced both metrics (ICR = 0.86).
- **N = 200:** Linguistic metrics remained high across outputs. Nova Pro achieved higher semantic accuracy (ICR = 0.75) than Sonnet 3.5 (0.39), while RTA retained near-complete coverage (ICR = 0.96).
- **N = 400:** LLMs maintained high linguistic similarity (Sonnet F1 = 0.90), but semantic accuracy was moderate (Sonnet 0.47; Nova 0.53). RTA showed slight drop in Recall but high semantic fidelity (ICR = 0.94).
- **N = 800:** Both models achieved their best performance with high linguistic similarity and improved ICR (Sonnet 0.65; Nova 0.76), yet RTA remained superior (ICR = 0.93), indicating larger datasets enhance LLM semantic alignment but do not fully match human interpretive analysis.

5 Discussion

This case study compared GenAI outputs to human-generated analyses of unstructured data using the ICR metric to quantify semantic fidelity. Across datasets, LLMs achieved high surface-level performance (cosine similarity, F1) but consistently lower semantic alignment, highlighting that LLMs replicate lexical patterns and topical overlap effectively but struggle with meaning, which requires relational reasoning, contextual grounding, and recognition of fluid conceptual structures. These results reinforce the semiotic distinction between signifiers (words) and signifieds (meaning).

Dataset size influenced semantic stability: while surface-level metrics remained consistent, ICR scores fluctuated with smaller datasets and stabilized only above 200 responses. Even at N = 800, LLMs did not reach human levels of interpretive coherence, suggesting that more data improves

Table 1: Evaluation metrics (Cosine similarity, Recall, Precision, F1, ICR) for human and LLM outputs across datasets

Dataset	Output	Cos	Rec	Prec	F1	ICR
N = 50	Human	0.68	0.60	1.00	0.75	1.00
	Sonnet 3.5	0.79	0.60	1.00	0.75	0.69
	Nova Pro	0.73	0.60	1.00	0.75	0.69
N = 100	Human	0.72	0.60	1.00	0.75	0.86
	Sonnet 3.5	0.89	0.89	0.93	0.91	0.35
	Nova Pro	0.64	0.83	0.87	0.84	0.48
N = 200	Human	0.72	0.65	0.86	0.72	0.96
	Sonnet 3.5	0.72	0.90	0.90	0.90	0.39
	Nova Pro	0.71	0.86	0.89	0.87	0.43
N = 400	Human	0.69	0.40	1.00	0.57	0.94
	Sonnet 3.5	0.78	0.89	0.92	0.90	0.47
	Nova Pro	0.68	0.83	0.83	0.83	0.53
N = 800	Human	0.75	0.60	1.00	0.75	0.93
	Sonnet 3.5	0.87	0.87	0.93	0.90	0.65
	Nova Pro	0.87	0.80	0.90	0.86	0.76

but does not guarantee semantic fidelity. Humans maintained high ICR scores across all dataset sizes, demonstrating robust interpretive reasoning.

Model-level analysis revealed that architectural sophistication or recency does not necessarily ensure better semantic performance. LLMs achieved high linguistic metrics yet showed substantial variability in synthesizing complex, multifaceted human meaning, emphasizing that surface fluency does not equate to semantic accuracy.

Finally, the findings have epistemological implications. Unlike human analysts, probabilistic LLMs lack reflexivity, historical awareness, and contextual understanding, transforming sign systems without situational sensitivity. The empirical ICR gap (human: 0.93; LLMs: 0.35-0.76) quantifies this deficit, supporting the view that GenAI simulates rather than generates meaning. This distinction is crucial when applying generative AI to interpretive work, where content similarity alone may mask deeper semantic misalignment.

5.1 Implications

This case study suggests several implications for theory, methodology, and practice:

- **Lexical vs. semantic alignment:** The results revealed that the LLMs, in this study, excelled at lexical content matching but struggled with deeper meaning. It also showed that lexical metrics may capture surface similarity but miss interpretive depth, highlighting the value of metrics like ICR for assessing semantic fidelity.

- **Dataset size and interpretive stability:** Semantic accuracy improves with larger datasets, but higher volume does not guarantee human-level interpretation. Human evaluation remains essential, particularly for small or conceptually nuanced datasets.

- **Epistemological considerations:** LLMs simulate meaning without reflexivity or contextual awareness. The human–AI semantic gap underscores that GenAI outputs are best used for pattern detection or preliminary synthesis, not as sources of understanding.

- **Interpretative evaluative approach:** Combining LLM outputs with interpretative baselines enables pattern detection while preserving interpretive rigor. ICR provides a transparent framework for evaluating meaning, but

requires researchers skilled in qualitative and interpretive methods.

Overall, the LLMs studied produce strong surface-level outputs but struggle with context-dependent meaning. Metrics like ICR help quantify semantic fidelity, revealing the degree to which outputs are truthful or potentially misleading.

6 Conclusion

This study demonstrates that while LLMs can reproduce surface-level lexical patterns, they often lack semantic depth and relational coherence. The ICR metric makes this distinction measurable by differentiating between surface-level linguistic simulation and deeper conceptual alignment, while also recognising that the evaluative baseline itself is a human construct, situated, plural, and interpretively contingent.

The findings reinforce the semiotic distinction between signifiers and signifieds, emphasizing the contextual, reflexive, and situated nature of meaning. Probabilistic LLM outputs simulate interpretive processes but lack the historical, cultural, and experiential grounding humans provide. The ICR gap highlights these epistemic limitations, offering a framework to assess interpretive reliability.

Meaning remains inherently multidimensional and plural, and LLM outputs may be locally plausible while still distorting or oversimplifying underlying concepts in ways that surface-level metrics miss. Without a human interpretive baseline, such outputs cannot be assumed to reliably reflect conceptual truth. We advocate for an interpretative, human-centered evaluation approach in which researchers provide contextual interpretation and ethical oversight. Metrics like ICR help distinguish between lexical imitation and genuine conceptual alignment, ensuring that GenAI systems support rather than obscure human meaning-making.

Limitations

This study has several limitations. First, it focuses on open-text survey responses to a single question, which constrains transferability across domains, text types, and NLG tasks. Future work should extend ICR to broader settings such as multi-document summarization, educational assessment, and clinical text synthesis to better characterize its performance boundaries. In addition, the interpretative baseline reflects situated human interpretation, which encodes the cultural, disciplinary, and contextual assumptions of the coders who construct it, resulting in an interpretative metric. In addition, the base ICR metric captures conceptual alignment through presence–absence structure, but does not include results such as discourse-level meaning. Although ICR is domain-agnostic in principle, its validity depends on the quality of the human-coded baseline and the expertise of coders. Applications in legal, clinical, or culturally specific domains require carefully selected coders and transparent attention to positionality to ensure interpretive and ethical robustness.

In addition, the evaluation was limited to two LLMs under specific prompting strategies and parameter settings, meaning results may not generalize to other models or configurations. Given the rapid evolution of GenAI, these findings should be interpreted as illustrative rather than definitive.

References

M. Alizadeh and A. Seilsepour. 2025. A novel self-supervised sentiment classification approach using semantic labeling based on contextual embeddings. *Multimedia Tools and Applications*, 84(12):10195–10220.

M. Apidianaki. 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523.

A. Arseniev-Koehler. 2024. Theoretical foundations and limits of word embeddings: what types of meaning can they capture? *Sociological Methods & Research*, 53(4):1753–1793.

R. Bedemariam, N. Perez, S. Bhaduri, S. Kapoor, A. Gil, E. Conjar, and N. Nayyar. 2025. Potential and perils of large language models as judges of unstructured textual data. *arXiv preprint arXiv:2501.08167*.

S. Bhaduri, S. Kapoor, A. Gil, A. Mittal, and R. Mulkar. 2024. Reconciling methodological paradigms: Employing large language models as novice qualitative research assistants in talent management research. *arXiv preprint*, abs/2408.11043.

A. D. Bhargavi. 2025. Comparative study of static and contextual text vectorization for sentiment analysis. *International Journal for Research in Applied Science & Engineering Technology*.

M. Borgstede and M. Scholz. 2021. Quantitative and qualitative approaches to generalization and replication – a representational perspective. *Frontiers in Psychology*, 12:605823.

V. Braun and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

V. Braun and V. Clarke. 2022. Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9(1):3.

J. W. Creswell and J. D. Creswell. 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.

J. W. Creswell and C. N. Poth. 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications.

E. Fitkov-Norris, N. Kocheva, and N. Bhimireddy. 2023. Are we there yet? thematic analysis, nlp and machine learning for academic research: a comparative review. In *FBSS Research Conference 2023*.

H. G. Gadamer. 1975. Hermeneutics and social science. *Cultural Hermeneutics*, 2(4):307–316.

R. Gangundi and R. Sridhar. 2025. Rbca-ets: enhancing extractive text summarization with contextual embedding and word-level attention. *International Journal of Information Technology*, 17(2):1127–1135.

J. Grindrod. 2024. Transformers, contextualism, and polysemy. *arXiv preprint arXiv:2404.09577*.

T. C. Guetterman, T. Chang, M. DeJonckheere, T. Basu, E. Scruggs, and V. V. Vydiswaran. 2018. Augmenting qualitative text analysis with natural language processing: methodological study. *Journal of Medical Internet Research*, 20(6):e231.

J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan. 2024. Does prompt formatting have any impact on LLM performance? *arXiv preprint arXiv:2411.10541*.

M. Hennink and B. N. Kaiser. 2022. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science and Medicine*, 292:114523.

J. Jacas, H. Winchester, A. E. Boyd, and B. Johnson. 2025. Architecture matters: Understanding how LLM model types influence harmful language detection in technical contexts. In *Proceedings of the 1st International Workshop on Responsible Software Engineering*, pages 7–12.

B. Paneru, B. Thapa, and B. Paneru. 2025. Sentiment analysis of movie reviews: a flask application using cnn with roberta embeddings. *Systems and Soft Computing*, 7:200192.

N. Perez, S. Bhaduri, and A. Chadha. 2026. Simulating meaning, nevermore! introducing ICR: A semiotic-hermeneutic metric for evaluating meaning in LLM text summaries. *arXiv preprint arXiv:2603.04413*.

E. A. Poe. 2013. *The Raven: Tales and Poems*. Penguin.

E. N. Purba and B. A. R. Damanik. 2025. Polysemy and homonymy in semantic interpretation. *Young Journal of Social Sciences and Humanities*, 1(3):101–109.

F. De Saussure. 1916. *Nature of the linguistic sign*. Bedford/St. Martin's Press.

S. Shan. 2025. Towards reflexive ai: A comprehensive exploration of enhancing social science research through nlp. In *Future of Information and Communication Conference*, pages 765–792. Springer Nature Switzerland.

F. Viegas, A. Pereira, W. Cunha, C. França, C. Andrade, E. Tuler, and M. A. Gonçalves. 2025. Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics. *Computational Linguistics*, pages 1–41.

Y. Zhang and B. M. Wildemuth. 2009. Qualitative analysis of content. In *Applications of social research methods to questions in information and library science*, volume 308(319), pages 1–12. Libraries Unltd Inc.

A Appendix: Example of GenAI Thematic Outputs in JSON

Below is an example of GenAI-generated thematic outputs formatted in JSON for consistency and comparability with human-generated themes:


```
{
  • "themes": [
      – { "name": "Climate Change Impacts",
        "summary": "Climate change is driving
        environmental shifts such as rising
        sea levels, extreme weather, and
        biodiversity loss." },
      – { "name": "Mitigation and Renewable
        Energy",
        "summary": "Mitigation focuses on
        reducing emissions through renewable
        energy, efficiency, and sustainable
        practices." },
      – { "name": "Policy and Global
        Cooperation",
        "summary": "Addressing climate change
        requires coordinated policy actions,
        global agreements, and regulatory
        frameworks." }
    ]
}
```


B Appendix: RTA Procedural Example via Climate Change Perceptions

The study followed a structured Reflective Thematic Analysis (RTA) process:

1. Data Familiarization Researchers reviewed 50 participant responses multiple times to immerse

themselves in the data. Initial observations highlighted recurring concepts such as climate impacts, policy concerns, and mitigation strategies.

2. Initial Coding Open coding was applied to meaningful text segments. Example codes included: *sea-level rise*, *renewable energy adoption*, *extreme weather events*, *global cooperation*. Multiple researchers coded independently to capture diverse interpretations.

3. Theme Identification Codes were grouped into broader themes:

- Climate Change Impacts
- Mitigation and Renewable Energy
- Policy and Global Cooperation
- Societal and Economic Consequences

4. Thematic Review Themes were refined for coherence and clarity. Overlapping concepts (e.g., types of environmental impacts) were separated to improve nuance and interpretability.

5. Defining Themes Each theme was clearly defined:

- *Climate Change Impacts*: Observed and projected environmental effects such as rising sea levels and extreme weather
- *Mitigation and Renewable Energy*: Actions to reduce emissions and promote sustainable energy
- *Policy and Global Cooperation*: Regulatory frameworks, international agreements, and policy measures
- *Societal and Economic Consequences*: Effects on communities, economies, and livelihoods

6. Generating Insights Relationships among themes were identified. For example, respondents associated stronger mitigation policies with reduced societal risk and perceived resilience, highlighting the interplay between structural actions and environmental outcomes.

Inter-Rater Reliability High agreement was observed (Cohen's $\kappa = 0.87$; Fleiss' $\kappa = 0.85$), supporting consistency in coding.

Interpretive Baseline The final thematic structure serves as the human-interpreted baseline, capturing key concepts, relationships, and contextual nuance. This baseline is used for subsequent ICR evaluation of GenAI outputs.

Appendix 4: ICA Output Example – Climate Change

A structured Inductive Content Analysis (ICA) was applied to GenAI outputs generated from the prompt: *"Summarize the experiences of 50 participants discussing climate change."*

1. Data Preparation Outputs were cleaned, segmented into sentences/phrases, and prepared for coding.

2. Unit of Analysis Each sentence or meaningful phrase was treated as a unit to capture distinct concepts. A comprehensive table of concepts and definitions was created.

3. Category Development Using an inductive approach (Zhang and Wildemuth, 2009), emergent categories included:

- Environmental Impacts
- Renewable Energy and Mitigation
- Policy and Governance
- Societal and Economic Effects
- Public Awareness and Engagement

4. Coding and Refinement Researchers iteratively coded samples, refined categories, and applied the final scheme to all outputs.

5. Consistency Check Inter-coder reliability was high (Cohen's $\kappa = 0.81$).

6. Comparison with RTA Baseline ICA categories were compared to RTA themes. GenAI outputs often:

- Captured broad themes (e.g., environmental impacts, policy)
- Underrepresented nuanced consequences (e.g., economic disparities)
- Overgeneralized mitigation strategies

7. ICR Calculation Conceptual alignment was classified as:

- TP: Correctly captured concepts
- FP: Unsupported additions
- FN: Missing concepts
- TN: Correct exclusions

Example:

- TP = 5, FP = 1, FN = 1, TN = 0

Using the Accuracy metric, the results show a score of ≈ 0.71

This score reflects moderate-to-strong alignment between the GenAI interpretation and the human baseline, indicating most key themes were captured while minor omissions and unsupported additions occurred. Higher ICR values indicate stronger semantic fidelity.