ARIC: A Cognitive Framework for Explanatory Narrative Evaluation in Conversational Information Seeking Systems

Vahid Sadiri Javadi
University of Bonn
Bonn, Germany
vahid.sadirijavadi@uni-bonn.de

Sadia Naseer
University of Bonn
Bonn, NRW, Germany
sadia.naseer@uni-bonn.de

Ali Ather
University of Bonn
Bonn, NRW, Germany
ali.ather@uni-bonn.de

Lucie Flek
University of Bonn
Bonn, NRW, Germany
flek@bit.uni-bonn.de

Johanne Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Abstract

Conversational information seeking (CIS) systems now generate explanations, but we still evaluate them using retrieval-focused metrics such as faithfulness, completeness, and source attribution. These checks are necessary, but they do not tell us whether a response helps users form a coherent mental model. Cognitive science treats *understanding* as an active construction guided by causal structure, coherence, and the organization of information. Evaluation should therefore test whether explanations support integration, inference, and retention. This paper aims to define the target form of communication and a way to assess it. We introduce **Explanatory Narratives** for CIS, which combine the organizing benefits of storytelling to enable explanation. We then propose **ARIC**, a cognitively grounded framework for evaluating explanatory narratives across four comprehension stages: **A**ttention, **R**epresentation, **I**ntegration, and **C**onsolidation. To demonstrate its value, we apply ARIC to human-authored explanatory narratives and show how stage-based analysis yields actionable diagnostic insights. This shifts CIS evaluation toward the question the IR community increasingly faces: whether system-generated explanations actually help users understand.

CCS Concepts

• **Human-centered computing** → *HCI design and evaluation.*

Keywords

Evaluation, Narrative, Human Cognition, Information Seeking

ACM Reference Format:
Vahid Sadiri Javadi, Sadia Naseer, Ali Ather, Lucie Flek, and Johanne Trippas. 2026. ARIC: A Cognitive Framework for Explanatory Narrative Evaluation in Conversational Information Seeking Systems. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26), July 25, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805713.3820401

ICTIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2600-2/2026/07
https://doi.org/10.1145/3805713.3820401

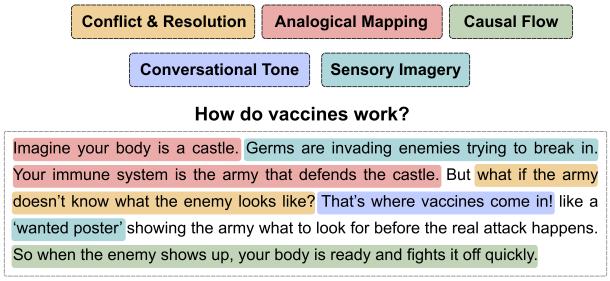

Figure 1: Decomposing an explanatory narrative into its constituent characteristics. Unlike direct factual responses, explanatory narratives employ analogical mapping, conflict and resolution, causal flow, sensory imagery, and conversational tone to align with human cognitive processing.

1 Introduction

The shift from ranked document lists to generated textual responses represents a transformation in information retrieval (IR) [53]. Conversational information seeking (CIS) systems no longer simply retrieve relevant documents. Instead, they synthesize, explain, and narrate information [71, 79]. When a user asks a CIS system *"How do vaccines work?"* or *"Why does inflation affect housing markets?"*, the system must generate a response that contains relevant information and *supports the user's understanding*. This paradigm shift, from retrieval to generation, carries implications for evaluation. The IR community has responded with nugget-based evaluation [1, 61], multi-dimensional utility frameworks [26], and LLM-as-judge approaches [6, 44, 62]. Yet these frameworks assess if responses are *faithful*, *complete*, and *well-attributed*, not whether they help users *understand* the information. This gap is where cognitive IR theory has long warned us to look: at the intersection of cognitive relevance [66], sense-making [19], and knowledge construction [41].

Parallel to this evaluation challenge, research across cognitive science and educational psychology shows that *how* information is structured shapes whether it is understood. Capturing the idea that narrative (i.e., the organization of information into temporally and causally structured sequences) is a cognitive tool and mechanism through which humans process information, attribute causes, and construct mental models of complex situations [10, 29, 40, 67]. Narratives improve comprehension, memory, and reasoning compared

to expository text across ages, reading abilities, and domains [27, 56]. Similar effects appear in CIS, where narrative framing increases engagement with complex explanations [36, 65]. This suggests that when CIS systems generate explanatory responses, the narrative structure of those responses is not a stylistic choice but a cognitive one, directly affecting if users achieve genuine understanding.

The IR evaluation gap and the cognitive power of narrative converge in what we term an **Explanatory Narrative**: a cognitively hybrid form of communication that integrates the structural strengths of storytelling (e.g., concreteness, imagery, causal sequencing) with the inferential power of explanation (e.g., causality, generalization, mechanism identification). Explanatory narratives are designed to recount events (i.e., first X happened, then Y) and to clarify *"why"* and *"how"* things happen. As shown in Figure 1, instead of presenting an abstract immunological description, the explanatory narrative uses a familiar analogy (i.e., your body as a castle) to ground the concept in concrete, relatable imagery and to introduce a conflict and resolution in which the immune system acts as a defending army. Explanatory narratives are ubiquitous in CIS: they are what systems produce (or should produce) when users pose "why" and "how" queries, i.e., non-factoid questions, which are challenging and poorly represented in existing evaluation benchmarks [9].

CIS evaluation still mainly focuses on relevance, faithfulness, and completeness rather than on whether answers actually support comprehension. Benchmarks such as TREC CAsT [17, 57], iKAT [2], and related RAG tracks largely ignore cognitive relevance, meaning how information changes a user's knowledge state [66]. As a result, automatic metrics do not predict human preferences for long-form answers [77], and they can even score incorrect conversational answers higher than correct ones [22]. Holistic ratings, such as coherence and engagement, and surface similarity metrics, such as BLEU, ROUGE, and BERTScore, collapse cognition into a single number and measure text properties rather than understanding. The result is a blunt evaluation that tells us whether something is good or bad, but not why, even though comprehension unfolds in stages and an explanation can succeed in attention but fail in integration or accuracy.

Information Processing Theory (IPT) models comprehension as a sequence of cognitive transformations in which information is attended to, encoded in working memory, integrated with prior knowledge, and stored in long-term memory [5]. In parallel, Mayer's **Cognitive Theory of Multimedia Learning (CTML)** describes meaningful learning as an active process in which learners select relevant information, organize it into coherent verbal and pictorial mental models, and integrate these with existing schemas [48].

The implication is direct: understanding is built, stage by stage, and it can fail in specific places. Evaluation of explanatory narratives, whether written by humans or generated by CIS systems, should measure support for these stages. Metrics that only check accuracy, similarity, or overall quality miss the cognitive mechanisms that determine whether users actually understand.

We introduce **ARIC**, a cognitively grounded four-stage evaluation framework: **Attention**, **Representation**, **Integration**, and **Consolidation**. We adapt this structure to evaluate CIS explanations. ARIC operationalizes each stage by identifying where a narrative supports or hinders understanding. Instead of a single overall score, ARIC identifies which stage succeeds or fails, providing a diagnostic analysis for CIS [26, 77].

This paper has four main contributions:

(1) We conceptualize **Explanatory Narratives** as a cognitively hybrid form of communication that integrates the strengths of storytelling with explanation.
(2) We introduce **ARIC**, a cognitively grounded evaluation framework for explanatory narratives in CIS. ARIC transfers established stage models of comprehension from cognitive theory into the CIS evaluation setting, a novel approach in a field dominated by retrieval and text-similarity metrics [5, 48]. It structures evaluation around four stages, **Attention**, **Representation**, **Integration**, and **Consolidation**, and **specifies fine-grained diagnostic criteria** to assess how well a narrative supports each stage.
(3) We demonstrate **ARIC's Utility** through human-authored explanatory narratives (TED-Ed Lessons[1]) and preliminary evaluations of LLM-generated responses to QA queries. We show that ARIC provides diagnostic insights into narrative effectiveness, identifying where narratives succeed or fall short. The code, dataset, annotation guidelines, and results of both studies are available.[2]
(4) By grounding explanatory narrative evaluation in cognitive mechanisms of comprehension, ARIC links *(1)* the IR community's recognition that generative systems require fundamentally new evaluation paradigms, and *(2)* cognitive science's understanding of the staged processes through which humans construct meaning.

2 Conceptualizing Explanatory Narratives

2.1 Narrative and Human Cognition

Narrative is a fundamental mode of human thought. Cognitive psychologist Jerome Bruner argues that cognition operates in two modes [11]. The paradigmatic mode supports logical and scientific reasoning through abstraction and categorization. The narrative mode organizes experience into sequences in terms of temporal order, causality, and human intentions. The paradigmatic mode aims for general truth conditions, whereas the narrative mode aims for presenting coherent, situated accounts that feel true. Bruner emphasizes that narrative thinking is not secondary to logic, but a primary means by which humans construct reality and meaning [11].

Narrative aligns well with how people process information. First, narratives are **temporally structured**. Events unfold in time, and narrative respects this unfolding, making information easier to follow and remember [7, 10, 34, 35, 39, 70]. Second, narratives are **causally organized**. They do not just list events but connect them through causal and intentional relations (e.g., this happened *because of that*). This causal structure aligns with how humans seek to understand the world by identifying causes, attributing intentions, and predicting consequences [37, 39, 46, 70, 76]. Third, narratives are **concrete and imagistic**. They ground abstract ideas in specific

[1] https://ed.ted.com/lessons
[2] https://github.com/vahidsj/ARIC

characters, settings, and actions, engaging perceptual simulations in the brain. Neuroimaging research suggests that reading narratives activates language areas and sensory regions, as if people are mentally simulating the described events [47, 58, 69]. Fourth, narratives are **perspectival**. They are told from a point of view, inviting users to adopt the goals, knowledge, and emotions of characters. This perspectival quality supports social cognition and empathy, allowing us to understand situations from multiple points [30, 38, 47, 64]. Empirical studies demonstrate that narratives are better comprehended and remembered than expository texts, with effects robust across age groups, reading abilities, and content domains [13, 16, 27]. The narrative advantage reflects a deep alignment between narrative structure and cognitive architecture.

This alignment has implications beyond literary or pedagogical contexts. When CIS systems generate responses to complex queries, they face a choice between expository text (i.e., presenting facts, definitions, and relationships in abstract, decontextualized form) and narratively structured text (i.e., embedding the same information in temporally sequenced, causally connected, concrete scenarios). The cognitive science reviewed above suggests that this choice is stylistic and consequential for whether users actually comprehend and retain the information provided. A system that structures its explanation as a narrative aligns its output with its user's cognitive architecture.

2.2 Explanation and Narrative

Explanation and narrative solve the same problem. They make events intelligible by organizing causes and consequences into a coherent sequence. Explanation answers *"why"* and *"how"* questions by linking outcomes to causes, mechanisms, and reasons [37, 46, 70, 76]. Narrative provides a practical structure for communicating those links in a form people can follow. This matters for CIS systems, since these are the query types that these systems are expected to handle [2, 17, 19, 66, 72, 79], and for which current evaluation methodology is least adequate [77]. Thus, understanding the connection between explanation and narrative is essential for generating and evaluating system responses to such queries.

Good explanations have several characteristics. First, they make the relevant causal structure explicit. They go beyond correlation and state the mechanisms by which causes produce effects [46, 76]. Narrative supports this by linking events with causal relations rather than just in chronological order [70]. Second, they are audience-directed. Effective explanations account for what the recipient already knows, what must be made explicit, and what can be safely left implicit [19, 37]. Skilled narratives do the same by adjusting detail and pacing, aligning with cognitive IR accounts that stress adapting information presentation to the user's cognitive state [33, 41]. Third, they ground abstractions in concrete analogies. Mapping an unfamiliar concept onto a familiar domain makes principles easier to grasp and remember, for example, describing a cell as a factory [24, 32].

Thus, explanation and narrative are closely linked, compatible, and cognitively synergistic. Explanations are constructed for an audience and must compress causal understanding into a form that is communicable [37, 46]. Narrative structure supports this by unfolding over time, highlighting causality, grounding abstractions in concrete imagery, and adopting an audience-appropriate perspective [10, 47, 58, 70]. Narrative provides the structural scaffolding, temporal sequencing, causal chaining, concrete grounding, and perspectival framing, that explanation requires to be effective [39, 76]. Explanation, in turn, gives narrative epistemic purpose: it transforms a sequence of events into a vehicle for understanding *why* and *how* things work [46]. When these two modes merge it is a communicative form, which we call **Explanatory Narrative**.

2.3 The Explanatory Narrative

We define an *Explanatory Narrative* as a narrative designed to explain. It uses storytelling structure such as temporal order, concrete imagery, and causal progression to present an explicit causal account, including mechanisms and general principles, so the audience can understand why and how something happens, not just what happened. We distinguish explanatory narratives from two adjacent forms: *(1)* **Pure explanations** convey causal or mechanistic information without narrative framing; *(2)* **Pure narratives** recount sequences of events without necessarily clarifying underlying mechanisms or principles. A chronological news report of a natural disaster, for instance, may describe what happened and when, without explaining the geological or meteorological processes that caused it. A textbook definition of the immune response, for example, may specify the roles of antigens and antibodies in abstract, decontextualized terms without embedding them in a concrete scenario. An explanatory narrative differs from these two by simultaneously narrating *and* explaining, i.e., it unfolds events in time while making the underlying causal structure visible, and it grounds abstract mechanisms in concrete imagery.

This distinction is directly relevant to CIS systems' goal. When a user asks a system *"How do vaccines work?"*, the system might respond with a pure explanation (i.e., a decontextualized description of antigen presentation) or a pure narrative (i.e., a chronological account of the vaccine's development without mechanistic detail), or an explanatory narrative that weaves the mechanism into a concrete, causally structured scenario.

This hybrid form matters because it changes how people process the explanation, not just how it sounds. In the vaccine example in Figure 1, the castle analogy activates familiar knowledge of enemies, defences, and training, providing users with a structure for learning new concepts. The narrative frame sets up a causal chain so each event explains the next, which supports coherence and inference. The concrete imagery of soldiers, walls, and scouts also prompts mental simulation, leading to richer encoding and better recall than abstract terms alone. And the conflict-resolution arc (i.e., threat → defense → preparedness) generates suspense and resolution, sustaining attention throughout the explanation. This passage works because it combines multiple cognitive processes.

This observation carries a critical implication for evaluation. If the effectiveness of explanatory narratives derives from their ability to engage a *sequence* of cognitive processes, then evaluating them requires methods that are sensitive to each of these processes individually. A narrative that captures attention brilliantly but builds a misleading mental model should not receive the same assessment

as one that is less engaging but supports accurate understanding. Similarly, clear representations without integration into prior knowledge yield fragile comprehension. Recognizing this staged cognitive architecture enables a shift from coarse quality judgments to diagnostically meaningful evaluation, a challenge we take up after examining the limitations of current approaches.

2.4 The Evaluation Gap

Current evaluations are misaligned with how explanatory narratives support understanding. Prior work relies on four main approaches, but each falls short for explanatory narratives.

Holistic Quality Judgments. A common quality evaluation approach asks evaluators, whether human or LLMs, to rate texts on global dimensions such as *"coherence"*, *"engagement"*, *"clarity"*, or *"overall quality"* [14, 18, 28, 40, 52]. In conversational IR, campaigns such as TREC iKAT evaluate responses on dimensions including relevance, completeness, groundedness, and naturalness [2]. While intuitive, such judgments conflate multiple cognitive processes into a single score. A narrative rated as *"clear"* may be clear at the surface level yet fail to support integration with prior knowledge. A CIS system response rated as *"complete"* may cover all relevant information yet structure it in a way that prevents coherent mental model construction. Holistic ratings provide an aggregate impression but cannot distinguish between these diagnostically critical differences. They tell us a narrative is perceived as good or bad, but not *why* or *where* in the comprehension process it succeeds or fails.

Surface-Level Metrics. Automated evaluation metrics widely used in natural language processing (NLP) and IR, such as those based on text overlap (e.g., BLEU, ROUGE), semantic similarity (e.g., BERTScore), or fluency scoring, assess linguistic properties rather than their cognitive effects [3, 43, 59, 80]. These metrics were designed for tasks such as machine translation and summarization, where fidelity to a reference text is the primary concern [43, 59]. In the generative IR context, Xu et al. [77] delivered a landmark finding that no existing automatic metrics are predictive of human preference judgments for long-form answers, while Frummet and Elsweiler [22] found that commonly employed metrics assign higher scores to incorrect conversational answers than correct ones. These results confirm that surface-level metrics evaluate textual similarity, not cognitive impact: a text may score highly on BERTScore while being cognitively opaque, or score poorly while being explanatorily effective through creative analogy or unconventional structure.

Outcome-Based Assessments. Some evaluation approaches measure downstream outcomes, such as recall, accuracy, comprehension test scores, knowledge transfer performance [15, 74, 75], or in IR contexts, task completion and knowledge gain [23, 73, 78]. Research in Search as Learning (SAL) has demonstrated that knowledge gain can be measured from search sessions [23, 73], and recent work has advanced toward personalized knowledge gain estimation [55]. These approaches capture *whether* comprehension occurred, but they do not diagnose *where* in the cognitive pipeline an explanatory narrative succeeded or failed. A low comprehension score could result from poor attention capture (i.e., the user never engaged with the text), inadequate representation (i.e., the user engaged but built an inaccurate mental model), failure to integrate with prior knowledge (i.e., the representation was accurate but isolated), or weak consolidation (i.e., the information was understood but not retained). Without stage-level diagnostics, outcome-based assessments provide aggregate results but not explanatory insight into the source of success or failure.

Dimension-Based Rubrics. More structured approaches decompose evaluation into multiple dimensions, such as "narrative structure", "factual accuracy", or "logical coherence" [14]. Recent IR frameworks evaluate generated answers using similar quality dimensions, mainly consistency, correctness, completeness, relevance, coherence, and source grounding [20, 26, 49, 60]. These represent an improvement over holistic ratings by acknowledging that quality is multifaceted. However, such dimensions are defined based on textual or informational properties rather than cognitive processes. They describe *what the text does* (e.g., whether it is faithful to sources or covers all relevant nuggets) rather than *what the text does to the user's mind* (e.g., whether it captures attention, builds a coherent mental model, or activates relevant prior knowledge). A system response can score highly on faithfulness, completeness, and coherence while still failing to support the cognitive processes required for genuine understanding.

Across cognitive science and IR, a fundamental disconnect persists between *evaluation* and *cognition*: explanatory narratives aim to support understanding but are rarely assessed in terms of the cognitive processes that enable it. In IR terms, current methods capture at most what Saracevic [66] termed *algorithmic* and *topical* relevance, while leaving *cognitive relevance*, the relation of information to the user's knowledge state, entirely unoperationalized. Most existing evaluation methods focus on text surface features, overall quality impressions, completeness, or outcomes such as learning gains [5, 39]. They do not indicate how explanatory narratives support understanding across the comprehension stages. ARIC addresses this gap by evaluating explanations through the cognitive architecture of comprehension rather than textual proxies or single global scores, identifying which cognitive stage succeeds or fails and providing fine-grained diagnostics for why a narrative works or breaks down.

3 The ARIC Framework

We introduce **ARIC**, a cognitively grounded evaluation framework for CIS. As illustrated in Figure 2, ARIC decomposes the evaluation of explanatory narratives into four stages: **Attention**, **Representation**, **Integration**, and **Consolidation**, each corresponding to a human information processing phase. Each stage is further decomposed into a set of diagnostic characteristics (shown in the left panel of Figure 2). The right panel of the figure previews ARIC's diagnostic capability: two explanatory narratives that holistic evaluation would rate similarly, a "Boring but Clear" narrative and an "Exciting but Confusing" narrative, receive sharply divergent stage-level profiles, revealing that their strengths and weaknesses lie at entirely different cognitive stages.

The theoretical foundation of ARIC draws on two complementary models of human cognition. **Information Processing Theory (IPT)** conceptualizes the mind as a system that processes information through sequential stages: sensory input is first filtered

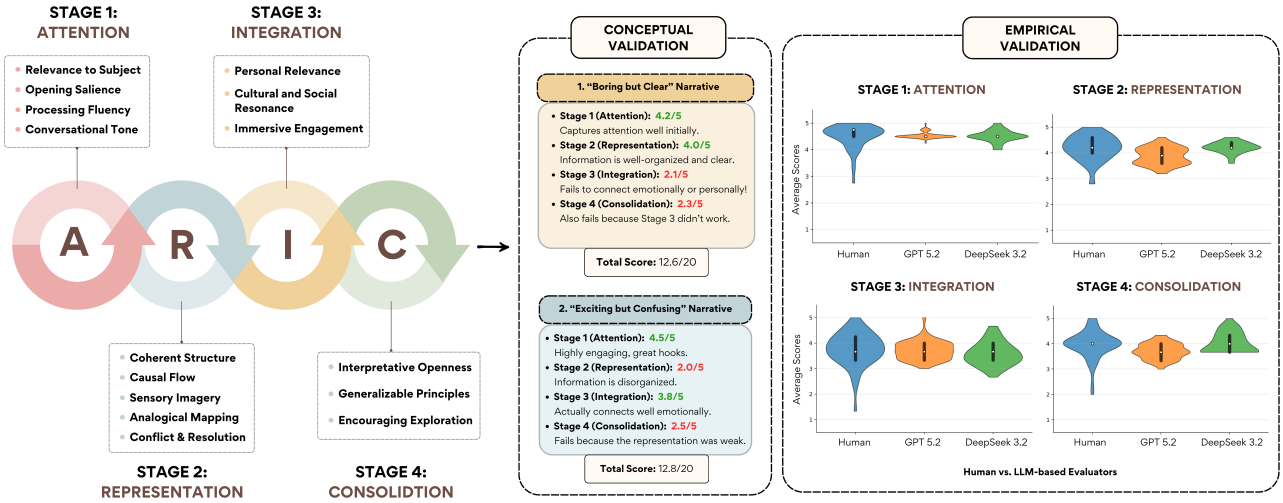

Figure 2: Left: The ARIC framework with its four cognitive stages and fifteen diagnostic characteristics. (See Section 3) Center: Conceptual validation showing how two narratives with nearly identical total scores (12.6 vs. 12.8 out of 20) receive divergent stage-level profiles, revealing that they fail at entirely different cognitive stages. Right: Empirical validation, conducted on human-authored explanatory narratives (TED-Ed Lessons), comparing human and LLM-based evaluations across all four ARIC stages. Agreement is strongest at early stages and diverges for subjective dimensions at later stages. (See Section 4)

Table 1: Stage 1 – Attention: Diagnostic characteristics, definitions, examples, and failure indicators. All examples reference the vaccine explanatory narrative from Figure 1.

Characteristic	Definition	Effective Example	Failure Indicator
Relevance to Subject	Establishes immediate connection to the explanatory topic and the user's goals or curiosity	Opens by addressing a common question: *"You may have wondered what happens inside your body after a vaccine..."*	Opens with tangential content (e.g., lengthy 18th-century epidemiology history) unrelated to the core explanation
Opening Salience	First moments are vivid, surprising, or provocative enough to trigger attentional capture	Uses a striking fact: *"Your immune system has more cells than there are stars in the Milky Way"*	Opens with flat, predictable framing: *"In this text, we will explain the mechanism of immune response"*
Processing Fluency	Language is perceptually and linguistically easy to process	Uses accessible, jargon-free language with clear sentence structure	Dense, jargon-heavy prose: *"APCs phagocytose the pathogen and display epitopes via MHC class II molecules..."*
Conversational Tone	Adopts an engaging, personalized voice that creates a sense of dialogue	Addresses the user directly: *"Imagine your body as a medieval castle..."*	Detached, impersonal register: *"The human body can be conceptualized as a fortified structure..."*

through attentional mechanisms, then encoded into working memory where it is organized and elaborated, integrated with prior knowledge stored in long-term memory, and ultimately consolidated into durable, retrievable representations [4, 5, 50]. **Mayer's Cognitive Theory of Multimedia Learning (CTML)** refines this architecture for learning contexts, proposing that meaningful learning requires three coordinated processes: *selecting* relevant information, *organizing* it into coherent verbal and pictorial mental models, and *integrating* these with existing knowledge schemas [48]. ARIC synthesizes these two theoretical traditions into a unified evaluative lens, where each stage captures a distinct cognitive function that an explanatory narrative must support for comprehension to succeed.

Critically, these stages are not independent. They form a *cascading dependency*: failure at an earlier stage constrains what can be achieved at later stages. If an explanatory narrative fails to capture attention, it makes it difficult to build the representation. If the representation is incoherent, there is less meaningful content to integrate. If integration does not occur, consolidation produces fragile, isolated knowledge.

3.1 Stage 1: Attention

Comprehension begins with attention. In IPT, attention serves as the gating mechanism that determines which incoming information enters working memory for further processing [5, 54]. Without attentional engagement, subsequent cognitive operations, encoding, elaboration, integration cannot begin. In Mayer's CTML, this corresponds to the *selection* phase, in which learners identify and attend to relevant verbal and visual information from the incoming stream [48]. The question for evaluation at this stage is: *Does the explanatory narrative recruit and sustain the user's attentional resources*

Table 2: Stage 2 – Representation: Diagnostic Characteristics, definitions, examples, and failure indicators.

Characteristic	Definition	Effective Example	Failure Indicator
Coherent Structure	Information is organized in a logical sequence that supports progressive understanding	Castle → enemy → army mobilization → training exercise: each element builds on the previous	Jumps between concepts (e.g., antibodies before antigens) or mixes analogy with technical jargon without transition
Causal Flow	Cause-and-effect relationships are explicit and traceable throughout the explanatory narrative	*Because* the vaccine introduces weakened pathogens, the immune system learns; *because* it learns, it responds faster	Lists disconnected facts: *"Vaccines contain antigens. The body produces antibodies. The person becomes immune"*
Sensory Imagery	Vivid, sensory-rich language supports mental simulation and dual coding	Thick castle walls, watchful soldiers, invading enemies, and scout messengers create a vivid mental scene	Abstract, imagistically barren: *"The adaptive immune system generates a targeted response through memory cell proliferation"*
Analogical Mapping	Analogies bridge abstract concepts to familiar domains with structurally sound mappings	Castle walls → skin, soldiers → immune cells, enemy → pathogens: each mapping preserves functional roles	Superficial analogy (e.g., "security alarm") that captures detection but misses the adaptive, learning dimension
Conflict & Resolution	A tension or problem drives the explanation forward and is resolved through explanatory content	Enemy threatens the castle (conflict); army trains in advance via the vaccine (resolution)	Flat exposition listing immune components without establishing why they matter or what problem they solve

sufficiently for cognitive processing to begin? ARIC operationalizes the Attention stage through four diagnostic characteristics:

Relevance to Subject. This characteristic assesses whether the explanatory narrative establishes a clear and immediate connection to its explanatory subject matter. Cognitive research on attention demonstrates that relevance is one of the strongest predictors of attentional allocation: people attend more readily to information that is perceived as pertinent to their current goals, questions, or concerns [8, 31, 45, 66]. An explanatory narrative about vaccine mechanisms that opens by addressing a common fear (e.g., *"You may have wondered what actually happens inside your body after you receive a vaccine"*) signals relevance immediately, aligning the explanatory narrative with the user's existing curiosity. *Failure* manifests when an explanatory narrative opens with tangential or overly broad content that does not signal what will be explained or why it matters, such as beginning a vaccine explanation with a lengthy historical overview of 18th-century epidemiology, causing users to disengage before reaching the explanatory core.

Opening Salience. Opening salience captures whether the first moments of the explanatory narrative are sufficiently vivid, surprising, or provocative to capture attention. This aligns with the psychological concept of *orienting response*, an automatic attentional shift toward novel or unexpected stimuli [54, 68]. In the context of explanatory narratives, opening salience might involve a striking fact (e.g., *"Your immune system contains more cells than there are stars in the Milky Way"*), a vivid scenario, or a thought-provoking question. The aim is to create an attentional "hook" that motivates continued processing. *Failure* occurs when the explanatory narrative opens with flat, predictable, or overly technical language that does not differentiate itself from background information, such as beginning with *"In this text, we will explain the mechanism of immune response"*, a framing that provides no cognitive incentive to continue reading.

Processing Fluency. Processing fluency refers to the ease with which the explanatory narrative can be perceptually and linguistically processed [3, 63]. Research in cognitive psychology shows that easily processed information is perceived as more truthful, more pleasant, and more attention worthy[3, 63]. For explanatory narratives, fluency is shaped by sentence complexity, vocabulary accessibility, syntactic clarity, and overall readability. An explanatory narrative that explains complex immunology using accessible, jargon-free language and short, well-structured sentences maintains attentional engagement. *Failure* manifests as dense, jargon-heavy prose that imposes excessive cognitive load at the surface level, causing users to abandon processing before meaning can be extracted, for example: *"The antigen-presenting cells phagocytose the pathogen and display epitopes via MHC class II molecules to naïve CD4+ T-helper lymphocytes"*, which, while accurate, overwhelms a general audience and disrupts attentional flow.

Conversational Tone. This characteristic assesses whether the explanatory narrative adopts a voice that feels engaging and accessible rather than distant or impersonal. Research on discourse processing shows that a conversational, personalized tone (e.g., using second person, posing rhetorical questions, adopting an informal register) increases cognitive engagement and learning outcomes, a phenomenon Mayer terms the *personalization principle* [48]. An explanatory narrative that addresses the user directly (e.g., *"Imagine your body as a medieval castle..."*) creates a sense of dialogue that sustains attention. *Failure* occurs when the explanatory narrative adopts a detached, impersonal academic register that creates psychological distance, such as *"The human body can be conceptualized as a fortified structure in which immunological agents perform defensive operations"*, which conveys similar content but without the relational warmth that sustains engagement. Table 1 summarizes the Attention stage characteristics with their definitions, examples of effective realization, and indicators of failure.

3.2 Stage 2: Representation

The next cognitive step is constructing a coherent mental representation of the information. In IPT, this corresponds to the encoding and organization of information within working memory, where incoming data is structured into meaningful patterns [4, 5]. In Mayer's CTML, this maps to the *organizing* phase, in which learners build coherent verbal and pictorial models from selected information [48].

Table 3: Stage 3 – Integration: Diagnostic characteristics, definitions, examples, and failure indicators.

Characteristic	Definition	Effective Example	Failure Indicator
Personal Relevance	Connects explanatory content to the user's own life, experiences, or concerns	*"That soreness after your flu shot was your castle's army beginning its training exercise"*	Remains entirely abstract and impersonal; never references anything the user has experienced
Cultural & Social Resonance	Draws on culturally familiar references and shared knowledge to facilitate schema-based connection	Castle-and-army metaphor activates a widely shared cultural schema of medieval defense	Uses culturally unfamiliar references (e.g., cricket metaphors for an audience that does not know cricket)
Immersive Engagement	Creates involvement, emotional resonance, or experiential presence that deepens connection	*"Imagine standing on the castle walls, watching the horizon for approaching enemies"*	Maintains a cold, distanced posture; presents explanation as something to observe rather than experience

For explanatory narratives, this stage is where the user moves from surface engagement to constructing a structured mental model of the explanation: understanding the components, their relationships, and the causal logic that connects them. The question for evaluation is: *Does the explanatory narrative support the construction of a coherent, structured, and accurate mental model of the explained phenomenon?* ARIC operationalizes the Representation stage through five characteristics:

Coherent Structure. This characteristic assesses whether the explanatory narrative presents information in a logical sequence that supports progressive understanding. Coherent structure enables the user to build a mental model incrementally, with each new piece of information fitting into an emerging framework. Research on text comprehension demonstrates that well-structured texts produce more coherent situation models and better recall [39, 52]. In the vaccine explanatory narrative, the progression from *castle* (the body) to *enemy at the gates* (the pathogen) to *army mobilization* (immune response) to *training exercise* (vaccination) follows a logical sequence where each element builds on the previous one. *Failure* manifests when the explanatory narrative jumps between concepts without logical connectives, for example, introducing antibodies before explaining what antigens are, or switching between the castle analogy and molecular biology terminology without transition, leaving the user with fragmented rather than integrated representations.

Causal Flow. Causal flow assesses whether the explanatory narrative makes cause-and-effect relationships explicit and traceable. In IPT, causal relations are among the most powerful organizers of information in working memory because they create inferential chains that link events meaningfully [39, 70]. An explanatory narrative with strong causal flow ensures that the user can trace *why* each event or mechanism leads to the next. In the vaccine example, the causal chain is explicit: because the vaccine introduces a weakened form of the pathogen, the immune system learns to recognize it; because it has learned, it can respond faster when the real pathogen arrives. *Failure* occurs when causal links are left implicit or are absent entirely, for instance, stating that *"Vaccines contain antigens. The body produces antibodies. The person becomes immune"* without connecting these steps causally, leaving the user with a list of facts rather than a causal model.

Sensory Imagery. This characteristic evaluates whether the explanatory narrative employs vivid, sensory-rich language that supports mental simulation. According to dual-coding theory [58], information encoded through both verbal and imaginal channels produces stronger and more durable representations than information encoded through a single channel. Neuroimaging studies confirm that narrative language activates sensory and motor cortices, suggesting that users mentally simulate described scenes [47, 69]. The castle analogy in the vaccine explanatory narrative is rich in sensory imagery: thick walls, watchful soldiers, invading enemies, and scout messengers, all of which create a vivid mental scene. *Failure* manifests as purely abstract, imagistically barren language, such as *"The adaptive immune system generates a targeted response to previously encountered pathogens through memory cell proliferation"*, which is informationally adequate but provides no perceptual anchors for mental model construction.

Analogical Mapping. Analogical mapping assesses if the explanatory narrative uses analogies or metaphors to bridge abstract concepts and familiar domains, and whether these mappings are structurally sound. Structure-mapping theory [24] holds that effective analogies preserve relational structure between the source (familiar) and target (unfamiliar) domains, i.e., the causal and relational architecture must correspond. The castle analogy maps systematically: walls → skin barriers, soldiers → immune cells, enemy army → pathogens, training exercise → vaccination. Each mapping preserves the functional role of the element. *Failure* occurs when analogies are superficial, misleading, or break down under scrutiny. For example, comparing the immune system to a "security alarm" may capture the detection function but fails to map the adaptive, learning dimension of immunity, potentially producing an incomplete or inaccurate mental model.

Conflict and Resolution. This characteristic assesses whether the explanatory narrative establishes a tension, problem, or question that drives the explanation forward and is resolved through the explanatory content. Narrative theory identifies conflict and resolution as a core structural element that creates coherence and directionality in stories [21, 42]. In explanatory narratives, this structure serves a cognitive purpose: the conflict signals a knowledge gap or a problem to be solved, and the resolution provides the explanatory answer, guiding the user's mental model toward closure [45]. In the vaccine example, the conflict (i.e., an enemy threatens the castle) creates a question in the user's mind ("How will

the castle defend itself?"), and the resolution (i.e., the army trains in advance using the vaccine) provides the explanatory answer. *Failure* manifests as flat, conflict-free exposition where information is presented without narrative tension, for instance, listing the components of the immune system without establishing why they matter or what problem they solve, resulting in a representation that lacks directionality and coherence. Table 2 summarizes the Representation stage characteristics.

3.3 Stage 3: Integration

A coherent representation is needed, but not sufficient, for deep understanding. The representation must be connected to the user's knowledge, experiences, and belief systems. In IPT, integration is the process by which new information is linked to prior knowledge in long-term memory, transforming isolated representations into interconnected, meaningful understanding [4, 5, 39]. In Mayer's CTML, this *integrating* phase is in which learners connect newly organized verbal and pictorial models with relevant prior knowledge activated from long-term memory [48]. Without integration, even a well-constructed representation remains inert: understood in isolation but not connected to the broader web of the user's knowledge and experience. The question for evaluation is: *Does the explanatory narrative facilitate the connection of new information to the user's existing knowledge, experiences, and personal context?* ARIC operationalizes the Integration stage through three characteristics:

Personal Relevance. This characteristic assesses whether the explanatory narrative connects its content to the user's life, experiences, or concerns. Research on self-referential processing shows that information processed in relation to the self is encoded more deeply and remembered better than impersonal information [38, 64]. The *self-reference effect* in memory research shows that relating new information to personal experience activates elaborative encoding processes that strengthen integration with existing knowledge structures [7, 64]. A vaccine explanatory narrative that says *"Think about the last time you got a flu shot, that slight soreness in your arm was your castle's army beginning its training exercise"* bridges the abstract mechanism to a felt bodily experience. *Failure* occurs when the explanatory narrative remains entirely abstract and impersonal, never connecting its explanatory content to anything the user has experienced or cares about, such as explaining vaccine mechanisms exclusively through molecular biology, leaving the understanding technically correct but personally disconnected.

Cultural and Social Resonance. This characteristic evaluates if the explanatory narrative draws on culturally familiar references, shared social knowledge, or collective experiences that facilitate connection. Schema theory holds that comprehension is an active process in which users use existing knowledge frameworks to interpret new information [4, 7]. When a narrative activates culturally shared schemas, it provides ready-made structures onto which new information can be mapped. The castle-and-army metaphor works precisely because most users share a cultural schema for medieval defense, regardless of their immunological knowledge. *Failure* manifests when the narrative relies on references, analogies, or cultural assumptions that are unfamiliar or alienating to the target audience, for instance, using cricket metaphors to explain immune response to an audience unfamiliar with the sport, or employing culturally specific humor that excludes rather than includes, thereby blocking rather than facilitating schema-based integration.

Immersive Engagement. Immersive engagement assesses whether the explanatory narrative creates a sense of involvement, emotional resonance, or experiential presence that strengthens the user's connection to the content. Research on narrative transportation demonstrates that users who feel "transported" into a narrative, experiencing it as vivid and emotionally engaging, show greater attitude change, deeper comprehension, and stronger memory than those who remain detached observers [30]. In the context of explanatory narratives, immersive engagement goes beyond entertainment. It serves the cognitive function of activating emotional and experiential knowledge structures that deepen integration [47]. A vaccine explanatory narrative that invites the user to *"imagine standing on the castle walls, watching the horizon for approaching enemies"* creates experiential presence that enriches the mental model. *Failure* occurs when the explanatory narrative maintains a cold, distanced posture that never invites experiential or emotional participation, presenting the explanation as something to be observed rather than experienced, thereby limiting integration to purely intellectual channels and forgoing the deeper encoding that emotional and experiential engagement affords. Table 3 summarizes the Integration stage characteristics.

3.4 Stage 4: Consolidation

The final cognitive stage concerns whether the understanding achieved through the preceding stages is consolidated into durable, transferable knowledge. In IPT, consolidation refers to the process by which information is stabilized in long-term memory, becoming resistant to forgetting and available for future retrieval and application [5, 50]. While Mayer's CTML does not use the term "consolidation" explicitly, it identifies the goal of meaningful learning as producing knowledge that supports both *retention* (remembering) and *transfer* (applying knowledge to new situations) [48], both of which depend on effective consolidation. In explanatory narratives, consolidation determines whether the understanding generated by the narrative persists and can be applied beyond the immediate reading context. The question for evaluation is: *Does the explanatory narrative support the formation of durable, retrievable, and transferable understanding?* ARIC operationalizes the Consolidation stage through three characteristics:

Interpretive Openness. This characteristic assesses whether the explanatory narrative leaves space for the user to interpret, question, and elaborate on the explanation rather than presenting it as a closed, fully resolved account. Research on generative learning demonstrates that comprehension and retention are enhanced when learners are prompted to actively process information, for example, through self-explanation, elaboration, or question generation, rather than passively receiving it [15, 75]. Constructivist theories of learning similarly emphasize that durable understanding emerges from active sense-making rather than passive absorption [12]. A vaccine explanatory narrative that concludes with *"But what if the enemy changes its disguise? How might the castle adapt?"* invites the user to extend the analogy, generating new inferences

Table 4: Stage 4 – Consolidation: Diagnostic characteristics, definitions, examples, and failure indicators.

Characteristic	Definition	Effective Example	Failure Indicator
Interpretive Openness	Leaves space for active interpretation, questioning, and elaboration rather than closed resolution	*"But what if the enemy changes its disguise? How might the castle adapt?"*	Closes all interpretive pathways: *"And that is exactly how vaccines work, with no exceptions"*
Generalizable Principles	Makes underlying explanatory principles explicit and applicable beyond the specific case	*"Your immune system works this way for all threats: it learns from exposure and remembers"*	Bound to one specific case (e.g., only COVID-19 vaccine mechanics) with no transferable principle
Encouraging Exploration	Motivates the user to seek further information, explore related topics, or apply understanding	*"Scientists are now exploring whether this principle could train your immune system to fight cancer"*	Creates false completeness; ends abruptly without signaling connections to broader questions

and strengthening consolidation through active elaboration. *Failure* occurs when the explanatory narrative closes down all interpretive pathways, presenting the explanation as complete and unquestionable, for instance, ending with *"And that is exactly how vaccines work, with no exceptions"*, which discourages the active processing that supports durable encoding.

Generalizable Principles. This characteristic evaluates whether the explanatory narrative makes its underlying explanatory principles explicit and applicable. Transfer research shows that learners who extract general principles from specific examples are more likely to apply their understanding to novel situations than those who encode only the specific instance [24, 25]. For explanatory narratives, this means the explanation should not only illuminate the particular case but also convey principles that the user can carry forward. A vaccine explanatory narrative that makes explicit the general principle, for instance, *"This is how your immune system works for all kinds of threats: it learns from exposure and remembers for the future"*, gives the user a transferable schema. *Failure* manifests when the explanatory narrative is so tightly bound to its specific example that no general principle emerges, such as explaining only the mechanics of a particular COVID-19 vaccine without conveying the broader logic of adaptive immunity, leaving the user with case-specific knowledge that does not transfer to understanding other vaccines or immune responses.

Encouraging Exploration. This characteristic assesses whether the narrative motivates users to seek further information, explore related topics, or apply their understanding. Research on curiosity and epistemic motivation shows that well-crafted explanations can create a sense of informed curiosity, where understanding one aspect of a phenomenon reveals new questions worth pursuing [8, 45]. This effect, sometimes termed the *"knowledge gap"* mechanism of curiosity [45], is particularly powerful when the explanatory narrative explicitly signals that there is more to learn. A vaccine explanatory narrative that concludes with *"Scientists are now exploring whether the same castle-defense principle could be used to train your immune system to fight cancer"* opens a pathway for continued learning. *Failure* occurs when the explanatory narrative creates a false sense of completeness, suggesting that the topic is fully exhausted and there is nothing more to learn, or when it ends abruptly without signaling connections to broader questions, effectively closing the door on further curiosity and exploration. Table 4 summarizes the Consolidation stage characteristics.

4 Application of ARIC

We demonstrate ARIC's utility in two ways shown in Figure 2: *(i)* conceptual validation of the diagnostic insights, and *(ii)* empirical validation that its stage-wise structure yields consistent, interpretable evaluations across human and automated evaluators.

Conceptual Validation. ARIC's diagnostic power is most evident when comparing narratives that holistic evaluation would rate equivalently. The authors validated the conceptual framework through manual annotation. The center panel of Figure 2 illustrates this with two explanatory narratives that receive nearly identical total scores (12.6 vs. 12.8 out of 20) yet show sharply divergent cognitive profiles. The *"Boring but Clear"* narrative scores well on Attention (4.2/5) and Representation (4.0/5) but collapses at Integration (2.1/5) and Consolidation (2.3/5), failing to connect with the user personally or culturally. The cascading dependency is visible: because Integration fails, Consolidation inherits the weakness, producing a fragile understanding that is unlikely to persist. The *"Exciting but Confusing"* narrative presents the opposite pattern: high Attention (4.5/5) driven by vivid hooks and engaging tone, but poor Representation (2.0/5) due to disorganized structure and unclear causal logic. Integration (3.8/5) partially succeeds through emotional resonance, yet Consolidation (2.5/5) remains weak because there is no coherent representation to retain. Holistic evaluation would rate both narratives as mediocre. ARIC reveals they are mediocre for *entirely different reasons*: one fails at connecting with the user, the other at organizing the explanation itself. This diagnostic specificity directly informs revision: the first narrative needs stronger personal and cultural bridging, not better explanatory content; the second needs disciplined structure, not more engagement.

Empirical Validation. We conducted an annotation study comparing human annotator ratings with LLM-based ratings. We wrote annotation guidelines that operationalize ARIC's 15 diagnostic characteristics as five-point Likert items with score anchors. We sampled 50 explanatory narratives from TED-Ed lessons across diverse topics. Three annotators (two female, one male, aged 25–28, all holding graduate degrees in computing-related disciplines) independently evaluated each narrative across all 15 ARIC diagnostic characteristics. Annotators received detailed guidelines to promote conceptual consistency; no calibration or consensus discussions were conducted. Inter-annotator agreement, measured by two-way agreement intraclass correlation for averaged ratings (ICC(A,k)) [51],

was excellent for Attention (ICC = 0.96, 95% CI [0.93, 0.97]) and good for Representation (0.82 [0.71, 0.89]), but moderate for Integration (0.53 [0.25, 0.72]) and Consolidation (0.45 [0.12, 0.67]). This gradient quantifies the increasingly subjective, experience-dependent character of the later stages and is expected given the absence of calibration rounds. Human scores were aggregated by averaging across annotators at the characteristic level and compared with scores from GPT-based *(gpt-5.2-2025-12-11)* and DeepSeek-based *(deepseek-v3.2)* ARIC evaluators. The right panel of Figure 2 presents the distributional comparison across all four stages. The results indicate consistent stage-wise alignment between human and LLM-based evaluations, with differences arising in variance rather than directional disagreement. Agreement is strongest in *Stage 1 (Attention)*, where all three evaluator types assign closely aligned scores with similar distributional shapes, suggesting that surface-level relevance and attentional cues are robustly recognized across both human and LLM-based assessment. In *Stage 2 (Representation)*, human annotators exhibit a wider expressive range while LLM-based evaluators apply more conservative calibration. In *Stage 3 (Integration)*, human judgment variability increases, mirroring the ICC gradient above. Despite this dispersion, LLM-based evaluations maintain similar central tendencies. A similar pattern holds in *Stage 4 (Consolidation)*: human annotations show broader dispersion for interpretive openness and encouraging exploration, while LLM-based evaluators produce tighter distributions with comparable central tendencies. Across all stages, human and LLM-based evaluations exhibit consistent stage-level profiles, with LLM evaluators applying stricter calibration. These results provide preliminary evidence that ARIC can serve as a consistent, interpretable, and scalable evaluation framework. We additionally conducted preliminary ARIC evaluations on LLM-generated responses to QA queries. Full prompts, decoding parameters, per-run outputs, and the QA-query results are available in our repository.

5 Conclusion and Future Directions

This paper advances CIS systems evaluation in two steps. We define Explanatory Narratives as explanations that use narrative structure to support understanding of "why" and "how". We then introduce ARIC, a cognitively grounded framework that evaluates explanatory narratives across four comprehension stages (**A**ttention, **R**epresentation, **I**ntegration, and **C**onsolidation). We support ARIC with conceptual analysis and initial empirical evidence showing that it provides stage-level diagnostics rather than a single quality score. Future work spans validation, adaptation, and the use of diagnostics to improve generation. Our main empirical study used human-authored educational narratives, which differ from CIS responses in length, intent, and constraints; our preliminary evaluation of LLM-generated responses to QA queries provides initial evidence that ARIC transfers to system-generated explanations. Extending this validation to broader CIS settings, such as RAG answers and multi-turn dialogue responses, and at a larger scale with human annotation, remains necessary to fully establish operational utility. ARIC could also be adapted to different audiences. Integration and Consolidation depend on the user's prior knowledge, goals, and cultural context, so evaluation criteria should be tailored to audience characteristics. This directly connects to personalized evaluation in Search as Learning. Finally, ARIC diagnostics could be used to improve generation. Stage-level scores can guide targeted revisions, such as strengthening causal links when Representation is weak or adding culturally grounded analogies when Integration falls short.

Acknowledgments

This work has been funded by the state of North Rhine-Westphalia, as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence and the InVirtuo 4.0: Experimental Research in Virtual Environments project, as well as the 2025 Bonn-Melbourne Research Excellence Fund. We would like to thank the anonymous reviewers for their valuable feedback.

References

[1] Zahra Abbasiantaeb, Simon Lupart, Leif Azzopardi, Jeffrey Dalton, and Mohammad Aliannejadi. 2025. Conversational gold: Evaluating personalized conversational search system using gold nuggets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3455–3465.

[2] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. 2024. Trec ikat 2023: The interactive knowledge assistance track overview. *arXiv preprint arXiv:2401.01330* (2024).

[3] Adam L Alter and Daniel M Oppenheimer. 2009. Uniting the tribes of fluency to form a metacognitive nation. *Personality and social psychology review* 13, 3 (2009), 219–235.

[4] John R Anderson. 2013. *The architecture of cognition*. Psychology Press.

[5] Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*. Vol. 2. Elsevier, 89–195.

[6] Krisztian Balog, Don Metzler, and Zhen Qin. 2025. Rankers, judges, and assistants: Towards understanding the interplay of LLMs in information retrieval evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3865–3875.

[7] Frederic Charles Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.

[8] Daniel E Berlyne. 1960. Conflict, arousal, and curiosity. (1960).

[9] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1196–1207.

[10] Jerome Bruner. 1991. The narrative construction of reality. *Critical inquiry* 18, 1 (1991), 1–21.

[11] Jerome Bruner. 2009. *Actual minds, possible worlds*. Vol. 1. Harvard university press.

[12] Jerome S Bruner. 1961. The act of discovery. *Harvard educational review* (1961).

[13] Olivia M Bullock, Hillary C Shulman, and Richard Huskey. 2021. Narratives are persuasive because they are easier to understand: examining processing fluency as a mechanism of narrative persuasion. *Frontiers in Communication* 6 (2021), 719615.

[14] Cyril Chhun, Pierre Colombo, Fabian Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*. 5794–5836.

[15] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science* 13, 2 (1989), 145–182.

[16] Michael F Dahlstrom. 2014. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the national academy of sciences* 111, supplement_4 (2014), 13614–13620.

[17] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624* (2020).

[18] Yashar Deldjoo, Nikhil Mehta, Maheswaran Sathiamoorthy, Shuai Zhang, Pablo Castells, and Julian McAuley. 2025. Toward holistic evaluation of recommender systems powered by generative models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3932–3942.

[19] Brenda Dervin. 1998. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of knowledge management* 2, 2 (1998), 36–46.

[20] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of*

the 18th conference of the european chapter of the association for computational linguistics: system demonstrations. 150–158.
[21] Gustav Freytag. 1895. *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs.
[22] Alexander Frummet and David Elsweiler. 2024. Decoding the metrics maze: Navigating the landscape of conversational question answering system evaluation in procedural tasks. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*. 81–90.
[23] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 2–11.
[24] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.
[25] Mary L Gick and Keith J Holyoak. 1983. Schema induction and analogical transfer. *Cognitive psychology* 15, 1 (1983), 1–38.
[26] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, et al. 2024. Evaluating generative ad hoc information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1916–1929.
[27] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. (2009).
[28] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.
[29] Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review* 101, 3 (1994), 371.
[30] Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology* 79, 5 (2000), 701.
[31] Suzanne Hidi and K Ann Renninger. 2006. The four-phase model of interest development. *Educational psychologist* 41, 2 (2006), 111–127.
[32] Douglas Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books (AZ).
[33] Peter Ingwersen. 1992. *Information retrieval interaction*. Vol. 246. Taylor Graham London.
[34] Vahid Sadiri Javadi, Martin Potthast, and Lucie Flek. 2023. OpinionConv: Conversational product search with grounded opinions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 66–76.
[35] Vahid Sadiri Javadi, Fryderyk Róg, Aska Aksa, Johanne R Trippas, Svitlana Vakulenko, and Lucie Flek. 2026. CHARISMA: Character-Based Interaction Simulation with Multi-LLM Agents Toward Computational Social Psychology. In *Proceedings of the ACM Conference on Information Interaction and Retrieval (CHIIR'26)*. 1–5.
[36] Vahid Sadiri Javadi, Johanne R Trippas, Yash Kumar Lal, and Lucie Flek. 2025. Can stories help LLMs reason? curating information space through narrative. In *The 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*. 92.
[37] Frank C Keil. 2006. Explanation and understanding. *Annu. Rev. Psychol.* 57, 1 (2006), 227–254.
[38] William M Kelley, C Neil Macrae, Carrie L Wyland, Selin Caglar, Souheil Inati, and Todd F Heatherton. 2002. Finding the self? An event-related fMRI study. *Journal of cognitive neuroscience* 14, 5 (2002), 785–794.
[39] Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review* 95, 2 (1988), 163.
[40] Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.
[41] Carol C Kuhlthau. 2025. *Seeking meaning: A process approach to library and information services*. Bloomsbury Publishing USA.
[42] William Labov. 1972. *Language in the inner city: Studies in the Black English vernacular*. Number 3. University of Pennsylvania Press.
[43] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[44] Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*. 47–58.
[45] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.
[46] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.
[47] Raymond A Mar. 2011. The neural bases of social cognition and story comprehension. *Annual review of psychology* 62, 1 (2011), 103–134.
[48] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139.
[49] James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, et al. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1904–1915.
[50] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* 102, 3 (1995), 419.
[51] Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods* 1, 1 (1996), 30.
[52] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
[53] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum*, Vol. 55. ACM New York, NY, USA, 1–27.
[54] Risto Näätänen. 2018. *Attention and brain function*. Routledge.
[55] Hadi Nasser, Célia da Costa Pereira, Cathy Escazut, and Andrea Tettamanzi. 2025. Personalized knowledge gain estimation through query-driven learning goal inference in search as learning. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*. 263–272.
[56] Aquiles Negrete and Cecilia Lartigue. 2004. Learning from education to communicate science as a good story. *Endeavour* 28, 3 (2004), 120–124.
[57] Paul Owoicho, Mohammad Aliannejadi, Leif Azzopardi, Johanne R Trippas, and Svitlana Vakulenko. [n. d.]. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation.
[58] Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford university press.
[59] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
[60] Sachin Pathiyan Cherumanal, Lin Tian, Futoon M Abushaqra, Angel Felipe Magnossão de Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R Trippas, Falk Scholer, and Damiano Spina. 2024. Walert: Putting conversational information seeking knowledge into action by building and evaluating a large language model-powered chatbot. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 401–405.
[61] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 180–190.
[62] Hossein A Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles LA Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. Llmjudge: Llms for relevance judgments. *arXiv preprint arXiv:2408.08896* (2024).
[63] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. 2004. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review* 8, 4 (2004), 364–382.
[64] Timothy B Rogers, Nicholas A Kuiper, and William S Kirker. 1977. Self-reference and the encoding of personal information. *Journal of personality and social psychology* 35, 9 (1977), 677.
[65] Vahid Sadiri Javadi, Johanne R Trippas, and Lucie Flek. 2024. Unveiling information through narrative in conversational information seeking. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–6.
[66] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for information Science and Technology* 58, 13 (2007), 2126–2144.
[67] Roger C Schank and Robert P Abelson. 2014. Knowledge and memory: The real story. In *Knowledge and memory: The real story*. Psychology Press, 1–85.
[68] EN Sokolov. 2000. Perception and the conditioning reflex: vector encoding. *International journal of psychophysiology* 35, 2-3 (2000), 197–217.
[69] Nicole K Speer, Jeremy R Reynolds, Khena M Swallow, and Jeffrey M Zacks. 2009. Reading stories activates neural representations of visual and motor experiences. *Psychological science* 20, 8 (2009), 989–999.
[70] Tom Trabasso and Paul Van Den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language* 24, 5 (1985), 612–630.
[71] Johanne R Trippas. 2024. Conversational search. In *Information Retrieval: Advanced Topics and Techniques*. 285–319.
[72] Johanne R Trippas, Luke Gallagher, and Joel Mackenzie. 2024. Re-evaluating the command-and-control paradigm in conversational search interactions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2260–2270.
[73] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18.
[74] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist* 46, 4 (2011), 197–221.

[75] Merlin C Wittrock. 1974. Learning as a generative process. *Educational psychologist* 11, 2 (1974), 87–95.
[76] James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.
[77] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3225–3245.
[78] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting user knowledge gain in informational search sessions. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 75–84.
[79] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval* 17, 3-4 (2023), 244–456.
[80] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. (2019).