Leveraging Graph-Centric Case-Based Reasoning for Enhancing Monotonicity & AI Coherency

Steve Chan
VTIRL, VT/DE-CAIR
Orlando, USA
stevec@de-cair.tech

***Abstract*—The Artificial Intelligence (AI) coherence (and concomitant AI hallucination) issue, which centers upon the issue of validity, among others, remains an ongoing challenge, and current mitigation methods have had limited efficacy for certain Large-Language Model (LLM)-based systems. This paper reviews the aforementioned points and presents some experimental conjoining of certain similarity measures along with AI-centric heuristic updating/generating in an endeavor to operationalize pathways that: (1) tend towards the monotonic arena, and (2) are, correspondingly, less likely to spawn towards the Non-deterministic Polynomial-time Hardness (NP-hard) non-continuous, non-polynomial, non-monotonic side.**

Keywords*—*Artificial Intelligence; Conversational AI Agents; AI Hallucinations; AI Coherence; Graph-Based Reasoning; Case-Based Reasoning; Analogical Reasoning; Isomorphism.

I. INTRODUCTION

This paper builds on [1], which covered Decision Quality (DQ) and certain Extrapolated Decision Quality (DQ) Thematics (EDQTs); DQ is a key underpinning element of Robust Dialogue Management (RDM), which is a non-trivial matter for Conversational AI Agents (CAs). For their conversational dialogues, CAs have the challenge of Sequential Decision-Making (SDM), which impacts maintaining logical flow, consistency, validity, and overall coherency during the course of not only a single conversation, but also, potentially, multi-turn conversations. This then segues into the matter of monotonic constraints, which are often looked to for the enforcement of logical flow and coherence. However, Real-World Scenarios (RWS) often tend to involve the non-monotonic side, wherein conclusions can be withdrawn/reversed/altered. While conventional reasoning involves deductive, inductive, and abductive forms, in contemporary times, Inductive Reasoning (IndR) seems to be among the more prevalently used, and there are Kahneman & Tversky underpinnings associated with this. When time is not of the essence, such as in a System 2 sense ("slow, deliberate, logical"), Deductive Reasoning (DedR) can be utilized for more complicated matters; however, when time is of the essence, such as in a System 1 sense (i.e., "fast, automatic, intuitive"), IndR might be more optimal. Within the category of IndR, Analogical Reasoning (AnaR) is often used, and within this, Case-Based Reasoning (CBR) seems to be among the more highly exercised. This constitutes an opportunity since AnaR is construed to be associated with the non-monotonic realm while CBR is cautiously monotonic; after all, the literature indicates that LLM-based CAs are likely to be better at example-based learning (e.g., CBR) rather than adhering to explicit guardrail or system prompt instructions. In turn, CBR can also be graph-based, thereby affording the opportunity to leverage Graph-Based Reasoning (GBR) and move from cautiously monotonic to monotonic. In this way, it is possible (for certain cases) to avail of the benefits of AnaR, via GBR/CBR, while staying more on the monotonic side.

This paper has been substantially re-written with new content (as contrasted to the conference paper). This paper is intended as a design paper, rather than as a complete empirical contribution, and focuses on reviewing the Artificial Intelligence (AI) coherence (and concomitant AI hallucination) issue, which centers upon the issue of validity, and delineates some of the current mitigation methods, which have had limited efficacy for particular Large-Language Model (LLM)-based systems. The paper presents some experimental conjoining of certain similarity measures along with AI-centric heuristic updating/generating in an endeavor to operationalize pathways that: (1) tend towards the monotonic arena, and (2) are, correspondingly, less likely to spawn towards the Nondeterministic Polynomial-time Hardness (NP-hard) noncontinuous, non-polynomial, non-monotonic side. Section I provided an overview regarding the challenges and complexities of RDM, which is a core requisite capability for CA. The remainder of the paper is organized as follows. Section II provides pertinent background information pertaining to the challenges of AI coherence (and hallucinations). Section III presents some theoretical foundations, experimental building blocks, and experimentation. Section IV provides a brief discussion and lists some of the limitations of the paper. Section V summarizes with concluding remarks, and proposed future work closes the paper.

II. BACKGROUND

There are a variety of AI LLM leaderboards, but, among others, some of the more highly recognized LLMs include: OpenAI's Generative Pre-trained Transformer (GPT)-5, Alibaba's Tongyi Qianwen (Qwen)-3-235B-A22B, xAI's Grok-3, Meta's Llama 4, Deepseek's V3.1, Anthropic's Claude 3.7, and Google's Gemini 2.5 Pro [2]. These LLMs are leveraged for a myriad of applications, such as CA, which strives to undertake natural human conversation through various modes (e.g., text, voice). Some of the more popular CAs include Apple's Siri, ChatGPT, Gemini, and ElevenLabs AI 2.0. Yet, these CAs are

beset by the technical issue of what is often referred to as "AI hallucinations," which OpenAI deems to be "plausible but false statements" and IBM deems to be "nonsensical or altogether inaccurate" outputs [3][4].

A. The Dilemma and Challenge

Against this backdrop of AI hallucinations, the referenced CA market is aggressively growing. The market research firm Grand View Research asserts that the CA "market size was estimated at USD 11,576.4 million in 2024 and is projected to reach USD 41,393.2 million by 2030, growing at a [Compound Annual Growth Rate] CAGR of 23.7% from 2025 to 2030" [5]. The firm Markets and Markets seems to be aligned with this assertion and notes that CA "is projected to be USD 49.80 billion by 2031, growing from 17.05 million in 2025" [6]. The firm Fortune Business Insights is even more aggressive in its assertion: [the CA] "market size is expected to grow from $12.24 billion in 2024 to $61.69 billion in 2032" [7]. In support of the described CA consumer market, Verified Market Research asserts that the CA "platform software market was valued at $234.82 million in 2024 and is projected to reach $589.76 million by 2031" [8]. As a summarization, the various market research firms cite tremendous growth for the CA ecosystem. Therein lies the dilemma; while CA consumer demand is accelerating (along with the underpinning platform infrastructure ecosystem), CA has been beset with technical issues, such as that of the referenced AI hallucinations.

Kate Irwin of *PC Magazine* reports on an interview that the *Washington Post* had with Apple CEO Tim Cook, wherein the takeaway was that "Apple's AI tools may not always be accurate" and may be subject to "AI hallucinations" [9]; William Gallagher of *AppleInsider* adds to this by noting that AI hallucinations may occur much more frequently than expected/desired [10]. A plethora of researchers have published studies regarding the frequency of AI hallucinations for CA, and in the case of Athaluri (as an exemplar), it is reported that ChatGPT's AI hallucinations resulted in (out of the 178 references cited): 28 fictitious references and 41 erroneous Digital Object Identifiers (DOIs) (i.e., a 39% hallucination rate) [11]. *Tech Crunch* has reported that "according to OpenAI's internal tests, o3 and o4-mini, which are so-called reasoning models, hallucinate *more often* than the company's previous reasoning models — o1, o1-mini, and o3-mini — as well as OpenAI's traditional, 'non-reasoning' models, such as GPT-4o" [12]. *Mashable* follows up on this by noting that while "o3's hallucination rate is 33 percent,"…"o4-mini's hallucination rate" (for a "more advanced version") is much higher at "48 percent" [13]. OpenAI, in its technical report, entitled "Open AI o3 and o4-mini System Card," notes that "more research is needed" as to why there are increasingly "more inaccurate/hallucinated claims" as the versions advance and as reasoning models are scaled up [14]. Shifting to Google, it self-notes that Gemini "can sometimes hallucinate," and Imad Khan at *CNET* notes that "Gemini can be slow, prone to hallucinate and links to incorrect pieces of information" [15][16]. Hugging Face further comments that the actual hallucination rate for the "latest Gemini model" may be higher than Google's reported rates [17]. In an endeavor to address this matter, during the AI Action Summit of February 2025, Google DeepMind and Giskard announced the Potential Harm Assessment & Risk Evaluation (PHARE) LLM benchmark, and the associated paper reports that "leading LLM systems consistently struggle with [AI] hallucination, exhibiting high variability across different contexts" [18][19][20]. Alan Weissberger remarks on PHARE, via the *IEEE Communications Society*, and notes that "some models" have "hallucination rates exceeding 30% in specialized fields [21]. Along this vein, reportage based upon the "RealHarm Dataset" of "problematic interactions" with CA indicates that "misinformation and fabrication represent approximately one-third of documented incidents, confirming that [AI] hallucination remains the primary challenge in production LLM systems despite significant research attention" [22][23]. Perhaps, OpenAI put it the best on 5 September 2025: "Even as language models become more capable, one challenge remains stubbornly hard to fully solve: [AI] hallucinations" [24].

B. Conventional Mitigation Approaches

Human oversight/validation is an axiomatic mitigation approach for addressing AI hallucinations. However, putting aside certain "high-stakes"/mission-critical applications (e.g., law, healthcare), the Human-in-the-Loop (HITL) approach does not necessarily scale well or at speed, as noted by Holzinger and others [25]. Apart from HITL, there are a variety of more automated approaches, among others, for addressing AI hallucinations: (1) Guardrails/System Prompts (GSP), (2) Fine-Tuned Models (FTM) for specialized domains, and (3) Real-time Retrieval Augmented Generation (RAG). These mitigation approaches are discussed in 1) through 3).

1) Guardrails/System Prompts (GSPs)

Andrew Cunningham of *Ars Technica* reminds us of existing system prompts, which — even as acknowledged by Apple itself — may not rise to the desired levels of mitigation against AI hallucinations; some of these system prompts are a constituent part of various metadata.json files within the "System/Library/AssetsV2/com_apple_MobileAsset_UAF_FM_GenerativeModels/purpose_auto folder on Macs running the macOS Sequoia 15.1 beta that have also opted into the Apple Intelligence beta" [26]. Upon examination, mitigating system prompts include: "'do not hallucinate,' 'do not make up factual information,' etc." [26]. Moving from system prompts to guardrails, there are OpenAI guiding principles (in their "Cookbook"), such as: "provide very descriptive metrics to evaluate whether a response is accurate," "ensure consistency across key terminology," "evaluate each sentence independently and then entire response as a whole," etc. [27]; on the Amazon front, guardrails include using a "contextual grounding check policy" to "detect and filter AI hallucinations in model responses that are not grounded in enterprise data" [28]. In short, while system prompts can serve as shaping operations to help ensure better alignment with the envisioned accuracy/precision and contextual relevance, guardrails can serve as a Quality Assurance/Quality Control (QA/QC) mechanism for the output of the system prompts.

2) Fine-Tuned Models(FTMs) for Specialized Domains

FTMs can indeed constitute a potential mitigation strategy for reducing AI hallucinations by leveraging/applying a pre-trained model to a particular "high-quality dataset" in the hopes of anchoring AI outputs via domain-specific knowledge (e.g., declarative, procedural, conditional, etc.) [29]. By way of background, declarative knowledge (e.g., recitals of fact, concepts) is "knowing what," procedural knowledge (e.g., step-by-step skills that are utilized/actions that are instinctively performed via "implicit memory" or "muscle memory") is "knowing how," and conditional knowledge is "knowing when and why" to strategically apply declarative and procedural knowledge [30][31][32]. Declarative knowledge can be encapsulated in both Known and Unknown forms (e.g., facts that exist but are not yet learned, facts that cannot be immediately recalled and are temporarily "unknown," "difficult-to-verbalize" recitals of fact/notions that are deemed to be implicit, etc.); likewise, procedural knowledge can be in both Known and Unknown (i.e., implicit) forms. In other words, when handling the "Unknown Unknowns" (UU) and "Known Unknowns" (KU) of the KU, UU, Unknown Knowns (UK), and "Known Knowns" (KK) epistemological (pertaining to the theory of knowledge) model leveraged by Shaker and Moore-Clingenpeel as well as others (and popularized by Donald Rumsfeld), such as shown in Table I below, there exists a dramatic distinction (as pertains to the rate of AI hallucinations) between the Knowns and the Unknowns.

TABLE I. EPISTEMOLOGICAL CONSTRUCTS [33][34]

Known Knowns (KK) "Things we are aware of and understand"	*Known Unknowns (KU)* "Things we are aware of and do not understand"
Unknown Knowns (UK) "Things we are not aware of, but understand"	*Unknown Unknowns (UU)* "Things we are not aware of and do not understand"

While FTMs can better mitigate against AI hallucinations on Known (e.g., KK) information, when contending with Unknown information (e.g., KU, UK, UU), they can, potentially, aggravate the AI hallucinations paradigm. To further accentuate this, Gekhman's Sampling-based Categorization of Knowledge (**S**li**CK**) Known/Unknown model, which is a variant of Table I, is shown in Table II, wherein the last column, *Resultant*, depicts how often greedy decoding predicts the correct answer. As part of the setup for Section III Experimentation herein, Gekhman's study setup is adopted. Hence, "given a fine-tuning dataset D and a pre-trained LLM M," M_D denotes "a model obtained by fine-tuning M on D" [35]. Gekhman had adopted the perspective that "M knows that the answer to q is a[,] if it generates a when prompted to answer q," and Gekhman also defined $P_{correct}(q,a;M,T)$ "as an estimate of how likely is M to accurately generate the correct answer a to q, when prompted with random few-shot exemplars and using decoding temperature T " [35][36][37].

TABLE II. GEKHMAN'S SLICK MODEL [35]

Type	***Category***	***Definition***	***Resultant***
Known	"Highly Known" (HK)	"$P_{correct}(q,a;M,T=0)=1$"	"Always"
	"Maybe Known" (MK)	"$P_{correct}(q,a;M,T=0) \in (0,1)$"	"Sometimes"
	"Weakly Known" (WK)	"$P_{correct}(q,a;M,T=0)=0 \wedge P_{correct}(q,a;M,T>0)>0$"	"Never with T=0, but Sometimes with T>0"
Unknown	"Unknown"	"$P_{correct}(q,a;M,T\geq 0)$"	"Never"

According to Gekhman's ascertainment as to the negative impact of Unknown examples, the finding is that a "higher Unknown ratio is proportional to performance degradation" [35]. Restated, "LLMs struggle to learn new factual information through unsupervised fine-tuning" and this can result in unanticipated/undesirable outcomes (e.g., a higher AI hallucination rate) [38]. Furthermore, as noted by Sun, "a common challenge when fine-tuning LLMs for domain-specific applications is the potential degradation of the model's generalization capabilities" [39]. However, it has been reported that complementary methods, such as Retrieval-Augmented Generation (RAG), may assist with improving performance reliability (e.g., accuracy).

3) Retrieval-Augmented Generation (RAG)

To obviate against simply relying upon internal static training data, RAG can connect to external dynamic sources of data. This helps to contend with potentially out-of-date training data and can, hopefully, enhance the overall contextual awareness for the purpose of lessening the probability of AI hallucinations. Yet, RAG is also plagued by issues, such as the retrieval of specious and/or irrelevant information as well as the occasional erroneous fusing/leveraging of external information.

C. An Alternative Approach: Moving from AI Hallucinations to AI Coherence

It should be noted that AI hallucinations and AI coherence are distinct and disparate notions (albeit interrelated). While AI hallucinations are a generated response consisting of inaccurate or specious information, AI coherence refers to the "logical consistency" of the output response. The conventional approaches toward the AI hallucination challenge (described in Section IIB) have had limited efficacy (as spotlighted in the latter part of Section IIA). Accordingly, this paper addresses matters upstream of AI hallucinations; hence, AI coherence is treated. The treatment of AI coherency is significant, as a collapse/degradation of coherence can be a harbinger of AI hallucinations (yet, not all AI hallucinations involve incoherence, as an AI hallucination may have logical flow, but still be false). Maintaining coherence (via, theoretically, stringent monotonic constraints) amidst RWS is a complicated matter, "as the CA might discern connections (particularly those that are non-monotonic) within the ever-evolving dataset, and non-monotonic facets may materialize as incoming information re-contextualizes and/or contradicts matters;" moreover, "enforcing a strict monotonic paradigm can segue to an unnatural rigidity and/or incorrect/irrelevant responses by the CA." Prior research has shown that enhanced insight into the CA

behavior at Monotonic/Non-monotonic Transition Zones (MNTZ) can, potentially, be quite meaningful for enhancing coherency and consistency (with the concomitant validity). Yet, the treatment of MNTZ is often not part of contemporary CA architectures. Accordingly, Section III will unpack this further.

III. EXPERIMENTATION

A. Theoretical Foundations

1) Reasoning Mechanisms (RMs)

For CA, the primary RMs, by validity ranking, are likely to be DedR, IndR, and then AbdR. Grote-Garcia states that DedR "is the process of using general premises to draw specific conclusions" (i.e., a "top-down paradigm") [40]. In contrast, for IndR, conclusions are derived by progressing from the specific to the general (i.e., a "bottom-up paradigm"); the University of Illinois Springfield puts it nicely: "inductive reasoning is the ability to combine pieces of information that may seem unrelated to form general rules or relationships" [41]. According to Minnameier, despite the recognized creative aspects (as Holyoak reminds us) of AnaR, it "is widely conceived of" as IndR (which is considered to be non-creative) [42][43]. In terms of validity ranking, AnaR ("the ability to perceive and use relational similarity between two situations or events") is subordinately situated below IndR, but it sits above AbdR (which may start with "puzzling observations" and segue to "inferring the most likely explanations") [44][45]. AnaR and AbdR may also leverage secondary RMs, such as CBR, while the former may also utilize GBR. Yan describes CBR as being "based on the cognitive assumption that similar problems have similar solutions" while Taylor & Francis describes CBR as being "a problem-solving approach that involves using past successful solutions to similar problems to solve new problems" [46][47]. Kolodneer notes that CBR is an AnaR method, and accordingly, CBR is subordinately situated below AnaR, as shown in Table III [48].

TABLE III. TYPES OF PRIMARY REASONING MECHANISMS (RMS) [49]

Information Available		***RM***	***Resultant***
Examples: " • Facts • Accepted Truths • Rules • Scientific Laws	• Mathematical Theorems • Established Principles • Logical Connections"	DedR	"Always true, if the premises and arguments are valid."
"Starts with the same set as Deductive Reasoning (if available) and involves a probabilistic approach."		IndR	"Likely to be true but could be false despite the observations being accurate."
		AnaR	
		CBR	
"Starts with the same set as Deductive Reasoning (if available), but also involves hypotheses, assessments, and best-fit approximations."		AbdR	"Sometimes true, as it is a plausible best guess approximation."

Meanwhile, GBR is a technique that buttresses reasoning capabilities by characterizing problems as graphs and then proceeding to resolve the problems via the ascertaining of connections and patterns. The various primary and secondary RMs discussed can also be sorted by the involved Reasoning Processes (RPs) (i.e., monotonic, non-monotonic).

2) Reasoning Processes (RPs)

As noted in prior work, RPs include Monotonic Reasoning (MR) and Non-Monotonic Reasoning (NMR). MR responses "remain consistent throughout time despite whatever new information might arrive" while NMR responses "allow for modification and/or retraction of prior assertions." As the thematics and priorities within the conversational dialogue change and as new information may alter the context and/or contradict prior information, the addressing of RDM and "MR and NMR becomes critical for maintaining consistency and interconnectedness, particularly for multi-turn conversations;" "if the constituent elements of a multi-turn conversation are indeed logically related, then the overall dialogue should be relatively free of contradictions," and this constitutes a paradigm of "conversational coherence." To validate conversational coherence, there are a variety of evaluation tools that can be leveraged. For example, the "ConversationCoherence Evaluator" is "a tool designed to check the coherence of conversations by an AI," as "it evaluates whether each response in a conversation logically follows from the previous messages, ensuring that the AI maintains context and relevance throughout the interaction" [50][51]. However, "'conversational coherence' is, for CA, quite difficult to maintain because the information supply changes temporally, and at some points, it may be sparse/incomplete and/or ambiguous/uncertain." Depending upon "what" and "when" the information is made available, a particular RM(s) may be more apropos. Table IV, which was reviewed in [49], provides a sampling of RMs as well as their associated RP (i.e., MR/NMR) categorization, which is illuminated using a Red-Orange-Yellow-Green (ROYG) color coding schema, wherein MR is indicated by green, NMR by red, weak MR by orange, and cautious MR by yellow [49][52][53].

TABLE IV. RM-CENTRIC AND RP-CENTRIC SORTING [49]

RM	***MR/NMR Categorization***	
DedR	*MR*	
GBR	*MR*	*NMR*
CBR	*Cautious[ly] MR*	*NMR*
IndR	*Weak[ly] MR*	*NMR*
AnaR	*NMR*	
AbdR	*NMR*	

Table IV is quite interesting as the often used AnaR is categorized as NMR. As can be seen, CBR is cautious[ly] MR while GBR is MR. Hence, if there is an opportunity to leverage GBR or GBR/CGR, then there is potential to stay more within the MR realm (i.e., higher coherence).

3) CA, IndR, AnaR, CBR, and GBR

As a follow-on, Chen points out that "one of the intriguing abilities of LLMs is reasoning" [54]. As an extension to this, "one of the intriguing abilities of" CA is also reasoning, which consists of the primary RMs of Table III: DedR, IndR (with its subordinate AnaR and CBR), and AbdR. While Chen asserts that IndR is central to LLMs[/CAs], in Chen's study, it was

found that LLM performance was somewhat sub-optimal for IndR [54]. Similarly, Luo reports that, when using the LogiGLUE benchmark (which is comprised of 24 datasets consisting of DedR, IndR, and AbdR), "the findings indicate that LLMs excel most in" AbdR, "followed by" DedR, "while they are least effective" for IndR [55]. Along this vein, Cheng points out that "while the" DedR "capabilities of LLMs, (i.e. their capacity to follow instructions in reasoning tasks), have received considerable attention, their abilities in true" IndR "remain largely unexplored" [56]. Cheng also claims that "LLMs demonstrate remarkable" IndR "capabilities through SolverLearner," but Eliot cautions that "depending upon how the generative AI was devised by an AI maker, such as the nature of the underlying foundation model, the capacity to undertake" IndR varies greatly [56][57]. Overall, "the findings suggest that LLMs might be better at learning by example and discovering patterns in data than at following explicit instructions," thereby hinting at the value-added proposition of CBR (and AnaR) [58]. Leake suggests that CBR provides a "basis for learning from few examples" (few-shot), and Qin asserts that AnaR can be used "to address unfamiliar challenges by transferring strategies from relevant past experiences"/examples [59][60]; as a reminder, CBR is a specialized form of AnaR [49][61].

CBR involves a similarity-based comparison, which might include various key features. An ensuing more robust comparison might entail finding cases/examples that are structurally identical and involve an isomorphic-based comparison; this can be achieved, such as by way of an Isomorphic Paradigm (IsoP) Comparator Similarity Measure (CSM) (ICSM), and prior work in this regard has shown that graph-based CBR (i.e., a GBR/CBR amalgam) can be quite suitable for IsoP. As pertains to the CSM, in some cases, such as for unordered sets, the ordering of the edges (and their weights) may not necessarily be relevant, as only the nodes and their values need to be compared; for example, when providing the temperature (T), relative humidity (RH), and wind speed (WSPD) (along with their associated values) for a particular point in time, the sequencing of T, RH, and/or WSPD is not of any particular import. In other cases, such as for ordered sets, the edges and the sequencing of the nodes is of significance, such as the [node] stops that are to be made along a delivery route. Some CSM considerations (e.g., Pre-IsoP, IsoP) from the prior work of [49] are presented in Table V, sample Partially Ordered Sets (POSETs) are shown in Table VI, and exemplar Isomorphic Variants (IVs) are shown in Table VII.

TABLE V. CSM (PRE-ISOP AND ISOP) CONSIDERATIONS [49]

Considerations	***Definition***
Unordered Set (UnS)	"A set of disparate constituents, wherein the order of the constituents is not relevant. By way of example," {T, RH, WSPD} equates to {WSPD, T, RH} and "{1, 2, 3, 4, 5} equates to {5, 3, 1, 4, 2}."
Equal Sets (EqS)	"A pair of sets S and S' is equal if and only if (iff) each constituent of S is also a constituent of S'; moreover, the order of the constituents is not relevant. By way of example, if S = [1, 2, 3, 8, 9, 10] and S' = {9, 3, 1, 2, 10, 8}, then S=S'."
Equivalent Sets (EquivS)	"A pair of sets S and S' is considered equivalent if the number of constituents in S and S' is the same (i.e., same cardinality). By way of example, if S = {1, 3, 5, 7, 9} and S' = {2, 4, 6, 8, 10}, then S and S' are considered to be equivalent."
Ordered Set (OrS)	"A set of disparate constituents, wherein the order of the constituents is relevant, and the constituents can be ordered and compared via operators, such as by <. By way of example, an ordered set might be {1, 2, 3, 5, 6, 8, 9, 10}, whereas an unordered set might be {6, 5, 1, 2, 3, 10, 9, 8}."
Partially Ordered Set (POSET)	"A set of disparate constituents, wherein the constituents might or might not be able to be ordered and compared, since operators such as <= can yield different variations. By way of example, Calcworkshop (https://calcworkshop.com/relations/partial-order/) provides some examples, which we extrapolate upon in the way of {a < b < c < d <= e <= f }, {a <= b <= c < d < e <= f < g}, and {a < b < c <= d <= e < f}, which are shown diagrammatically in Table VI."
Unordered Sets with Isomorphism (UnS-Iso)	"For a set of disparate constituents, wherein the constituents are unordered, if there is a one-to-one relationship (i.e., bijection), then the unordered sets are likely isomorphic. By way of example, if S={1, 2, 3, 4, 5}, S'={a, b, c, d, e}, and 1<->a, 2<->b, 3<->c, 4<->d, and 5<->e (wherein each constituent in S relates to a unique constituent in S'), then S and S' are considered to be isomorphic."
POSETs with Isomorphism (POSET-Iso)	"For a set of disparate constituents, wherein the constituents are considered to be within a POSET, if there is a bijection, then the POSETs are likely isomorphic. By way of example, if S={S_1, S_2, S_3, S_4, S_5}, S'={S'_2, S'_4, S'_5, S'_3, S'_1}, and S_1<->S'_2, S_2<->S'_5, S_3<->S'_1, S_4<->S'_4, and S_5<->S'_3, (wherein each constituent in S relates to a unique constituent in S'), then S and S' are considered to be isomorphic. To demonstrate this, online tools are available, such as https://graphonline.top/en/?graph=xPLjwOkrglDRgYe, among others."
Isomorphism Variants (IV)	"Extrapolating upon the POSETs with Isomorphism (POSET-Iso), there are also permutations that are actually isomorphism variants (e.g., automorphism, which is a particular type of isomorphism that has a symmetrical structure), which Lemons nicely depicts and for which examples are shown in Table VII as IV#1 through #3" [62].

TABLE VI. ISOMORPHIC PARTIALLY-ORDERED SETS (POSETS) [49]

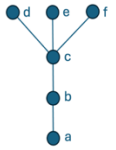
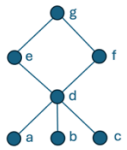
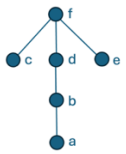
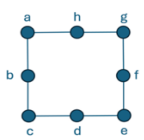
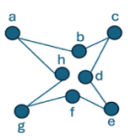
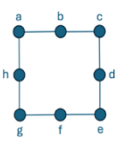
POSET #1	***POSET #2***	***POSET #3***
		
Maximal: d, e, f	Maximal: g	Maximal: f
Greatest: none	Greatest: g	Greatest: f
Minimal: a	Minimal: a, b, c	Minimal: a
Least: a	Least: none	Least: a

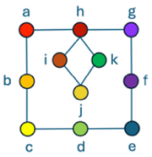
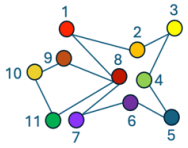
TABLE VII. ISOMORPHIC VARIANTS (IVS) [49]

IV #1	***IV #2***	***IV #3 (automorphism)***
		

With regards to GBR, if there exists a one-to-one correspondence between the vertices of S and S', then S and S'

are isomorphic. This is shown in Table VIII (which was previously presented in [49]), and this can be affirmed via a variety of tools, such as the one available at https://graphonline.top/en/?graph=xPLjwOkrglDRgYeS. Also, rather than the graphs themselves, adjacency matrices can also be utilized to determine isomorphism. This can be affirmed via a variety of tools, such as the one available at https://graphonline.top/en/create_graph_by_matrix. Given this versatility, the leveraging of graph-based CBR can be quite advantageous; in fact, when GBR and CBR are conjoined (i.e., GBR/CBR), it is possible to, potentially, move towards the green MR side (as contrasted to the yellow cautious[ly] MR side) of Table IV. Li, Xu, and many others seem to be a proponent of this approach [63][64]. By moving from the yellow to the green, it is likely contributory towards reducing the propensity for spawning towards the NP-hard, non-continuous, non-polynomial, and NMR side.

TABLE VIII. EXEMPLAR ISOMORPHISM BETWEEN S AND S' [49]

S and S' Isomorphism	***Graph S***	***Graph S'***
f(a) = 1 *f(b)* =2 *f(c)* =3 *f(d)* =4 *f(e)* =5 *f(f)* =6 *f(g)* =7 *f(h)* =8 *f(i)* =9		

It would also, likely, enhance AnaR via the GBR/CBR amalgam. This is important, as analogies are prevalent in conversation; metaphors are also prevalent. As a quick primer, simile "is a comparison of two disparate entities, via words, such as 'like' or 'as'," metaphor "is a direct comparison and asserts that two disparate entities are the same, via words, such as 'is,' 'was,' etc. (wherein the words 'like' or 'as' are not utilized), analogy "creates a comparison of how a seemingly disparate entity is akin to, relates to, or is similar to another disparate entity for the purpose of explaining/demonstrating," and allegory "embodies a more complex/symbolic comparison and leverages a narrative to convey an abstract notion/concept." In a study by Casarett, it was found that "the use of metaphors and analogies may enhance physicians' ability to communicate," as metaphors appeared in 64% of the conversations, while analogies were used in 31% of the conversation, and on average, doctors used 1.6 metaphors and 0.6 analogies per conversation [65]. According to Kanthan, similes, metaphors, and analogies "bridge the Known to the Unknown" [66]; hence, in accordance with Table II, the utilization of metaphors and analogies (which facilitates the movement from the Unknown to the Known) segues to a paradigm, wherein the resultant validity is likely to be higher (with regards to Table II, the validity for HK is "always," MK is "sometimes," and WK is "sometimes" when T>0). Moreover, the handling can be accomplished by CBR or GBR/CBR rather than simply AnaR; the key distinction here is that GBR/CBR reside in the green/yellow of MR and cautious[ly] MR, respectively, while AnaR resides in the red of NMR.

B. Experimental Building Blocks

Pragmatically, GBR/CBR can leverage various similarity metrics for the determination of similar cases/examples. An alphabetized sampling of measures is shown in Table IX, and these were then sorted by their linear/non-linear paradigms and their categorization with regards to monotonicity/non-monotonicity, such as shown in Table X; this builds upon the work from [67].

TABLE IX. VARIOUS MEASURES FOR MONOTONIC/NON-MONOTONIC AND LINEAR/NON-LINEAR PARADIGMS [67]

Measure	***Descriptor***
Distance Correlation Coefficient (dCor) [67]	"dCor is 'better at revealing complex…relationships…compared with other correlation metrics' by 'integrating both linear and non-linear dependence'" [71].
Hoeffding's D Correlation Coefficient (D) [67][68]	"D can reflect a certain degree of concordance and discordance."
Information Coefficient of Correlation (ICC) [69]	"ICC can provide a gauge of alignment between the posited and actual value."
Kendall's Tau Correlation Coefficient (tau) [67][70]	"Tau can illuminate correlations of significance when the distributions of the sample set and population are not necessarily known."
Maximal Correlation (MC) [69]	"MC pertains to transformations of the data, which are considered to maximize the correlation."
Maximum Information Coefficient (MIC) [69]	"MIC encompasses both linear and nonlinear correlations between the 'variable pairs'."
Mutual Information (MI) [69]	"MI is a paradigm, wherein one of the variables conveys a quantifiable amount of information about the other."
Pearson's [Product]-Moment Correlation Coefficient (PPMCC) [67]	"PPMCC measures the relationship strength and direction between the 'variable pairs'."
Percentage Bend Correlation Coefficient (PBCC) [67]	"PBCC refers to a paradigm, wherein a specified percentage of marginal observations deviating from the median are weighted downward" [72].
Spearman's Rho Correlation Coefficient (rho) [67][70]	"Rho scrutinizes the dependence between two random variables" [73].

TABLE X. EXEMPLAR USAGE OF VARIOUS MEASURES [67]

		Monotonic	***Non-monotonic***
Linear		D [67] rho [67] tau [67][70] PPMCC [67][69][70] PBCC [67] dCor [67]	N/A[2]
Non-linear		PPMCC[1][67][69]70] rho [67] tau [67][70] PBCC [67] dCor [67] D [67]	MC [69] dCor [67] D [67] PPMCC[3] [67][69] Rho[3] [67][69]
	Curvilinear	rho [67] PBCC [67] dCor [67] PPMCC[1][67] tau [67]	dCor [67] D [67] PPMCC[3] [69] Rho[3] [67][69]

[1] "Heuvel notes the efficacy of PPMCC with 'families of bivariate distribution functions with non-linear monotonic associations'" [70].

[2] "Technically, non-monotonic cannot be linear; however, as noted by Nicolaou, linear dynamics may experience transient segueing 'toward non-monotonic dynamics'" [74].
[3] "Of note, given symmetry, it does not 'find non-monotonic dependence'" [69].

The experimentation conducted yielded results that seemed to align with the findings of Mirtagioglu [67]. For example, the following seem to hold: (1) "in cases where there is no relationship (i.e., 0) between the variables" (e.g., non-functional relationship, wherein "there is no function of one variable that interacts with the other and vice versa"), dCor, D, tau, PPMCC, PBCC, rho, as well as MC "have given very satisfactory results," (2) "very low values (i.e., close to 0)" of rho, tau, PPMCC, and PBCC is emblematic of a "random relationship between the variables," and (3) "very low values (i.e., close to 0)" of tau, PPMCC, and PBCC, and rho when conjoined with "very high values (close to 1)" of dCor is emblematic of a non-monotonic relationship between/among variables, such as shown in Table XI [67][70].

TABLE XI. EXEMPLAR FINDINGS FROM MEASURES & POSITS [67]

Close to 0	***Close to 1***	***Close to -1***	***Relationship***
dCor, D, tau, PPMCC, PBCC, rho	N/A	N/A	None
rho, tau, PPMCC, PBCC	N/A	N/A	Random
N/A	rho, PPMCC	N/A	Strong Positive Monotonic
N/A	N/A	rho, PPMCC	Strong Negative Monotonic
tau, PPMCC, PBCC, rho,	dCor	N/A	Non-monotonic

The results also somewhat align with the findings of Rainio, and Heuvel. However, the rankings and sortings, such as offered by Mirtagioglu (M), Rainio (R), and Heuvel (H) somewhat differ, as shown in Table XII below. In terms of computational complexity, D is at O(n log n) while dCor is at $O(n^2)$ [75][76]; hence D is faster than dCor, and these are the validation measures for NMR. Moreover, D is less sensitive to outliers and can be well suited for ties. Next, rho is at O(n log n) while tau can be between O(n log n) and $O(n^2)$ [77][78][79]; hence, rho can be faster than tau for handling non-linear MR (however, tau tends to be less sensitive to outliers and can be well suited for ties). PBCC is considered to be at O(n log n), but it is subject to the involved sorting algorithms. MIC is at $n^{2.4}$[80]. Finally, PPMCC can be considered to be between O(n) and O(n log n). Hence, PPMCC can be faster than rho (in some cases) (however, rho tends to be less sensitive to outliers and has higher efficacy than PPMCC). Using a ROYG color-coding schema, the complexities are shown in color within Table XII, which stems from the work of [67].

TABLE XII. POSITED RANKING/SORTINGS BY M, R, AND H [67]

		M [68]	***R*** [70]	***H*** [71]
Monotonic	***Linear***	rho PBCC PPMCC dCor tau	PPMCC[1] rho[2] tau[2]	PPMCC MIC
	Non-linear	rho	rho	PPMCC
		PBCC dCor tau D	tau PPMCC	MIC
Non-monotonic	***Non-linear (e.g., curvilinear)***	dCor D	N/A	PPMCC[3] rho[3] MIC

[1] "more oriented for 'linear association'" [71].
[2] "more oriented for 'monotonic association'" [71].
[3] "however, this is N/A when the non-monotonic dependence is symmetric" [71].

Other Similarity Measures (SMs) include the Heterogeneous Euclidean-Overlap Metric (HEOM) and Value Difference Metric (VDM), which can accommodate situations with mixed numerical and categorical data [82][83]; for example, dimensions constitute numerical data, but colors are categorical data. Measures, such as these as well as others, will be considered in future work. For the experiments herein, the Table XIII complexity considerations for the SMs of Tables IX-XII were utilized. Then, those SMs that were the most computationally tractable (and more on the monotonic side), for both the initial foray and validation shown in Figure 1 (which builds upon the work of [67]), were arranged into Lower-Level Heuristic (LLH) amalgams, as shown in Table XIV.

TABLE XIII. COMPLEXITY CONSIDERATIONS FOR SMS

O(n)	***O(n)*** <=SM<= ***O(n log n)***	***O(n log n)***	***O(n log n)*** <=SM<= ***O(n2)****	***O(n2)****	***O(n2.4)****
	PPMCC	D rho PBCC	tau	dCor	MIC

*for sufficiently large values of n

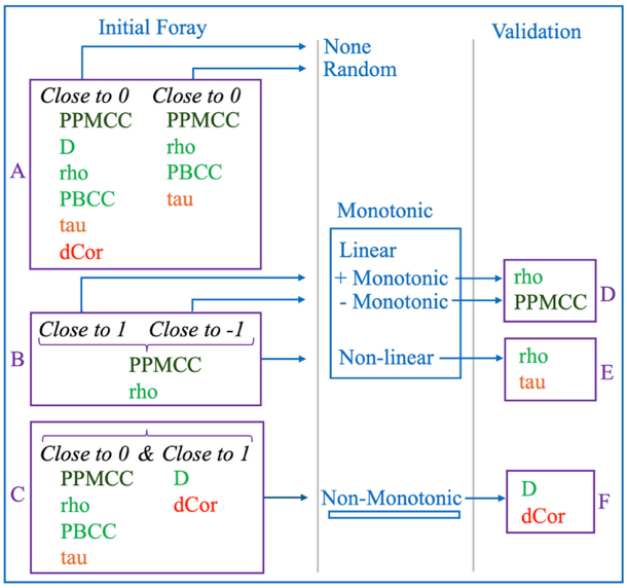


Fig. 1. Construct Utilized for Experimental Runs [67]

TABLE XIV. LLH AMALGAMS FOR EXPERIMENTAL RUNS

LLH A	***SM(s) A***	***LLH B***	***SM(s) B***	***LLH D***	***SM(s) D***
A1	PPMCC	B1	PPMCC	D1	PPMCC
A2	rho	B2	rho	D2	rho
A3	PBCC	***LLH C***	***SM(s) C***	***LLH E***	***SM(s) E***
A4	D-rho	C1	PPMCC-D	E1	rho

A5	D-PBCC	C2	rho-D	***LLH F***	***SM(s) F***
		C3	PBCC-D	F1	D

These LLHs are conjoined alongside the Hyper-Heuristics (HH) and Metaheuristics (MH) experimented with in prior work. This is shown in Figure 2, which is an extrapolation of Bouazza's work [84].

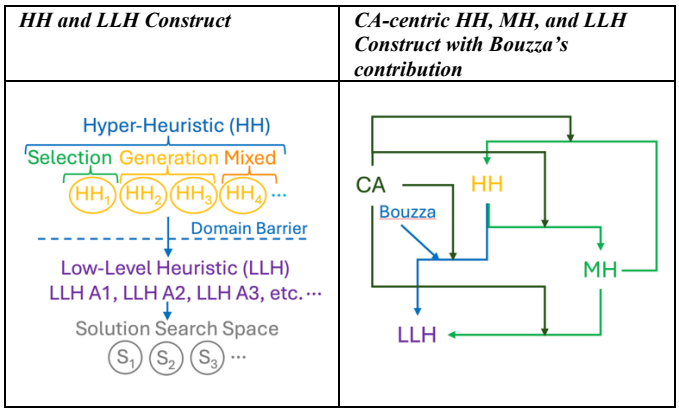


Fig. 2. Construct Utilized for Experimental Runs [85]

Experimentation was conducted on open LLM models, such as Mistral AI's Mixtral 8x7B (Apache 2.0 license) (LLM1), Berkeley's Neural Engineering Systems Technology's (NEST) Starling-LM-7B-Alpha (Apache 2.0 license with the additional condition that the model is not used to compete with OpenAI) (LLM2), and Mintplex Lab's AnythingLLM (MIT license) (LLM3, which used both LLM1 and LLM2). The relative performance from the experimental runs of Table IV and Figure 2 are shown in Figure 3.

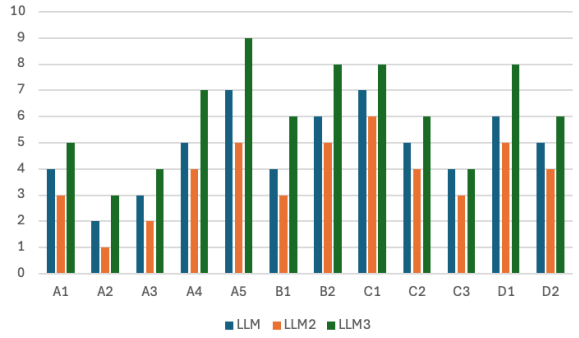


Fig. 3. LLH Performance Results from Experimental Runs

E1 and F1 were excluded, as they were solitary LLHs. LLM1 seemed to exhibit superior results to LLM2, and interestingly, LLM3 performed better than LLM1. LLM1's Sparse Mixture of Experts (SMoE) might have accounted for its performance, and LLM3's seemingly successful amalgamation of LLM1 and LLM2 warrants further investigation. The performance of LLM A4 and A5 was not surprising given D, and for this set of experimental runs, PBCC seemed to perform better than rho. In the case of the LLH C set, PPMCC seemed to perform better than rho and PBCC; along this vein, the LLH D set emulates this trend (however, LLH B set does not). Future work will involve further experimentation in this regard.

Then, as an extrapolation of Table IV, Table XIII is derived. Under an Uncompressed Decision Cycle (UDC) paradigm (more akin to System 2), DedR can be utilized. Under a Compressed Decision Cycle (CDC) paradigm (more akin to System 1), AbdR or IndR can be used; using a ROYG color coding schema, green has the fastest relative performance while red has the slowest relative performance. In essence, as DedR is more analytical, it tends to be slower; as IndR endeavors to establish a pattern, it tends to be slower than AbdR, which can be faster in putting forth an "inference to the best explanation." As IndR tends to be prevalent for CA, AnaR, CBR, and GBR are compared. As can be seen in Table XIII, given comparable performances, it seems prudent to opt for GBR given that it has a heightened probability of staying within the green MR zone.

TABLE XV. RMs AND RPs UNDER UDC AND CDC

UDC		***CDC***		
DedR	*MR*	IndR	*W MR*	*NMR*
		AnaR	*NMR*	
		CBR	*C MR*	*NMR*
		GBR	*MR*	*NMR*
		AbdR	*NMR*	

IV. DISCUSSION

This paper advances a design framework, wherein which GBR-facilitated CBR and order-theoretic formalisms (e.g., posets, isomorphisms, monotonicity) are leveraged to facilitate the assessment of coherence in LLM outputs. Among other aims, one research Line of Effort (LOE) explores whether model outputs can be mapped to structured representations whose relations are partially ordered (i.e., in a poset fashion), whereby desirable reasoning behavior can be delineated as monotone movement within that order and quantified with designated rank and/or dependence statistics.

The paper depicts how certain cases, exemplars, and their relations might be represented as nodes and order relations, and it leverages the use of correlation and/or rank concordance measures (e.g., Pearson, Spearman, Kendall, Hoeffding) as prospective diagnostics of monotonic progress. As a limitation (and intended as future work), this particular paper does not explicitly specify the operational pipeline; this will be presented in a follow-on piece of work that is already in-progress. That work will also clearly map the limitations of the various measures used. For example, Pearson correlation assumes linear dependence and is sensitive to scale, Spearman and Kendall are able to capture monotone associations, but do not suffice in affirming causal monotonicity in reasoning, and Hoeffding's measure can have high efficacy in some cases but is also data-hungry and can be unstable with small samples. As can be gleaned, this work-in-progress will contain a lengthy survey section with known limitations. Future work will be further discussed in Section V.

V. CONCLUDING REMARKS

RWS tend to be more in the NMR realm. Consequently, enforcing MR constraints to maintain coherence for CA is

challenging, and achieving RDM is non-trivial. This AI coherence issue tends to devolve to incidents of AI hallucinations, and despite various contemporary mitigation approaches (GSP, FTM, RAG, etc.), the problem seems to be, as reported in the literature, worsening as CA versions advance. In essence, as opined by a number of researchers in the arena, the treatment of the AI hallucination/AI coherence issue has been sub-optimal for various LLM-based CA systems; AI hallucinations, AI coherence, and validity seem to be linked to the state of the information-at-hand. For the case of the Known (e.g., KK, particularly HK, as contrasted to MK or WK), the validity seems to remain higher (relatively speaking). For the case of the Unknown (e.g., KU, UK, UU), the validity declines. Of course, the decline in validity is accompanied by a dramatic performance degradation with regards to AI coherence (and hallucinations). Generally speaking, CA conversations occurring in real-time, from the CA vantage point, tend to reside somewhere between CDC and UDC. Hence, the involved primary RM would likely tend more towards IndR (rather than AbdR or DedR, respectfully). Along this vein of IndR (and its subordinate oft used AnaR), GBR/CBR seems to be optimal. Moreover, in the context of CA, Zheng and others have long asserted: to "build emotional bond with users," more "advanced linguistic features" should be incorporated [86]. Within the literature, historical studies (e.g., Glucksberg, Kaall, Roberts) have shown "that figurative language…are key to interesting and engaging conversations" [86]. In addition, Hofstadter has argued that analogy is the "core of cognition," and Holyoak seems to affirm [87][88]. Hofstadter further states, "without concepts there can be no thought, and without analogies there can be no concepts" [89]. At its core, the conveyance of related/similar concepts/notions is central for RDM. Accordingly, the secondary RMs of CBR and GBR become significant; after all, LLMs are "optimally suited at learning by example," and the issue then becomes how to best operationalize matters. While there is of course, interest in assessing the involved computational complexity, the goal of staying as much in the MR realm (as opposed to NMR), amidst computational practicality, rises to the forefront (and the contextualization of the MNTZ becomes crucial). The experimentation within demonstrates that there are some heuristic amalgams (involving certain similarity measures) that seem to be apropos for this goal. Future work will involve more quantitative and qualitative experimentation in addition to that proposed in Section IIIB.

REFERENCES

[1] S. Chan, "Treatment of the Multi-Attribute Decision-Making Rank Reversal Problem for Real-World Systems," *The Seventeenth International Conference on Future Computational Technologies and Applications (Future Computing)*, Apr. 2025, pp. 1-10.

[2] "Tongyi Qianwen (Qwen)," *Alibaba Cloud*, [Online]. Accessed: Sep. 5, 2025. Available: https://www.alibabacloud.com/en/solutions/generative-ai/qwen?_p_lc=1.

[3] A. Kalai, O. Nachum, S. Vempala, and E. Zhang, "Why Language Models Hallucinate," *Arxiv.org*, Sep. 4, 2025 [Online]. Accessed: Sep. 5, 2025. Available: https://arxiv.org/abs/2509.04664.

[4] "What are AI hallucinations," *IBM*, Sep. 1, 2023. [Online]. Accessed: Sep. 1, 2025. Available: https://www.ibm.com/think/topics/ai-hallucinations.

[5] "Global Conversational AI Market Size & Outlook, 2024-2030," *Grand View Horizon*, [Online]. Accessed: Sep. 1, 2025, Available: https://www.grandviewresearch.com/horizon/outlook/conversational-ai-market-size/global.

[6] "Conversational AI Market worth $49.80 billion by 2031," *Markets and Markets*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.marketsandmarkets.com/PressReleases/conversational-ai.asp.

[7] "Conversational AI Market Size," *Fortune Business Insights*, Aug. 25, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://www.fortunebusinessinsights.com/conversational-ai-market-109850.

[8] "Conversational AI Platform Software Market Size and Forecast," *Verified Market Research*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.verifiedmarketresearch.com/product/conversational-ai-platform-software-market/.

[9] K. Irwin, "Tim Cook: Apple Intelligence Isn't Immune to Hallucinations," *PC Magazine*, Jun. 11, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://www.pcmag.com/news/tim-cook-apple-intelligence-isnt-immune-to-hallucinations-wwdc.

[10] W. Gallagher, "Siri Chatbot prototype nears ChatGPT quality, but hallucinates more than Apple wants," *AppleInsider*, Jun. 1, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://appleinsider.com/articles/25/06/01/siri-chatbot-prototype-nears-chatgpt-quality-but-hallucinates-more-than-apple-wants.

[11] S. Athaluri et al., "Exploring the Boundaries of Reality: Investigating the Phenomenon of Artifical Intelligence Hallucination in Scientific Writing Through ChatGPT References," *Cureus*, vol. 15, pp. 1-5.

[12] M. Zeff, "OpenAI's new reasoning AI models hallucinate more," *TechCrunch*, Apr. 18, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/.

[13] C. Mauran, "OpenAI's o3 and o4-mini hallucinate way higher than previous models," *Mashable*, Apr. 19, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://mashable.com/article/openai-o3-o4-mini-hallucinate-higher-previous-models.

[14] "OpenAI o3 and o4-mini System Card," *OpenAI,* Apr. 16, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

[15] K. Anish, "Getting the Most from Gemini: Understanding its Knowledge and Creativity," *Gemini Apps Help*, Nov. 26, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://support.google.com/gemini/community-guide/309961349/getting-the-most-from-gemini-understanding-its-knowledge-and-creativity?hl=en.

[16] I. Khan, "Google Gemini Chatbot Review: Hallucination Station," *Cnet*, Apr. 2, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://www.cnet.com/tech/services-and-software/google-gemini-chatbot-review-hallucination-station/.

[17] A. Zilber, "Google's AI is 'hallucinating,' spreading dangerous info — including a suggestion to add glue to pizza sauce," *New York Post*, Jun. 6, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://nypost.com/2025/06/06/business/googles-ai-overviews-are-hallucinating-by-spreading-false-info/.

[18] M. Dora, "Giskard announces Phare, a new open & multi-lingual LLM Benchmark," *Giskard*, [Online]. Accessed: Sep. 1., 2025, Available: https://www.giskard.ai/knowledge/giskard-announces-phare-a-new-llm-evaluation-benchmark#:~:text=This%20new%20LLM%20benchmark%2C%20called,samples%20will%20be%20open%2Dsource.

[19] "Phare: A Safety Probe for Large Language Models," *Hugging Face*, May 16, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://huggingface.co/papers/2505.11365.

[20] P. Jeune, B. Malezieux, W. Xiao, and M. Dora, "Phare: A Safety Probe for Large Language Models," *Arxiv.org*, May 26, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/abs/2505.11365.

[21] A. Weissberger, "Sources: AI is Getting Smarter, but Hallucinations Are Getting Worse," *IEEE ComSoc Technol.*, May 10, 2025. [Online]. Accessed: Sep. 1, 2025. Available: https://techblog.comsoc.org/2025/05/10/nyt-ai-is-getting-smarter-but-hallucinations-are-getting-worse/.

[22] P. Jeune, J. Liu, L. Rossi, and M. Dora, "RealHarm: A Collection of Real-World Language Model Application Failures," *RealHarm Dataset*, [Online]. Accessed: Sep 4, 2025. Available: https://realharm.giskard.ai/.

[23] P. Jeune, J. Liu, L. Rossi, and M. Dora, "RealHarm: A Collection of Real-World Language Model Application Failures," *Proc. of the The First Workshop on LLM Security (LLMSEC)*, Aug. 2025, pp. 87-100.

[24] "Why language models hallucinate," *OpenAI*, Sep. 5, 2025. [Online]. Accessed: Sep. 5, 2025. Available: https://openai.com/index/why-language-models-hallucinate/.

[25] A. Holzinger, K. Zatloukal, and H. Muller, "Is human oversight to AI systems still possible?" *New Biotechnology*, vol. 85, pp. 59-62, Mar. 2025.

[26] A. Cunningham, ""Do not hallucinate": Testers find prompts meant to keep Apple Intelligence on the rails," *ArsTechnica*, Aug. 6, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arstechnica.com/gadgets/2024/08/do-not-hallucinate-testers-find-prompts-meant-to-keep-apple-intelligence-on-the-rails/.

[27] R. Ziv, "Developing Hallucination Guardrails," *OpenAI Cookbook*, May 29, 2024. [Online] Accessed: Sep. 1, 2025, Available: https://cookbook.openai.com/examples/developing_hallucination_guardrails.

[28] A. Gutpa, "Use Guardrails to prevent hallucinations in generative AI applications," *AWS*, Jul. 11, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://builder.aws.com/content/2i12ntqFx3xAaDLfvrjH7278sEW/use-guardrails-to-prevent-hallucinations-in-generative-ai-applications.

[29] "Domain specific knowledge and skills," *ScienceDirect*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.sciencedirect.com/topics/psychology/domain-knowledge.

[30] A. Huaulme, "Surgical declarative knowledge learning: concept and acceptability study," *Comput. Assisted Surgery*, vol. 27, pp. 74-83, 2022.

[31] L. Fisher, B. Halima, and K. Yerian, "Procedural and Declarative Knowledge," *Learning How to Learn Languages*, 2024 [Online]. Accessed: Sep. 5, 2025. Available: https://opentext.uoregon.edu/languagelearningedition1/chapter/procedural-and-declarative-knowledge/.

[32] I. Zsigmond, "Role of Conditional Knowledge in Conscious Reading: The Integrating Model of Metacognition," *Proc. of the 16th European Conf. on Reading and 1st Ibero-American Forum on Literacies*, Jan. 2009, pp. 1-8.

[33] M. Shaker and M. Moore-Clingenpeel, "The known knowns, known unknowns, and unknown unknowns of surveys and sleep," *Ann. Allergy Asthma Immunol*., vol. 129, pp. 669-670, Dec. 2022.

[34] S. McGregor, "Learning with Donald Rumsfeld – flexible learning: the relevance and resonance of multiprofessional learning in primary care," *Br. J. Gen Pract*., vol. 54, pp. 722-723, Sep. 2004.

[35] Z. Gekhman et al., "Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?" *Proc. of the 2024 Conf. on Empirical Methods in Natural Lang. Process*., Nov. 2024, pp. 7765-7784.

[36] S. Kadavath et al., "Language Models (Mostly) Know What They Know," *Arxiv.org*, Nov. 21, 2022. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/pdf/2207.05221.

[37] P. Mankul, A. Liusie, M. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," *Proc. of the 2023 Conf. on Empirical Methods in Natural Lang.Process.*, Dec. 2023, pp. 9004-9017.

[38] O. Elisha, "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs," *Arxiv.org*, Jan. 2024. [Online]. Accessed: Sep. 14, 2025. Available: https://arxiv.org/pdf/2312.05934.

[39] J. Sun et al., "Dial-insight: Fine-tuning Large Language Models with High-Quality Domain-Specific Data Preventing Capability Collapse," *Arxiv.org*, Mar. 14, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/abs/2403.09167.

[40] S. Grote-Garcia, "Deductive Reasoning," in *Encyclopedia of Child Behavior and Development*. Boston, MA. Springer, 2011, pp. 477-478.

[41] "Inductive Reasoning," *University of Illinois Springfield, ION Professional eLearning Programs*, [Online]. Accessed: Sep. 1, 2025. Available: https://www.uis.edu/ion/resources/oiai/inductive-reasoning.

[42] G. Minnameier, "Abduction, Induction, and Analogy," in *Model-Based Reasoning in Science and Technology. Studies in Computational Intelligence*, vol. 314, pp. 107-119, 2010.

[43] K. Holyoak, N. Ichien, and H. Lu, "Analogy and the Generation of Ideas," *Creativity Research Journal*, vol. 36, pp. 532-543, Jul. 2023.

[44] D. Gentner and L. Smith, "Analogical Reasoning," in *Encyclopedia of Human Behavior*. Oxford, UK. Elsevier, 2012, pp. 130-136.

[45] A. Sandoval-Hernandez and D. Rutkowski, "Embracing complexity: abductive reasoning as a versatile tool for analyzing international large-scale assessments," in *Educ. Assessment Eval. And Accountability*, vol. 37, pp. 255-271, Dec. 2024.

[46] A. Yan and Z. Cheng, "A Review of the Development and Future Challenges of Case-Based Reasoning," *Appl. Sci.*, vol. 14, pp. 1-22, Aug. 2024.

[47] "Case-based reasoning," *Taylor & Francis*, [Online]. Accessed: Sep. 1, 2025. Available: https://taylorandfrancis.com/knowledge/Engineering_and_technology/Artificial_intelligence/Case-based_reasoning/.

[48] J. Kolodner, "Improving Human Decision Making through Case-Based Decision Aiding," *AI Magazine*, vol. 12, pp. 52-68, 1991.

[49] S. Chan, "Interstitial b-SHAP-Owen Amalgam for the Enhancement of Artificial Intelligence System-Centric Sequential Decision-Making," *International Journal on Advances in Intelligent System*s, v18, in press.

[50] "How to Evaluate AI Chats Using Conversation Coherence Evaluator," *Athina AI*, Apr. 17, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://blog.athina.ai/how-to-evaluate-ai-chats-using-conversation-coherence-evaluator.

[51] "Athina-ai/athina-evals," *Github*, [Online]. Accessed: Sep. 1, 2025. Available: https://github.com/athina-ai/athina-evals/blob/main/examples/conversation_coherence.ipynb.

[52] K. Jantke, "Monotonic and non-monotonic inductive inference of functions and patterns," *Lecture Notes in Computer Science*, vol. 543, pp. 161-177, Jan. 2005.

[53] G. Paulino-Passos and F. Toni, "Monotonicity and Noise-Tolerance in Case-Based Reasoning with Abstract Argumentation," *Proc. of the 18th Int. Conf. on Princ. of Knowl. Representation and Reasoning*, Nov. 2021, pp. 508-518.

[54] B. Chen, R. Saetre, and Y. Miyao, "A Comprehensive Evaluation of Inductive Reasoning Capabilities and Problem Solving in Large Models," *Findings of the Assoc. for Comput. Linguistics*, pp. 323-339, Mar. 2024.

[55] M. Luo et al., "Towards LogiGLUE: A Brief Survey and A Benchmark for Analyzing Logical Reasoning Capabilities of Language Models," *Arxiv.org*, Mar. 31, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/html/2310.00836v3.

[56] K. Cheng, "Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs," *Arxiv.org*, Aug. 7, 2024. [Online]. Accessed: Sep. 5, 2025. Available: https://arxiv.org/abs/2408.00114.

[57] L. Eliot, "On Whether Generative AI And Large Language Models Are Better At Inductive Reasoning Or Deductive Reasoning And What This Foretells About The Future Of AI," *Forbes*, Aug. 11, 2024, [Online]. Accessed: Sep. 1, 2025. Available: https://www.forbes.com/sites/lanceeliot/2024/08/11/on-whether-generative-ai-and-large-language-models-are-better-at-inductive-reasoning-or-deductive-reasoning-and-what-this-foretells-about-the-future-of-ai/.

[58] B. Dickson, "LLMs excel at inductive reasoning but struggle with deductive tasks, new research shows," *Venture Beat*, Aug 15, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://venturebeat.com/ai/llms-excel-at-inductive-reasoning-but-struggle-with-deductive-tasks-new-research-shows.

[59] D. Leake and D. Crandall, "On Bringing Case-Based Reasoning Methodology to Deep Learning," *Case-Based Reasoning Methodology to Deep Learning. Lecture Notes in Computer Science*, vol. 12311, pp. 343-348, Oct. 2020.

[60] C. Qin et al., "Relevant or Random: Can LLMs Truly Perform Analogical Reasoning?" *Findings of the Assoc. for Comput. Linguistics*, Jul. 2025, pp. 23993-24010.

[61] M. Voskoglou and A. Salem, "Analogy-Based and Case-Based Reasoning: Two sides of the same coin," *Int. J. of Appl. Of Fuzzy Sets and Artif. Intell*., vol. 4, pp. 5-51, 2014.

[62] N. Lemons, B. Hu, W. Hlavacek, "Hierarchical graphs for rule-based modeling of biochemical systems," *BMC Bioinformatics*, vol. 12, pp. 1-13, Feb. 2011.

[63] J. Li, X. Luo, and G. Lu, "GS-CBR-KBQA: Graph-structured case-based reasoning for knowledge base question answering," *Expert Syst. with Appl.*, vol. 257, pp. 1-37, Dec. 2024.

[64] H. Xu, Y. Wei, Y. Cai, and B. Xing, Knowledge graph and CBR-based approach for automated analysis of bridge operational accidents: Case representation and retrieval," *PLOS One*, Nov. 2023, pp. 1-21.

[65] D. Casarett, "Can Metaphors and Analogies Improve Communication with Seriously Ill Patients," *J. Palliat Med*., vol. 13, pp. 1-6, Mar. 2010.

[66] R. Kanthan and S. Mills, "Using Metaphors, Analogies and Similes as Aids in Teaching Pathology to Medical Students," *Medical Sci. Educ*., vol. 16, pp. 102-116, Dec. 2017.

[67] S. Chan, "A Prospective Monotonic/Non-Monotonic Transition Zone Impediment for Concept Model-Centric Artificial Intelligence Systems," *The Second International Conference on AI-based Systems and Services (AISys)*, in press.

[68] H. Mirtagioglu and M. Mendes, "On Monotonic Relationships," *Biostatistics and Biometrics*, vol. 10, pp. 1-11, May 2022.

[69] A. Fujita, J. Sato, M. Demasi, "Comparing Pearson, Spearman, and Hoeffding's D measure for gene expression association analysis," *J. of Bioinformatics and Comput. Biology*, vol. 7, pp. 663-684, Sep. 2009.

[70] O. Rainio, "Different Coefficients for Studying Dependence," *The Indian J. of Stat*., vol. 84-B, pp. 895-914, Nov. 2022.

[71] E. Heuvel and Z. Zhan, "Myths About Linear and Monotonic Associations: Pearson's r, Spearman's ρ, and Kendall's τ," *The Amer. Stat.*, vol. 76, pp. 44-52, Nov. 2021.

[72] J. Hou, et al., "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, pp. 1-24, Feb 2022.

[73] C. Pernet, R. Wilcox, and G. Rousselet, "Robust correlation analyses: false positive and power validation using a new open source Matlab toolbox," *Front. Psychol*., vol. 3, pp. 1-18, Jan. 2013.

[74] Y. Dodge, "Spearman Rank Correlation Coefficient," In: *The Concise Encyclopedia of Statist.* New York, NY: Springer, 2008, pp. 502-505.

[75] Z. Nicolaou, T. Nishikawa, S. Nicholson, J. Green, and A. Motter, "Non-normality and non-monotonic dynamics in complex reaction networks," *Phys. Rev. Res. 2*, pp. 1-15, Oct. 2020.

[76] C. Even-Zohar, "Independence: Fast Rank Tests," *Arxiv.org*, Oct. 2020. [Online]. Accessed: Sep. 5, 2025. Available: https://arxiv.org/pdf/2010.09712.

[77] J. Hou et al., "Distance correlation application to gene co-expression network analysis," *BMC Bioinformatics*, vol. 23, pp. 1-24, Feb. 2022.

[78] M. Stephanou and M. Varughese, "Sequential estimation of Spearman rank correlation using Hermite series estimators," *Arxiv.org*, Jul. 2021. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/pdf/2012.06287.

[79] M. Sepulveda, "Kendallknight: An R Package for Efficient Implementation of Kendall's Correlation Coefficient Computation," *Arxiv.org*, Dec. 8, 2024. [Online]. Accessed: Sep. 1, 2025. Available: https://arxiv.org/pdf/2408.09618.

[80] D. Christensen, "Fast algorithms for the calculation of Kendall's τ," *Comput. Stat*., vol. 201, pp. 51-62, Mar. 2005.

[81] F. Shao and H. Liu, "The Theoretical and Experimental Analysis of the Maximal Information Coefficient Approximate Algorithm," *De Gruyter*, Mar. 2021, pp. 1-10.

[82] C. Li and H. Li, "Correlation weighted heterogeneous euclidean-overlap metric," *Int. J. of Comput*., vol. 33, pp. 341-346, Jul. 2015.

[83] L. Jiang and C. Li, "Two improved attribute weighting schemes for value difference metric, " *Knowledge and Inf. Syst*., vol. 60, pp. 949-970, 2019.

[84] W. Bouazza, "Machine Learning-Based Hyper-Heuristics: A Clear Insight," *Proc. of the 2024 7th Int. Conf. on Comput. Intell. And Intell. Syst. (CIIC)*, Feb. 2025, pp. 29-37.

[85] S. Chan, "AI-Centric Hyper-Heuristic and Self-Exclusion Mechanism for the Updating of a Heuristic," *IEEE World AI IoT Congress (AIIoT)*, May 2025, pp. 0536-0545.

[86] D. Zheng, R. Song, T. Hu, H. Fu, J. Zhou, "Love is as Complex as Math: Metaphor Generation System for Social Chatbox," *Chinese Lexical Semantics. Lecture Notes in Comput. Sci.*, vol. 11831, pp. 337-347, Jan. 2020.

[87] D. Hofstadter, "Epilogue: Analogy as the Core of Cognition," in *The Analogical Mind*. Cambridge, MA. MIT Press, 2001.

[88] K. Holyoak, "The Human Edge: Analogy and the Roots of Creative Intelligence," Cambridge, MA. MIT Press, pp. 39-41, 2025.

[89] D. Hofstadter and E. Sanger, "Surfaces and essences: Analogy as the fuel and fire of thinking," *Basic Books*, pp. 1-578, April 2013.