Prototype Pseudo-Metacognition Module for AI Reasoning

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

***Abstract*—A prototype pseudo-Metacognition Module (MM) is presented. It consists of various presets, which have been enhanced for this paper, as well as a newly presented construct. Architecturally, this bespoke MM diverges from a prototypical quasi-Mixture of Experts (MoE) approach (e.g., a multi-layered LLM atop LLM atop LLM-type approach). Rather, it incorporates the principles of the Knudsen and Konishi findings and utilizes an embodiment approach; instead of a component by component build, it endeavors to weave in considerations for a more intrinsic construct. A Watrobski-inspired test dataset is formulated, and an Dong-inspired ablation study is conducted. Preliminary findings indicate that the proposed MM might actually be more useful than not.**

Keywords-Conversational AI Agent (CAA), Case-Based Reasoning (CBR), Metacognition Module (MM).

I. Introduction

A Conversational AI Agent (CAA) is considered to be a Real-World Scenario (RWS) application for which Robust Dialogue Management (RDM) is needed. RDM involves, among other facets, real-time reasoning and decision-making for a naturally flowing, coherent conversation. For the most part, CAA are rooted in the use of Large Language Models (LLMs). While computer scientists, such as Chen, assert that certain Reasoning Mechanisms (RMs), such as Inductive Reasoning (IndR) are central to CAA LLMs, in Chen's study, it was found that the performance was often, ironically, sub-optimal for IndR [1]. Along this vein, Luo's study involving the LogiGlue Benchmark, which consists of 24 Deductive Reasoning (DedR), IndR, and Abductive Reasoning (AbdR) datasets, revealed that, interestingly, "LLMs excel most in" AbdR "followed by" DedR "while they are least effective" for IndR [2]. Cheng put it succinctly; "while the" DedR "capabilities of LLMs…have received considerable attention, their abilities in" IndR "remain largely unexplored" [3]. This is also true for the enhancements to LLMs, such as Large Reasoning Models (LRMs). Hence, one of the goals of this paper is to delve into the CAA IndR matter further with the proposed MM. The paper also sets out to delineate an "embodiment" architecture (vs. a layered Mixture of Experts or MOE-style stack) that leverages Lower Ambiguity, Higher Uncertainty (LAHU) and Higher Ambiguity, Lower Uncertainty (HALU) framing to manage ambiguity vs. uncertainty under compressed vs. uncompressed decision cycles. The paper further posits a "prototype pseudo–Metacognition Module (MM)" intended to improve RDM for CAA by explicitly orchestrating reasoning mechanisms (DedR/IndR/AbdR) and related processes (monotonic vs non-monotonic), with an emphasis on Case-Based Reasoning (CBR) and Graph-Based Reasoning (GBR) as computationally tractable bridges for probabilistic and temporal reasoning.

Section I provides an overview regarding the challenges and complexities surrounding RDM, which is a core capability needed for CAA. The remainder of the paper is organized as follows. Section II provides pertinent background information pertaining to several CAA technical challenges and alludes to the notion that underlying architectural issues may need to be examined. An overview of various reasoning mechanisms/processes is provided, and those that are most suitable (e.g., reasonable computational tractability) for CAA are discussed. The section wraps up with an evaluation of whether a semblance of "commonsense reasoning" (for the CAA) can be achieved with the proposed MM. Section III revisits the need for an MM, puts forth some theoretical foundations, proposes an architecture, and delineates the test dataset and ablation study used for the experimentation. Section IV summarizes with concluding remarks, and proposed future work closes the paper.

II. Background

CAA are beset by a variety of technical challenges, which include among others, goals processing (i.e., intent recognition), context persistence (particularly for multi-turn conversations/multi-session dialogues), validity (pertaining to the accuracy of the information conveyed), and the follow-on notion of construct validity (i.e., the ability of a benchmark to gauge the adequacy by which a notion is captured/expressed). While much research has focused on the technical challenges, individually (e.g., those presented in Section I), in many instances, far less attention has been paid to the prospective underlying issue(s) (e.g., the underlying CAA architectural approach, the heuristical schema underpinning the CAA architecture, etc.), which may profoundly impact the aforementioned challenges. Taking one specialty domain as an example, the medical field, in a study by Casarett, it was ascertained that physicians had used metaphors in 64% and analogies in 31% of their conversations with patients [4]. Johnston affirms the efficacy of figurative language for triggering "multiple learning pathways" and facilitating clearer medical communication [5]. By way of background information, a metaphor tends to be implicit/direct while an analogy is explicit/extended; the amalgam constitutes substantial coverage for the involved domain knowledge communication channel. Accordingly, this begets the question as to whether, say, a Large Concept

Model (LCM) architectural approach might be a better approach than a LLM or LRM for figurative language (e.g., metaphors, analogies) since LCMs are more concept-centric rather than the token-centric (e.g., words, subwords, characters) LLMs and even LRMs (with their lengthier sequences of tokens or "chains of thought"). In a number of cases, LLMs have been known for sub-optimal performance with regards to intent, context, and validity; to underscore this point, Hussain finds that current LLM approaches (e.g., safety) can be sub-optimal in gleaning user intent; Du finds that even context length may adversely affect LLM performance, and in his study of "5 open- and closed-source LLMs," performance degradation was in the range of "13.9%-85%;" and Bean's study of 445 LLM benchmarks determined that there was a significant gap between empirical test results and the target phenomena they were supposed to measure (i.e., low construct validity) [6][7][8].

On the surface, it seems that simply shifting to an LCM-centric architecture may, perhaps, address the referenced challenges of discerning intent, maintaining context, and preserving validity. Intuitively, the LCM higher-level concept embeddings seem to be more representative than the LLM token-centric contextual embeddings. This then begets certain questions regarding the ensuing layer, such as the architecture underlying the LCM's causal graph approach (which is known to be, potentially, more advantageous for zero-shot and multi-modal taskings) as contrasted to LLM's correlation (i.e., pattern-matching) approach [9][10]. In either case, the RMs and Reasoning Processes (RPs) involved should be better understood; some of these primary RMs, among others, include: (1) DedR, (2) Probabilistic Reasoning (ProbR), (3) Temporal Reasoning (TempR), (4) IndR, (5) Analogical Reasoning (AnaR), and (6) AbdR. Some of these primary RMs may also leverage secondary RMs, such as Case-Based Reasoning (CBR) and Graph-Based Reasoning (GBR); the latter is especially important for the previously referenced causal graph approach.

A. Overview of RMs & RPs

Initially, the RMs focused upon will be DedR, IndR (which includes AnaR, CBR, and GBR), and AbdR. Grote-Garcia describes DedR as "the process of using general premises to draw specific conclusions" (i.e., a top-down paradigm) [11]. In contrast, Davidson describes IndR as a process wherein "specific observations are often used to draw general conclusions" (i.e., a "bottom-up" paradigm) [12]. Gentner describes AnaR as "the ability to perceive and use relational similarity between two situations or events," and Smaling notes that AnaR can be considered to be hierarchically situated below IndR (however, AnaR is situated above AbdR, which is a bit more nebulous and incomplete) [13][14]. Sandoval-Hernandez describes AbdR as the ability to move from "puzzling observations" to "inferring the most likely explanations," Thagard describes AbdR as "explanatory hypotheses" that "are formed and evaluated," and Belzen depicts AbdR as "explaining a phenomenon by a cause" [15][16][17]. Kolodneer notes that situated below AnaR (but above AbdR) is CBR, which Momem describes as utilizing "the knowledge obtained in past situations, referred to as cases, to solve new problems" [18][19]. In turn, Das and others note that GBR, which Zhang describes as "exploring the relationships between nodes and edges in a graph and making inferences based on these relationships," can be a form of CBR [20][21].

For CAA, the involved primary RMs, by validity ranking, are likely to be DedR, IndR, and then AbdR. As Boger and Cheng remind us, Aristotle's DedR leads to conclusions that are *definitively true*, if the premises and argument are valid [22][23]. Then, as Glass reminds us, Bacon's IndR empirical method, while *probable*, can indeed lead to conclusions that might be either true or false (e.g., there could be experimental discrepancies) [24]. Next, as Lu reminds us, hypotheses (both hypothesis generation and evaluation) form the core of Peirce's AbdR—"intelligent guessing"—but they are simply *possible* outcomes and can be either true or false [25]. However, there is also a temporal facet that is shaped by "what" and "when" information becomes available. Along this vein, generally speaking, DedR is considered to, generally, accompany the rubric of Uncompressed Decision Cycles (UDC) while IndR (as well as AnaR, CBR, and GBR) and AbdR are considered to, generally, accompany the rubric of Compressed Decision Cycles (CDC). Hence, apart from the ideal/desired validity requirement (e.g., guaranteed, probable, possible), in actuality, depending upon the amount of time available (i.e., UDC, CDC), a particular RM may be more apropos, and this is delineated in Table I by columns 2 and 3.

TABLE I. RM, VALIDITY, AND TEMPORAL SPAN

RM	***Validity***	***Temporal Span***	
DedR	*Guaranteed to be true*, if the premises and argument are valid.	Typically back-loaded, as it unfolds ineratively in a "bottom-up" fashion.	UDC
IndR (can include AnaR, CBR, and GBR)	*Likely to be true*, but it could be false despite the observations being accurate.	Typically front-loaded, but it can also unfold in a "bottom-up" fashion.	CDC
AbdR	*Can be true* (but might not be), as it involves a plausible best guess approximation or a posit as to the optimal explanation.	Typically front-loaded, but as it has various sensitivites (e.g., uncertainty/information gaps/ambiguity/ multiplicity, etc.) it can also unfold in a "bottom-up" fashion.	

The referenced RMs can also be sorted by the RPs of Monotonic Reasoning (MR) and Non-Monotonic Reasoning (NMR). Taking the logic of Xiu, MR can lead to conclusions that become invalid over time, as it is unable to accommodate new evidence [26]. In contrast, Brewka notes that NMR allows for modification and/or "retraction of prior conclusions" [27]. With regards to DedR, Bundy and Wallen note "the *monotonicity* of deductive logic," wherein "the addition of new axioms to a set of axioms can never decrease the set of theorems or facts" [28]. Fuhrmann affirms by noting that "deductive inference, at least according to the canons of classical logic, is *monotonic*; if a conclusion is reached on the basis of a certain set of premises, then that conclusion still holds if more premises are added" [29]. Continuing on, IndR can be construed to be *non-monotonic* as well as *weakly monotonic*. By way of example, with regards to IndR, Janke describes how *non-monotonic* reasoning "is inherently required in several approaches to inductive inference," and how IndR can also be *weakly monotonic* [30]. Proceeding to AnaR, as it is situated below IndR, Kerber asserts that AnaR is *non-monotonic*, and Passos and Amgoud, respectively, assert that CBR can be both *cautiously monotonic* and *non-monotonic* [31][32][33]. Next, with regards to AbdR, Hentenryck notes that AbdR is "closely related to *non-monotonic* reasoning" and is "a form of reasoning appropriate for handling incomplete information" [34]. Paul affirms by noting that "abduction is a form of *non-monotonic* reasoning" [35]. In the context of RWS, the RMs can be organized by their varying temporal constraints (e.g., UDC, CDC) as well hierarchically sorted by their associated RPs (e.g., MR, NMR). The UDC and CDC facets can also be further coupled with the notions of Higher Ambiguity, Lower Uncertainty (HALU) and Lower Ambiguity, Higher Uncertainty (LAHU), such as described in Table II.

TABLE II. LAHU/HALU MODULE (LHM)

Ambiguity/ Uncertainty	***Descriptor***
Higher Ambiguity, Lower Uncertainty (HALU)	Under a paradigm of UDC, and for the situation wherein prior cases do *not* exist (i.e., a paradigm of higher ambiguity), there needs to be a proactive seeking of more data (since time is readily available under UDC) so as "to lower uncertainty" and move towards a more acceptable state (i.e., a paradigm of lower uncertainty) [36][37].
Lower Ambiguity, Higher Uncertainty (LAHU)	Under a paradigm of CDC, and for the situation wherein prior cases do indeed exist (i.e., a paradigm of lower ambiguity), there is more tolerance for sparse data/no data (a paradigm of "higher uncertainty"), particularly if time is of the essence) [36][37].

Building upon Table II, Table III can be constructed. A Red-Orange-Yellow-Green (ROYG) color coding schema is utilized, wherein green denotes the best performance (either validity or computational performance) while red indicates the worst performance. For example, with regards to computational performance, MR is indicated by green, NMR by red, weak MR (W MR) by orange, and cautious MR (C MR) by yellow.

TABLE III. RMS AND RPS UNDER UDC AND CDC (WITH ZERO-SHOT CIRCUMSTANCES)

Validity	***HALU-centric***		***Validity***	***LAHU-centric***		
	UDC			***CDC***		
Guaranteed	DedR	*MR*	Probable	IndR	*W MR*	*NMR*
				AnaR	*NMR*	
				CBR	*C MR*	*NMR*
				GBR	*MR*	*NMR*
			Possible	AbdR		*NMR*

It should be noted that Table III depicts a specific instance/circumstance, wherein the paradigm is zero-shot (i.e., prior cases do *not* exist). Under few-shot circumstances, Table III can take a different form, as the degree of being HALU-centric/LAHU-centric can change temporally. As shown by Table III, as DedR is more analytical, its computational performance tends to be slower. Since IndR is focused upon establishing specific patterns, its tends to be slower than AbdR, which can be somewhat faster by simply putting forth a "best guess." As noted by Chen, IndR tends to be prevalent for RWS applications, such as CAA [38]; accordingly, the hierarchical subordinates of IndR (e.g., AnaR, CBR, and GBR—which Castaneda would construe as ProbR, as they involve "the retrieval of prior knowledge" for Momem's "cases"—are scrutinized/compared [19][39]. The "cases" can also involve what Xiong describes as TempR paradigms: "sequencing, duration, frequency, simultaneity, temporal relation, comparative analysis and facts extraction" [40]. For example, an aberration that occured with simultaneity (with comparable duration, frequency, etc.) in several different geographic regions might constitute a prior Indicators & Warnings (I&W) "case." Leeuwenberg would likely affirm this vantage point, as he posits that TempR is "the process of combining different temporal cues into a coherent temporal view" [41]. Cai cautions that employing TempR over sparse/incomplete and/or ambiguous/uncertain can be problematic [42]; the author has previously noted this as well: "for CAA, conversational coherence is 'quite difficult to maintain because the information supply changes temporally, and at some points, it may be sparse/incomplete and/or ambiguous/uncertain'" [43]. Accordingly, LAHU/HALU is utilized to mitigate against some of Cai's concerns by well considering the issues surrounding TempR; also, AnaR, CBR, and GBR serve to capture the essence of ProbR. In this way, the RMs previously referenced in the overview of Section II (as well as the opening thread of Section II) have all been addressed.

B. *A Winnowing of the RMs/RPs for CAA*

Pertaining to the CAA's RDM, as the discussion topics and prioritization of the thematics may shift (and as incoming information may potentially conflict with prior information), the handling of MR and NMR becomes crucial for maintaining context, validity, and intent. After all, MR can be brittle, as it is unable to revise prior conclusions in the face of new, contradictory information; in contrast, NMR can be somewhat more resilient, as it is able to retract and adapt. Yet, while it may be better suited for RDM, the computational requirements for NMR can be intractable; fortunately, as can be gleaned from Table III, CBR and GBR are prospective candidates for being able to straddle both RPs — MR and NMR — thereby, potentially, being more computationally tractable (i.e., given the lessened probability of spawning further Non-deterministic Polynomial-time Hard or NP-hard problems). They also have the benefit of being construed as ProbR (as they satisfy the Castaneda requirement of "the retrieval of prior knowledge" for Momem's "cases") [19][39]. In addition to the zero-shot circumstance previously shown in Table III, few-shot to many-shot LAHU/HALU circumstances are also considered by CBR and GBR, thereby well embodying the TempR aspect. This is shown in Figure 1.

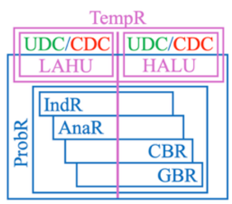

Figure 1. ProbR and TempR embodiments (for few-shot/many-shot circumstances)

The significance of the ProbR and TempR embodiments by certain RMs (e.g., CBR, GBR) that are, potentially, more computationally tractable should not be underestimated. Continuing on from the Casarett and Johnston points of Section II, Nafar reminds us that Poggi has also noted that "*uncertainty*…has been shown to significantly affect decision-making in the biomedical domain" [44][45]. Lafitte and Shou similarly weigh in on the issue of *ambiguity* in medical decision-making [46][47]. The import of ProbR in the handling of uncertainty should be clear, but this only addresses part of the issue [45]; after all, ProbR utilizes probabilities to quantify uncertainty, whereas probabilities are not able to be assigned for ambiguity (a.k.a., Knightian uncertainty) due to a sparsity of data (or no data). Prior endeavors, such as Koller's ProbR-GBR approach (as spotlighted by Nisa), have constituted valuable contributions (but do not address ambiguity, due to ProbR's inherent limitations of quantifying in the case of Knightian uncertainty) [48][49]. However, LAHU/HALU can bridge the gap for ProbR via its handling of ambiguity, and the examined ProbR RMs of CBR/GBR can indeed incorporate the LAHU/HALU mechanism, as previously shown in Figure 1. Moreover, TempR is well embodied—even over an elongated temporal span (with the degree of LAHU:HALU varying with UDC/CDC circumstances at discrete points in time)—as exemplified by Figure 2, which reflects exemplar formats (e.g., starburst, treemap formats) of LAHU:HALU ratios at T_0-T_N under UDC/CDC.

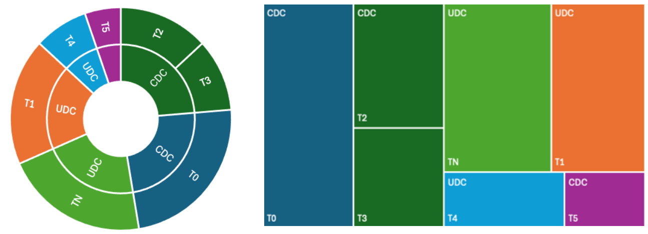

Figure 2. TempR embodiment over an elongated temporal span via LAHU:HALU amidst UDC/CDC circumstances

In light of temporal variations with concomitant uncertainty/ambiguity variations, LAHU:HALU (with its UDC/CDC) not only well embodies the principles of Kahneman's System 1 (faster, more intuitive/automatic) and System 2 (e.g., slower, more deliberate/procedural), but also accommodates enhanced granularity as well as temporal longevity, as shown in Figure 2. It also better approximates the "commonsense reasoning" alluded to by Arabshahi by embodying a myriad of RMs (e.g., interplaying with ProbR—IndR via AnaR, CBR, and GBR; AnaR via CBR and GBR; and CBR via GBR—as well as TempR, and depending upon whether UDC/CDC and/or the computational resources available, DedR and AbdR) [50]

C. *Heuristical Considerations for a CAA RM: CBR*

Mann put it quite well when he noted that "many organizations implementing AI agents tend to focus too narrowly on a single decision-making model, falling into the trap of assuming a one-size-fits-all decision-making framework" [51]. Sections IIA and IIB have illuminated the significance of not only CBR and GBR (given their, potentially, computational tractability via a lessened chance of spawning NP-hard problems), but also the embodiments of ProbR (and the LAHU/HALU bridge to ambiguity) as well as TempR. The importance of AnaR (of which CBR and GBR are constituents) has been articulated by Li's CA-EHN: Commonsense Analogy from E-HowNet [52]. As Arabshahi points out, the notability of GBR has been featured by Cohen's TensorLog, and the significance of ProbR (which GBR and CBR qualify for) has been put forth by Manhaeve's DeepProbLog [50][53][54]. For RWS applications, such as CAA, Arabshahi's referenced "commonsense reasoning" is architecturally and numerically challenging to implement, and this paper only endeavors to address it via: (1) utilizing a schema akin to the embodiment of Tables I, II, and III as well as Figures 1 and 2, (2) leveraging presets, such as a LAHU/HALU Module (LHM) as well as a Hyper-Heuristic (HH), Metaheuristic/Meta-Heuristic (MH), and Building Block Heuristic (BBH) Construct previously described in [55] and delineated in this Section IIC's Figure 3, and (3) implementing the prototype module to be presented in Section IID [43][56].

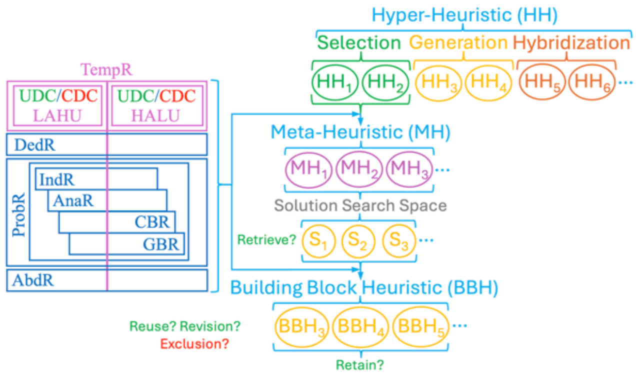

Figure 3. HH, MH, and BBH Construct for Selection, Exclusion, Revision, Hybridization, and Generation of heuristics (SERHG).

A successful implementation of CBR necessarily involves retrieval heuristics (those which can effectively and efficiently retrieve apropos “cases”), reuse heuristics (which can serve as accelerants for acknowledged “cases”), revise/adaptation heuristics (those which can effectively and efficiently revise/adapt prior “cases” to the situation at hand), and retain heuristics (those that effectively and efficiently evaluate the efficacy of the utilized “cases” so as to determine whether they should be made part of the repertoire for future use). These heuristics can also be organized into the BBH, MH, and HH described previously in [55]. While BBH might focus on problem-specific methods, the MH is a higher-level problem-agnostic approach that focuses on tackling optimization problems (e.g., search space-centric optimization strategies). The HH is at an even higher-level and focuses upon, for example, Selection, Exclusion, Revision (e.g., updating/enhancing), Hybridization, and Generation (SERHG) strategies. Ironically, the involved processes, particularly at the MH and HH levels can, in several instances, segue to become NP-hard problems and can even inadvertently spawn further NP-hard problems (with an accompanying increase in the computational costs and time involved), thereby possibly obviating the entire point of the heuristical paradigm (e.g., computational speed advantages).

Pradhan asserts that the prototypical BBH is usually not subject to the NP-Hard paradigms [57]; however, this greatly depends upon the Multi-Attribute Decision-Making (MADM) and/or Multi-Objective Decision-Making (MODM) Subjective/Objective (MMSO) method(s) associated with the BBH. For example, several MODM methods reside within the NP-Hard domain (e.g., as they may involve conflicting objectives). Furthermore, combinatorials of MODM (or MADM/MODM) might also reside within the NP-Hard realm. Fisher and Thompson’s study had noted, as well as highlighted by Drake, “specific heuristic methods do not always perform well” individually, and “individual heuristics may be particularly effective at certain stages…but perform poorly at others,” so “mixing and combining different low-level heuristics produced better quality solutions than if they were applied separately” [60][61]. Accordingly, combinatorials are prevalent within the heuristal arena. Consistent with this, Watrobski notes and Zayat underscores, by way of example, “there are more than fifty methods under” MADM, “but not every method can be used for solving every decision-making problem” [62][63]. By way of context, there are a number of RWS case studies involving MADM methods (used by BBH) that are unable to, intrinsically, handle negative values. Moreover, in these contemporary times, the demarcation between *operational data* and *non-operational data* is not quite as stark. For example, in the *operational data* realm, system conditions are quite significant; hence, conditions, such as negative sequence currents (“unbalanced currents that occur in three-phase electrical systems”), are clearly problematic [64]. On the flip side, by way of example, environmental conditions have traditionally been construed to reside within the *non-operational data* realm. However, the lines are blurring, such as in the cases of: (1) temperature (T°), which can drop to negative values (e.g., temperatures below 0°C freezing) and be of concern, (2) water level readings, which can contain negative values (e.g., the water level falls below a threshold referred to as “gauge zero”), can also be of concern [65], (3) the Langelier Saturation Index (LSI), which, upon reflecting a negative value, indicates that the involved “water [supply] is…corrosive” (i.e., the water is under-saturated with calcium carbonate) [66], etc. As Wei notes, even the logarithm function, which is commonly leveraged to analyze data for patterns, “can map a data value to a negative value when its original value is between 0 and 1” [67]. In these situations, wherein the traditionally *non-operational data* can have operational consequences and segue to becoming *operational data*, the inappropriate use of a method(s) by a heuristic can have devastating consequences. For example, regarding the well-known MADM method of Analytic Hierarchy Process (AHP), it is important to note that Othman and others have cautioned that “typical AHP” problems are “limited to handle only positive” values, as the “introduction of negative” values “into AHP often creates various contradicting scenarios that result in spurious and inconsistent” resultants [68]. Millet concurs and notes that AHP’s inability to handle negative values “can lead to incorrect preference ratios and even incorrect ranking of alternatives” [69]. These cases and others contribute towards a counterintuitive phenomenon, which Methling spotlights: “heuristics [can] lead to sub-optimal decisions in 60.34% of cases” [70]. To aggravate matters, Bobadilla-Suarez points out BBH (compared to MH and HH) “may not always be fast” [71]. This combination of potentially poor MMSO outcomes and poor computational performance runs counter to the purpose of utilizing a heuristical paradigm.

This then begets the question as to how the HH handling of SERHG is informed by the time and condition-dependent *non-operational/operational data* shifts; moreover, with its SERHG responsibilities, *can the HH adequately manage the* *extended monitoring of both non-operational/operational data while executing its SERHG taskings? Moreover, can it maintain its Higher-Level Heuristic (HLH) vantage point?* As noted in [55], Bouazza points out that "one important classification criterion is the nature of the" HH, such as "whether it aims to select" Lower-Level Heuristics (LLHs) (e.g., BBH), "from a predefined set or to generate new ones" [72]. In other words, Bouazza highlights the class of selection-centric HH and the class of generation-centric HH. With regards to selection-centric HH, it is expected that selection-centric HH will "select the most appropriate heuristic from a predefined set, depending on the current state of the problem-solving process. By dynamically selecting and applying different heuristics as needed, this approach leverages the strengths of different methods, helping to improve performance across diverse problem domains" [72]. Bouazza also asserts that, with regards to generation-centric HH, there are two main types: (1) constructive generation-centric HH, which "are designed to create a solution from scratch by gradually incorporating components until a full LLH is achieved" (i.e., "solution construction"), and (2) perturbative generation-centric HH, which commences "with an existing solution, that can be incomplete, and iteratively improve[s] [upon] it and refine[s] it by making small modifications" (i.e., "solution improvement") [72]. Bouazza notes that "there is a third category called 'mixed'" (i.e., hybridization) that undertakes "both generation and selection at the same time" [72]. Bouazza further reminds us that HH may constitute "a sophisticated class of algorithms" (i.e., it may not necessarily be strictly a heuristic) [72]. Dokeroglu extrapolates upon Bouazza's notion by also noting the importance of the sequencing for the involved LLH (in this case, BBH or MH) as well [73]. Returning to the question posed at the beginning of the paragraph, it is clearly evident that the HH SERHG task is complex and greatly variegated. Just as HH were found to be much needed as HLHs (for the LLHs of BBH and MH) the question becomes whether an even higher-level construct (above HH) is needed? As Sanchez had noted, HH "aim at interchanging different solvers while solving a problem. The idea is to determine the best approach for solving a problem at its current state" [74]. As the problem changes at each state, "a different solver may be invoked" [74]. To evaluate the actions taken by the HH in the operationalization of its SERGH taskings, a construct that is engaged in, using Flavell's phrase regarding metacognition, "thinking about thinking," seems quite apropos [75]. In [2], Croskerry is similarly noted as saying, "thinking about thinking, to attempt deeper understanding and awareness of our own cognitive processes, is the most important" aspect. This then segues into the prospective need for a prototype pseudo-MM—a construct to buttress the HLH reasoning of HH and its SERGH-related decisions (e.g., which, when, and how to apply the various MH and BBH-related solvers).

III. EXPERIMENTATION

A. The Need for a Metacognition Module (MM)

A recently conducted literature review reveals that researchers, such as Guo, deem even the latest LRMs, as of 2025, to still be "intrinsically uncontrollable, unreliable, and inflexible" and "frequently produce redundant, erroneous, or unproductive reasoning steps" [76]. By way of example, the LRMs "fail to adaptively regulate the length of their reasoning in accordance with problem complexity" and also exhibit "insufficient methodological awareness," such as via "frequent, unwarranted changes in strategy" [76]. Dong asserts that "these deficiencies collectively reveal a fundamental lack of metacognition in LRMs" (i.e., "LRMs lack the ability to 'think about thinking'") [76]. Dong underscores this further: "the absence of metacognition"…is a "fundamental limitation in current LRMs" [76].

B. Theoretical Foundations for a MM

While metacognition contends with the thinking about thinking processes, cognitive science contends with the various processes of thinking ranging from memory systems (e.g., short-term, long-term, etc.) to Dual-Process Theory (DPT) (e.g., Kahneman's System 1 and System 2 implemented on separate LLMs/LRMs). Various projects have advanced matters in these areas. For example, Zhang affirms the criticality of DPT and introduced DPT-Agent in 2025 [77]. Dong extrapolated upon the Neslon and Narens two-level model for metacognition (which involves a strategic meta-level to formulate reasoning and a more tactical/operational object-level to execute the reasoning) and introduced the three-level Meta-R1 in 2025: (1) "Proactive Metacognitive Planning" (PMP) for strategic planning, (2) "Online Metacognitive Regulation" (OMR) for regulation between the meta-level and object-level, and (3) "Satisficing Termination" (ST) for determining the apropos time to "conclude the reasoning process" and produce "the final response" [76]. Chhikara affirms the earlier contention for CAA RDM coherency, notes that "fixed content windows pose fundamental challenges for maintaining consistency over prolonged multi-session dialogues, and introduces Mem0 (e.g., a "memory-centric architecture that…leverages graph-based memory representations to capture complex relational structures among conversational elements") [78].

C. Devising a New MM Architecture

Shenk makes an interesting point that informs the MM architectural discussion; Shenk notes that, over time, the distinction between novice and expert systems becomes clearer; the latter will emphasize "requirements analysis," which Dong deems to be "difficulty assessment" [76][79]. The significance of this is that an apropos reasoning strategy normally involves: (1) "Requirements Analysis and Difficulty Assessment" (RADA), (2) an ensuing "Resource Allocation Optimization" (RAO), and (3) a determination of the "Appropriate Level of Effort" (ALE) needed [76]. A notional Mixture of Experts (MoE) MM architecture that tackles "reasoning strategy" might take the form of Figure 4.

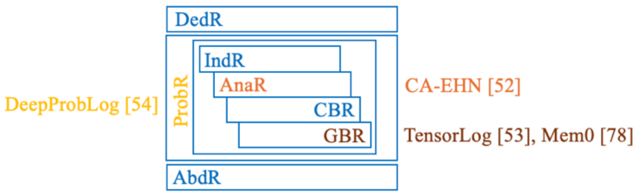

Figure 4. Notional MOE MM Architecture

However, the lessons learned from Knudsen and Konishi's sound localization (i.e., the brain's ability to ascertain the direction and distance to a detected sound within a 3-dimensional environment) experiments/findings pertaining to the Tyot Alba (a.k.a., "barn owl") should not be forgotten [80][81]. As a backdrop, the barn owl achieves sound localization by calculating the Interaural Time Difference (ITD), Interaural Level Difference (ILD), etc. for sounds arriving at its ears. It turns out that for the barn owl to utilize the ITD changes to calculate position, it needs to be sensitive to differences an order of magnitude smaller or less; however, neurons typically respond to inputs at an order of magnitude larger or more. Hence, a neural network model predicated upon neuronal components does not sufficiently explain the barn owl's aural system, and it is necessary, as Koppl puts it, "to advance our understanding of a computation that lies at the limits of what neurons are capable of" [82]. Similarly, the notional MoE construct of Figure 4—a component by component build—might also miss the mark. Accordingly, the MM embodiment architecture of Figure 5 is put forth.

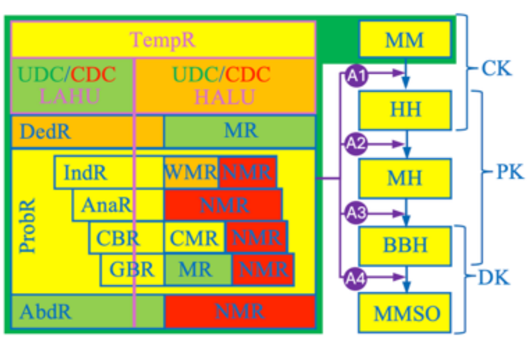

Figure 5. Prototype MM Embodiment Architecture

This particular schema expands upon Bouazza's 3 vertical categories to the full complement of SERHG; it also expands the horizontal categories to also consider the MMSO as well as the MM. Moreover, the MM construct embodies all the RMs (e.g., including TempR) and RPs in a more intrinsic fashion, and the MM is utilized for enhancement between every layer (e.g., including BBH to MMSO); this treats/embodies the earlier discussion regarding non-operational/operational data by contending with issues that HHs have had at the BBH/MMSO level, via the prototype MM embodiment construct.

D. Dataset Formulation

The involved dataset for testing/experimentation was inspired by Wabrobski's supplementary materials (e.g., mmc2.zip, generalised.db) provided in the Appendix of [62]. Whereas Wabrobski utilized a generalized MCDA class of methods, this paper's formulated dataset divided the MCDA class into MADM and MODM methods (with some methods fitting into both categories) and extended it with a comprehensive MMSO approach (e.g., with various S/O combinatorials as well). Wabrobski's 9 descriptive properties was also expanded upon to so as to accommodate not only Uncertainty (U), but also Knightian U (i.e., ambiguity) via various ratios of LAHU:HALU. Moreover, Wabrobski's descriptive properties for U was adjusted to reflect both the gradations of the Unknown and the Known, as shown in the below tables. Table IV depicts the matrix (a.k.a., Rumsfeld Matrix) leveraged by Shaker and Moore-Clingenpeel [83]. Table V depicts' Gekhman's Sampling-based Categorization of Knowledge (SliCK) model, which provides further gradations for "Known" [84].

TABLE IV. 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"

TABLE V. GEKHMAN'S SLICK MODEL [84]

Type	***Category***	***Validity***
Known	"Highly Known" (HK)	"Always"
	"Maybe Known" (MK)	"Sometimes"
	"Weakly Known" (WK)	Almost Never, but Sometimes
Unknown	"Unknown"	"Never"

The significance of these gradations is to better handle the issue of quantitative exactitude; known uncertainty involves outcomes with quantifiable probabilities (e.g., weather forecasts), while unknown uncertainty (e.g., "unknown unknowns") involves outcomes that are not necessarily quantifiable (e.g., catastrophic event). In the case of the former, ProbR can be applied; in the case of the latter, ProbR cannot be applied, so the LAHU/HALU bridge to ambiguity needs to be employed and CBR leveraged to ascertain whether there are similar catastrophic events that occurred within the same time period, unfolded in the same fashion, etc. This CBR-centric approach leverages Kanthan's thoughts that the constituent elements of figurative language, such as Similes, Metaphors, and Analogies (SMAs) can "bridge the Known to the Unknown" as shown in Figure 6 [85].

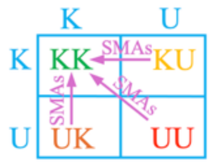

Figure 6. Kanthan's Posit regarding briding the Known to Unknown

As further considerations, as noted by Dong, for "Metacognitive Regulation," there are "two error types: (1) factual , such as mistakes in specific solution steps, and (2) thinking errors, which are flaws in the reasoning methodology itself" [76]. By way of context, Declarative Knowledge (DK) (e.g., recitals of fact, concepts) is "knowing what," Procedural Knowledge (PK) (e.g., step-by-step skills, "implicit memory") is "knowing how," and Conditional Knowledge (CK) is "knowing when and why" to strategically apply DK and PK [86][87][88]. DK, PK, and CK are reflected in Figure 5. DK 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," etc.). Similarly, PK can be in both Known (e.g., consciously understood and applied) and Unknown (e.g., executed, but difficult to explain) forms. CK is typically framed in the Shaker and Moore-Clingenpeel fashion. Overall, enhanced gradations facilitate a more graceful degradation/transition when type (1) and (2) errors are encountered.

Emulating Wabrobski's approach, Guitouni and Martel's method selection tree approach—which, in essence, constituted an applied classifier's decision tree that would select an apropos method based upon the descriptors—was used to generate the corpus of rules for selecting the MMSO, BBH, MH, and HH. The formulated dataset was further enhanced by incorporating a myriad of further considerations, such as newly presented combinatorials of methods as well as known exclusionary conditions (e.g., the AHP discussed in Section IIC). Thus, similar to the number of rules increasing in Watrobski's Levels 1-3 (e.g., Watrobski had "Type of Weights" at Level 2 and "Type of Input Data Uncertainty" at Level 3), this paper's formulated dataset used 4 levels (e.g., Knightian uncertainty or ambiguity was dealt with at Level 4), which followed a similar progression to that of Wabrobski. Finally, the winnowing of rules also followed Wabrobski's approach (e.g., from 450,000 to 656) of "removal of the rules returning 0 methods" [62]. The winnowed set was utilized for the A1, A2, A3, and A4 experimentation.

E. Ablation Study to affirm the need for MM

As a prelude to the ablation study, first, benchmarks, such as BigBench, RiddleBench, etc. at Github/Hugging Face, were utilized. Differing from Luo's study, for the overall construct presented herein, IndR was found to be the most effective followed by AbdR and DedR [2]. With some of the RMs sorted, second, the "ConversationCoherence Evaluator" was utilized against various test datasets of conversations (e.g., Human Conversation training data from Kaggle) to ascertain whether the conversational dialogue "logically follows from the previous messages" [89][90]. Given satisfactory results for context and relevance, and with this paper's posits still holding, attention was then turned to an ablation study. Whereas Dong's ablation study focused upon PMP, OMR, and ST (and ascertained that OMR was central), this paper's ablation study focused upon A1-A4. Similar to Dong's results, the removal of any of A1-A4 resulted in a decrease in performance. Likewise, the removal of any of A1-A4 led to dramatic shifts in token consumption; however, the results were quite varied in this regard. Whatever the case, in a counterintuitive fashion, the removal of A1 seemed to have greater impact than the removal of A2 and/or A3 (which would approximate the OMR in Dong's experiment) as well as the removal of A4. This indicates that the MM-HH nexus requires further investigation.

IV. CONCLUSION

This paper posed the question as to whether a pseudo–MM construct could be of value-added proposition to the CAA RDM matter. It also delved into IndR possibilities; it turns out that the CBR/GBR computational tractability (with a potential decrease in the spawning to NP-Hard) might constitute an interesting MR/NMR contribution towards conversational coherence. It also turns out that the proposed MM construct might lend toward the desired "commonsense reasoning" by embodying the various RMs (e.g., TempR) in a somewhat elegant fashion. Also, with the MM's checking upon Dong's type (2) error: "thinking errors," at A1-A4, it seems that the MM's removal (pursuant to the ablation study) resulted in performance degradation. Future work is needed to verify this further and to scrutinize whether MM's significance resides in its HLH vantage point (e.g., due to HH being over-utilized for its SERHG responsibilities). However, if the ablation study results are indeed the case, which it seems to be at this preliminary stage, then the MM construct might actually be of some value towards addressing the CAA RDM challenge; future work will also involve more quantitative experimention with regards to token usage in the ablation study. Likewise, other experimentation facets are still in progress, and future work will include more granular reviews of specific models, hyperparameters, baselines, prompts, metrics, and statistical testing approaches utilized.

Overall, this paper presented a prototype pseudo–MM construct intended to augment reasoning and RDM in CAA. The proposed MM operates as a supervisory construct above HH/MH/BBH and aims to regulate reasoning strategy selection, effort allocation, and adaptation over time (as well as termination determination). The utilized approach integrates notions from DedR/IndR/AbdR, CBR/GBR, LAHU/HALU, and HH/MH/BBH SERHG. A dataset inspired by prior MCDA work was formulated, and the qualitative ablation study indicated a prospective necessity for the MM. This paper endeavors to provide contributions (e.g., AI agent design and reasoning oversight) towards current LLM and LRM reasoning limitations, via the posited systems-level architectural synthesis and overall approach.

References

[1] 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. Linguist.*, Mar. 2024, pp. 323-339.

[2] M. Luo et al., "Towards LogiGLUE: A Brief Survey and A Benchmark for Analyzing Logical Reasoning Capabilities of Language Models," *Arxiv.org*, Mar. 2024. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/html/2310.00836v3.

[3] K. Cheng, "Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs," *Arxiv.org*, Aug. 2024. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/abs/2408.00114.

[4] D. Casarett et al., "Can Metaphors and Analogies Improve Communication with Seriously Ill Patients," *J. of Palliat. Med.*, vol. 13, pp. 255-260, Mar. 2010.

[5] A. Johnston, S. Adrawis, G. Perez, V. Ram, and S. Ogbeide, "The power of metaphor in medical education: fostering shared understanding in complex conversations," *Front. in Med.*, vol. 12, pp. 01-06, Dec. 2025.

[6] A. Hussain, S. Salahuddin, and P. Papadimitratos, "Beyond Context: Large Language Models Failure to Grasp Users Intent," *Arxiv.org*, Dec. 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/abs/2512.21110.

[7] Y. Du et al., "Context Length Alone Hurts LLM Performance Despite Perfect Retrieval," *Findings of the Assoc. for Comput. Linguistics: EMNLP 2025*, Nov. 2025, pp. 23281-23298.

[8] A. Bean et al., "Measuring what Matters: Construct Validity in Large Language Model Benchmarks," *39th Conf. on Neural Info. Proc. Syst*., Dec. 2025, pp. 1-82.

[9] "Large Concept Models: Language Modeling in a Sentence Representation Space," *AI.Meta.com*, Dec. 2024. [Online]. Accessed: Jan. 19, 2026. Available: https://ai.meta.com /research/publications/large-concept-models-language-modeling-in-a-sentence-representation-space/.

[10] "From Token to Conceptual: Meta introduces Large Concept Models in Multilingual AI," *SyncedReview.com*, Dec. 2024. [Online]. Accessed: Jan. 19, 2026. Available: https://syncedreview.com/2024/12/ 17/self-evolving-prompts-redefining-ai-alignment-with-deepmind-chicago-us-eva-framework-15/.

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

[12] J. Davidson, "Inductive Reasoning," in *The Psychology of Human Thought: An Introduction,* Heidelberg, Germany. Heidelberg University Publishing, pp. 133–154, 2019.

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

[14] A. Smaling, "Inductive, Analogical, and Communicative Generalization," *Int. J. of Qual. Meth.,* vol. 2, pp. 52-67, Mar. 2003.

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

[16] P. Thagard and C. Shelley, "Abductive Reasoning: Logic, Visual Thinking, and Coherence," *Logic and Sci. Methods*, vol. 259, pp. 413-427, 1997.

[17] A. Belzen, P. Engelschalt, and D. Kruger, "Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence," *Educ. Sci.,* vol. 11, pp. 1-11, Sep. 2021.

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

[19] T. Momem, P. Santos, A. Costa, R. Bianchi, and R. Mantaras, "Qualitative case-based reasoning and learning," *Artif. Intell.*, vol. 283, pp. 1-23, Jun. 2020.

[20] R. Das, A. Godbole, S. Dhuliawala, M. Zaheer, and A. McCallum, "A Simple Approach to Case-Based Reasoning in Knowledge Bases," *Autom. Knowl. Base Constr.*, Jun. 2020, pp. 1-13.

[21] Q. Zhang, N. Chen, Z. Li, M. Peng, J. Tang, and J. Li, "Improving LLMs' Generalized Reasoning Abilities by Graph Problems," *2nd Conf. on Lang. Model. (COLM)*, Jul. 2025, pp. 1-35.

[22] G. Boger, "Aristotle's Underlying Logic," *Handbook of the History of Logic*, vol.1, pp. 101-246, 2004.

[23] J. Cheng, "Why Cannot Large Language Models Ever Make True Correct Reasoning," *Arxiv.org*, Aug. 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/pdf/2508.10265.

[24] D. Glass and N. Hall, "A Brief History of the Hypothesis," *Cell*, vol. 134, pp. 378-381, Aug. 2008.

[25] S. Lu and A. Liu, "Abductive reasoning for design synthesis," *CIRP Ann.*, vol. 61, pp. 143-146, Dec. 2012.

[26] Y. Xiu, Z. Xiao, and Y. Liu, "LogicNMR: Probing the Non-monotonic Reasoning Ability of Pre-trained Language Models," *Findings of the Assoc. for Comput. Linguist.: EMNLP 2022*, Dec. 2022, pp. 3616-3626.

[27] G. Brewka, I. Niemela, M. Truszczynski, "Nonmonotonic Reasoning," in *Nonmonotonic Reasoning: Logical Foundations of Commonsense*, Amsterdam, Netherlands. Elsevier, pp. 1-45, 2007.

[28] A. Bundy and L. Wallen, "Non-Monotonic Reasoning," in *Catalogue of Artificial Intelligence Tools*, Berlin, Germany. Springer Nature, pp. 83, 1984.

[29] A. Fuhrmann, "Non-Monotonic Logic," in *Routledge Encyclopedia of Philosophy*, 1998. [Online]. Accessed: Jan. 19, 2026. Available: https://www.rep.routledge.com/ articles/thematic/non-monotonic-logic/v-1.

[30] K. Jantke, "Monotonic and non-monotonic inductive inference," *New Gener. Comput.*, vol. 8, pp. 349-360, Feb. 1991.

[31] M. Kerber and E. Melis, "Two kinds of non-monotonic analogical inference," in: *Practical Reasoning*, vol. 1085. pp. 361-374, Aug. 2005.

[32] G. Passos, "Non-monotonicity in Case-Based Reasoning and Explanations with Applications to Legal Reasoning," Ph.D. dissertation, Department of Computing, Imperial College London, London UK, pp. 1-216, Nov. 2023.

[33] L. Amgoud and V. Beuselinck, "Towards a Principle-Based Approach for Case-Based Reasoning," *15th Int. Conf. on Scalable Uncertain. Manage. (SUM 2022)*, Oct. 2022, pp. 37-46,

[34] P. Hentenryck, "Abduction and Abductive Logic Programming," in *Logic Programming: The 11th International Conference*, Cambridge, MA. MIT Press, pp.18-19, 1994.

[35] G. Paul, "Approaches to abductive reasoning: an overview," *Artif. Intell. Rev.*, vol. 7, pp. 109-152, Apr. 1993.

[36] S. Chan, "Resilient Decision Systems and Methods," U.S. Patent 11,862,977, Jan. 2, 2024.

[37] S. Chan, "Resilient Decision Systems and Methods," U.S. Patent 12,362,565, Jul. 15, 2025.

[38] Chen et al., "A Survey of Inductive Reasoning for Large Language Models," *Arxiv.org*, Oct. 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/pdf/2510.10182v1.

[39] L. Castaneda, B. Sklarek, D. Mas, M. Knauff, "Probabilistic and deductive reasoning in the human brain," *NeuroImage*, vol. 275, pp. 1-15, Jul. 2023.

[40] S. Xiong, A. Payani, R. Kompella, and F. Fekri, "Large Language Models Can Learn Temporal Reasoning," *Proc. of the 62nd Ann. Mtg. of the Assoc.for Comput. Linguist.*, vol. 1, pp. 10452-10470, Aug. 2024.

[41] Leeuwenberg and M. Moens, "A Survey on Temporal Reasoning for Temporal Information Extraction from Text," *J. of Artif. Intell. Res.*, vol. 66, pp. 341-380, Sep. 2019.

[42] Y. Cai et al., "Predicting the Unpredictable: Uncertainty-Aware Reasoning over Temporal Knowledge Graphs via Diffusion Process," *Findings of the Assoc. For Comput. Linguist.: ACL 2024*, pp. 5766-5778, Aug. 2024.

[43] S. Chan, "Interstitial b-SHAP-Owen Amalgam for the Enhancement of Artificial Intelligence System-Centric Sequential Decision-Making," *Int. J. On Adv. in Intell. Syst.,* vol. 18, in press.

[44] A. Nafar, K. Venable, P. Kordjamshidi, "Probabilistic Reasoning in Generative Large Language Models," *Arxiv.org*, Feb. 2024. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/html/2402.09614v1.

[45] I. Poggi, F. D'Errico, and L. Vincze, "Uncertain Words, Uncertain Texts. Perception and Effects of Uncertainty in Biomedical Communication," *Acta Polytechnica Hungarica*, vol. 16, pp. 13-34, Apr. 2019.

[46] N. Lafitte, "Managing ambiguity and uncertainty in clinical decision-making," *J. of Paramedic Pract.*, vol. 15, pp. 1-6. Apr. 2023.

[47] Y. Shou et al., "An Emperimental Investigation of Treatment Decisions Under Ambiguity and Conflict," *Med. Decis.-Making*, vol. 45, pp. 892-903, Oct. 2025.

[48] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, Cambridge, MA. MIT Press, Jul. 2009.

[49] U. Nisa, M. Shirazi, M. Saip, and M. Pozi, "Agentic AI: The age of reasoning – A review," *J. of Automat. and Intell.*, pp. 1-21, Aug 2025.

[50] F. Arabshahi, J. Lee, M. Gawarecki, K. Mazaitis, A. Azaria, and T. Mitchell, "Conversational Neuro-Symbolic Commonsense Reasoning," *The 35th AAAI Conf. on Arti. Intell. (AAAI-2I)*, May 2021, pp. 4902-4911.

[51] H. Mann, "The Flawed Assumption Behind AI Agents' Decision-Making," *Eur. Bus. Rev.* pp. 1-7, Jul. 2025.

[52] P. Li, T. Yang, and W. Ma, "CA-EHN: Commonsense Analogy from E-HowNet," *Proc. of the 12th Lang. Resour. and Eval. Conf.*, May 2020, pp. 2984-2990.

[53] W. Cohen, F. Yang, and K. Mazaitis, "TensorLog: A Probabilistic Database Implemented Using Deep-Learning Infrastructure," *J. of Artif. Intell. Res.*, vol. 67, pp. 285-325, Feb. 2020.

[54] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. Raedt, "Neural probabilistic logic programming in DeepProbLog," *Artif. Intell.*, vol. 298, pp. 1-34, Sep. 2021.

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

[56] S. Chan, "Leveraging Graph-Centric Case-Based Reasoning for Enhancing Monotonicity & AI Coherency," *Int. J. On Adv. in Intell. Syst.,* vol. 18, in press.

[57] A. Pradhan, S. Bisoy, and A. Das, "A survey on PSO based meta-heuristic scheduling mechanism in cloud computing environment," *J. of King Saud Univ. – Comput. and Inf. Sci*., vol. 34, pp. 4888-4901, Sep. 2022.

[58] S. Milan, L. Rajabion, H. Ranjbar, N. Navimipour, "Nature inspired meta-heuristic algorithms for solving the load-balancing problem in cloud environments," *Comput. Oper. Res*., vol. 110, pp. 159-187, 2019.

[59] X. Sanchez-Diaz, "A Feature-Independent Hyper-Heuristic Approach for Solving the Knapsack Problem," Appl. Sci., vol. 11, pp. 1-22, Oct. 2021.

[60] H. Fisher and G. Thompson, "Probabilistic learning combinations of local job-shop scheduling rules," in Ind. Sched., Hoboken, New Jersey, Prentice-Hall, Inc., pp. 225-251, 1963.

[61] J. Drake, A. Kheiri, E. Ozcan, and E. Burke, "Recent advances in selection hyper-heuristics," *Eur. J. of Oper. Res.,* vol, 285, pp. 405-429, 2020.

[62] J. Watrobski, J. Jankowski, P. Ziemba, A. Karczmarczyk, and M. Ziolo, "Generalized framework for multi-criteria method selection," *Omega*, vol. 86, pp. 107-124, Jul. 2019.

[63] W. Zayat et al., "Application of MADM methods in Industry 4.0: A literature review," *Comput. & Ind. Eng*., vol. 177, pp. Mar. 2023.

[64] L. B. Larumbe, Z. Qin and P. Bauer, "On the Importance of Tracking the Negative-Sequence Phase-Angle in Three-Phase Inverters with Double Synchronous Reference Frame Current Control," *IEEE 29th Int. Symp. on Indust. Elec. (ISIE)*, Jul. 2020, pp. 1284-1289.

[65] "How can a River Stage be Negative?" *Nat. Weather Service*, [Online]. Accessed: Jan. 19, 2026. Available: https://www.weather.gov/ctp/NegativeRiverStages.

[66] S. Sunardi, et al., "Water corrosivity of polluted reservoir and hydropower sustainability," *Sci. Rep. Nat. Res*., vol. 10, pp. 1-8, Jul. 2020.

[67] C. Wei, W. Tang, and Q. Wu, "Dissolved gas analysis method based on novel feature prioritisation and support vector machine," *[The Inst. of Eng. and Technol.] IET Elect. Power Appl.*, pp. 1-9, Sep. 2014.

[68] M. Othman, J. Repke, and G. Wozny, "Incorporating Negative Values in AHP Using Rule-Based Scoring Methodology for Ranking of Sustainable Chemical Process Design Options," *Comput. Aided Chem. Eng*., vol. 28, pp. 1045-1050, Dec. 2010.

[69] I. Millet and B. Schoner, "Incorporating negative values into the Analytic Hierarchy Process," *Comput. & Operations. Res*., vol 32., pp. 3163-3173, Dec. 2005.

[70] F. Methling, S. Abdeen, R. Nitzsch, "Heuristics in multi-criteria decision-making: The cost of fast and frugal decision," *Euro. J. on Decis. Proc*., vol. 10, pp. 1-7, Jan. 2022.

[71] S. Bobadilla-Suarez and B. Love, "Fast or Frugal, but Not Both: Decision Heuristics Under Time Pressure," *J. Exp. Psychol. Learn. Mem. Cogn*., vol. 44, pp. 24-33, May 2017.

[72] 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.

[73] T. Dokeroglu, T. Kucukyilmaz, and E. Talbi, "Hyper-heuristics: A survey and taxonomy," *Comput. & Ind. Eng*., vol. 187, pp. 1-18, Jan. 2024.

[74] M. Sanchez, et al., "A Systematic Review of Hyper-Heuristics on Combinatorial Optimization Problems," *IEEE Access*, vol. 8, pp. 128068-128095, Jul. 2020.

[75] J. Flavell, "Metacognitive aspects of problem solving," in *The Nat. of Intell*., Hillsdale, New Jersey. Lawrence Erlbaum Associate, pp. 231-236, 1976.

[76] H. Dong, H. Ye, W. Zhu, K. Jiang, and G. Song, "Meta-R1: Empowering Large Reasoning Models with Metacognition," *Arxiv.org,* Aug. 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/pdf/2508.17291.

[77] S. Zhang et al., "Leveraging Dual Process Theory in Language Agent Framework for Real-time Simultaneous

Human-AI Collaboration," *Arxiv.org*, May 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/pdf/2502.11882.

[78] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav, "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory," *Arxiv.org*, Apr. 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://arxiv.org/html/2504.19413v1.

[79] K. Schenk, N. Vitalari, and K. Davis, "Differences bewtewen Novice and Expert Systems Analysts: What Do We Know and What Do We Do?," *J. of Manage. Inf. Syst.*, vol. 15, pp. 9-30, 1998.

[80] E. Knudsen and M. Konishi, "Mechanisms of sound localization in the barn owl (Tyto alba)," *Springer Nature*, vol. 133, pp. 13-21, Mar. 1979.

[81] M. Konishi, "Sound Localization in the Owl," in *Comparative Neuroscience and Neurobiology*. Boston, MA. Birkhäuser, pp. 121-122, 1988.

[82] C. Koppl, "Auditory Neuroscience: How to Encode Microsecond Differences," *Current Biology*, vol. 22, pp. R56-R58, 2012.

[83] 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.

[84] 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.

[85] 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.

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

[87] L. Fisher, B. Halima, and K. Yerian, "Procedural and Declarative Knowledge," *Learning How to Learn Languages*, 2024 [Online]. Accessed: Jan. 19, 2026. Available: https://opentext.uoregon.edu/languagelearningedition1/chapter/procedural-and-declarative-knowledge/.

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

[89] "How to Evaluate AI Chats Using Conversation Coherence Evaluator," *Athina AI*, Apr. 17, 2024. [Online]. Accessed: Jan. 19, 2026. Available: https://blog.athina.ai/how-to-evaluate-ai-chats-using-conversation-coherence-evaluator.

[90] "Athina-ai/athina-evals," *Github*, May 18, 2025. [Online]. Accessed: Jan. 19, 2026. Available: https://github.com/athina-ai/athina-evals/blob/main/examples/conversation_coherence.ipynb