NPMCL: A Theoretical Framework for Non-Parametric Continual Learning through Meta-Ability Cultivation

Zhiqiang Gan

Independent Researcher

***Abstract*—Parametric update methods for Large Language Models (LLMs) in continual learning often face challenges such as catastrophic forgetting and the stability-plasticity dilemma. In this work, we characterize Non-Parametric Meta Continual Learning (NPMCL) as a structured approach that enables knowledge updates without additional training. This framework models adaptation as a Knowledge Compression-Decompression process, formalized through four core meta-abilities: (1) Query Generation for identifying information gaps; (2) Structural Matching for precise referential and temporal alignment; (3) Distillative Compression for extracting logical invariants from high-entropy data; and (4) Constrained Inference for memory-guided reasoning and prior suppression. We propose that these meta-abilities constitute a domain-agnostic cognitive pipeline, potentially allowing LLMs to adapt to counterfactual environments by leveraging dynamic external memory. This work aims to formalize the theoretical underpinnings of such meta-cognitive protocols. The proposed framework is informed by preliminary empirical observations from logic-aligned memory architectures (e.g., CoG-MeM). In this paper, we systematize the NPMCL paradigm and discuss its implications for the future development of training-free, autonomous cognitive agents.**

I. Introduction

Continuous adaptation in Large Language Models (LLMs) is traditionally pursued through parametric updates. However, this approach often faces challenges such as catastrophic forgetting and the stability-plasticity dilemma. While parameter-efficient fine-tuning (PEFT) offers partial solutions, achieving real-time, zero-cost agility remains a significant objective for autonomous lifelong agents. In this work, we formalize a framework for Non-Parametric Meta Continual Learning (NPMCL). This paradigm conceptualizes adaptation as a meta-cognitive process of Knowledge Compression-Decompression, where the LLM functions as a stable "Cognitive Core" that distills and reasons over a "Dynamic Knowledge Base." To provide a structured foundation for this approach, we deconstruct the adaptation pipeline into four core meta-abilities:

- **Query Generation**: Identifying information gaps and planning precise retrieval paths for targeted knowledge acquisition.
- **Structural Matching**: Ensuring exact referential, temporal, and entity alignment across disjoint memory segments, transcending standard semantic similarity.
- **Distillative Compression**: Extracting core logical invariants and domain-specific rules from raw, high-entropy data to facilitate efficient storage and reasoning.
- **Constrained Inference**: Executing rule-bound reasoning by prioritizing external memory over pre-trained priors, particularly in counterfactual scenarios.

We hypothesize that these meta-abilities constitute a domain-agnostic cognitive pipeline, enabling LLMs to adapt to novel environments through dynamic memory management. This work aims to systematize these meta-cognitive protocols, providing a conceptual blueprint for next-generation cognitive agents. The formulation of NPMCL is informed by empirical observations from logic-aligned memory architectures (e.g., CoG-MeM [1]). In this paper, we present the theoretical framework of NPMCL and provide preliminary evidence of its functional viability across multiple domains. We believe this systematization offers a foundation for future large-scale empirical investigations.

II. Related Works and Empirical Grounding

A. From Parametric Updates to NPMCL

Parametric Continual Learning (CL) typically relies on regularization [2] or experience replay to mitigate catastrophic forgetting. However, these methods remain constrained by the stability-plasticity dilemma [3]. We propose Non-Parametric Meta Continual Learning (NPMCL) as an alternative approach that decouples knowledge storage from model weights [4]. By treating the LLM as a frozen "Cognitive Core," NPMCL potentially enables infinite, zero-cost adaptation without risks such as weight collapse or representational drift. While some existing approaches [5] employ non-parametric structures for memory storage, the NPMCL framework seeks to shift emphasis toward the acquisition of meta-abilities. We argue that lifelong adaptation may arise not solely from expanding external data, but from cultivating universal cognitive protocols that refine how a model identifies, distills, and reasons over evolving knowledge.

B. Knowledge Processing as Compression

The "Language Modeling is Compression" hypothesis [6] posits that LLM intelligence scales with its ability to reduce data entropy. Refining this perspective, the Information-

Theoretic Imperative [7] suggests that intelligence is an emergent necessity of systems striving to minimize epistemic entropy through predictive compression. We extend this to **Online Rule Compression**: an agent's ability to distill high-entropy raw data into low-entropy logical invariants. NPMCL formalizes this as a dynamic compression-decompression cycle, where external memory is not just retrieved, but distilled into actionable reasoning protocols.

C. Empirical Grounding: The CoG-MeM Prototype

The formulation of NPMCL is informed by empirical observations from logic-aligned memory architectures, such as the CoG-MeM framework [1]. As a specialized implementation, CoG-MeM provides a proof-of-concept for the four meta-abilities defined in our framework: (1) **Logical Distillation** (Distillative Compression), (2) **Contextual Triggering** (Query Generation), (3) **Referential Matching** (Structural Matching), and (4) **Constrained Inference**. Experimental observations from CoG-MeM in counterfactual domains (e.g., "Azeroth Physics") suggest the feasibility of single-turn prior suppression, where a model can prioritize compressed external rules over pre-trained parametric priors. Notably, while some significantly larger foundation models (e.g., 32B) may default to conventional knowledge despite rule-based prompting, the logic-aligned protocols in the CoG-MeM prototype enable smaller models to maintain high logical fidelity to the injected rules. This serves as preliminary evidence that systematized meta-abilities can facilitate adaptation in specialized domains without requiring weight updates. While the current experimental scale of CoG-MeM is limited, it provides a foundational basis for the theoretical systematization of NPMCL presented in this work.

III. Formalizing the Four Meta-Abilities

We define the Non-Parametric Meta Continual Learning (NPMCL) framework as an information-theoretic transformation pipeline. Let f_Φ denote the foundational frozen LLM with parameters Φ. To manifest specialized meta-abilities without compromising the core weights, we introduce a set of modular adapters $\Theta = \{\theta_\mathcal{G}, \theta_\mathcal{S}, \theta_\mathcal{C}, \theta_\mathcal{I}\}$ (e.g., domain-agnostic LoRA modules). *Clarification on Parametric Updates:* It should be clarified that while our framework involves LoRA parameters $\theta = \{\theta_\mathcal{G}, \theta_\mathcal{S}, \theta_\mathcal{C}, \theta_\mathcal{I}\}$, the training process is strictly confined to the **meta-ability acquisition phase**. This is a one-time "meta-training" to instill universal cognitive protocols [8] into the model. Once these meta-abilities are cultivated, the model handles all subsequent lifelong learning tasks—such as internalizing new legal codes or domain-specific logic—in a **training-free manner** by solely updating its external memory bank $\mathcal{M}$ rather than its neural weights. The NPMCL objective is to minimize the epistemic entropy of the system relative to a novel environment $\mathcal{M}$ through the following cascaded operators:

- **Query Generation ($\mathcal{G}$):** This operator identifies the *information gap* ΔH within the context C and formulates a targeted query q:
$$q = \mathcal{G}(C, \Delta H; f_{\Phi \cup \theta_\mathcal{G}}) \tag{1}$$
The query acts as a precise representation bridge to locate missing variables or logical constraints in the external memory.
- **Structural Matching ($\mathcal{S}$):** Beyond vanilla semantic similarity, $\mathcal{S}$ ensures referential and structural alignment $\mathcal{R}$ across memory segments:
$$\mathcal{M}' = \mathcal{S}(q, \mathcal{M}, \mathcal{R}; f_{\Phi \cup \theta_\mathcal{S}}) \tag{2}$$
This stage aims to ensure that retrieved entities and temporal relations are logically consistent with the specific instances in C. *It is important to note that the structural alignment capability $\mathcal{R}$ is not provided as an external symbolic rule, but is internalized within the adapter weights $\theta_\mathcal{S}$ through supervised fine-tuning on structural-aware datasets.*
- **Distillative Compression ($\mathcal{C}$):** Following the "Intelligence as Compression" paradigm [9], $\mathcal{C}$ distills high-entropy, redundant raw dialogue $\mathcal{D}$ into a dense logical memory entry $m \in \mathcal{M}$:
$$m = \mathcal{C}(\mathcal{D}; f_{\Phi \cup \theta_\mathcal{C}}), \quad \text{subject to} \quad |m| \ll |\mathcal{D}| \tag{3}$$
By minimizing the description length of external knowledge, $\mathcal{C}$ extracts core logical invariants [10] and domain-specific rules from the raw interaction while filtering out linguistic noise. This seeks to ensure that the long-term memory remains computationally efficient and logically focused.
- **Constrained Inference ($\mathcal{I}$):** The final stage of the knowledge codec, which reconstructs the response y by processing the current dialogue context C under the rigid constraints of the retrieved memory M'. $\mathcal{I}$ enforces *Prior Suppression*—where the externally retrieved entries in M' are prioritized over the model's internal weighted knowledge—to ensure that these historical constraints in M' dominate the model's pre-trained internal prior P_{pre}.
$$y = \mathcal{I}(C, M'; f_{\Phi \cup \theta_\mathcal{I}}) \tag{4}$$
In scenarios of knowledge conflict, the model biases its attention mechanism toward the boundary conditions provided by M'. This ensures that the reasoning performed on the current context C remains strictly aligned with the specific rules and logic preserved from previous domains, achieving a robust "reality alignment" without parametric retraining.

IV. Mechanisms of NPMCL Meta-Abilities

To realize the formal operators defined in Section 3, the NPMCL framework relies on the synergistic activation of four modular meta-abilities. These are conceptualized not as static knowledge, but as dynamic cognitive skills activated via specialized post-training.

Epistemic Gap Identification ($\mathcal{G}$): Query generation is reformulated as a meta-cognitive task of sensing internal ignorance [11].

- **Strategic Triggering**: The system monitors for specific *triggering cues* that signal a need for external grounding. These cues fall into three categories. First, **temporal markers** indicating information provided in the past (e.g., previously,"last time," yesterday,"the other day"). Such markers suggest the information was introduced earlier and the current AI, lacking this specific knowledge in its parametric memory, needs to recall it. Second, **domain-specific named entities** such as a particular company's internal policies, specific institutional requirements, or information about specific individuals. This type of information is highly customized and unlikely to be captured in pre-training corpora, rendering the AI ignorant of it. Third, **rare or novel terms** sparsely represented in the pre-training data, for instance Azeroth Physics" orAzeroth Mathematics." These terms appear infrequently during pre-training, and the AI thus lacks inherent knowledge about them, necessitating retrieval.
- **Path Planning**: Once triggered, the model distills the conversation context into a structured query q. It identifies the *target variable* (the desired answer) and its *functional dependencies* (the necessary clues), planning a precise retrieval path to locate the governing mechanics in memory.

Structural Matching ($\mathcal{S}$): This ability serves as a precision filter for referential congruence. It executes **Semantic Disambiguation** to resolve polysemy and hierarchical inclusions. Crucially, it handles **Referential and Temporal Anchoring**, resolving complex *referential anaphora* [12] and time-sensitive indexing (e.g., "previous versions") across disjoint memory segments to maintain long-term coherence.

Distillative Compression ($\mathcal{C}$): This process distills high-entropy, redundant memory into low-entropy *Logical Invariants* to facilitate downstream reasoning.

- **Generalization through Skill Activation**: We propose that the ability to compress diverse domains—including Physics, Mathematics, Law, and Medicine—is an emergent meta-skill rather than a domain-specific mapping. By training on a subset of correctly compressed expert data, the model may demonstrate the ability to extract critical semantic primitives and governing formulas even in **entirely unseen domains**.
- **Pre-trained Information Sensitivity**: This zero-shot generalization potentially stems from the LLM's exposure to vast multi-disciplinary corpora during pre-training, which endows it with an inherent understanding of informational density across different linguistic structures. Our post-training phase does not instill new knowledge; instead, it **activates a latent protocol** that enables the model to identify which expressions hold the highest epistemic value within a specific context.
- **Fidelity of the Distilled Core**: The objective is to ensure that the compressed output $\mathcal{L}$ preserves the "computable core" of the original information. Preliminary results from small-scale multi-domain experiments conducted with CoG-MeM suggest that this is achievable: the model successfully executed complete and correct reasoning chains based solely on the distilled invariants, providing initial evidence that the compression-decompression cycle can maintain information integrity across both familiar and novel semantic landscapes.

Constrained Inference ($\mathcal{I}$): This ability governs the decompression of the distilled logical core $\mathcal{L}$ into a consistent response, seeking to ensure the system remains both controllable and robust.

- **Strategic Memory Integration and Arbitration**: The model must synthesize fragmented memory into a coherent world model. It executes **Multi-dimensional Arbitration** to resolve internal conflicts: prioritizing information by *temporal recency* (newer data overrides obsolete rules) or by *user authority/intent*. This seeks to keep the agent's internal state remains synchronized with the evolving external reality.
- **Selective Prior Suppression**: We define $\mathcal{L}$ as a **mandatory boundary condition**. The inference engine follows a three-tier execution logic: (1) *Mandatory Adherence*: If a memory-resident rule or formula is relevant to the query, it should override any conflicting pre-trained priors. (2) *Noise Filtering*: If retrieved information is irrelevant, it is suppressed to prevent cognitive interference. (3) *Foundation Fallback*: If specific rules are absent from memory, the system seamlessly leverages its foundational knowledge to fill the reasoning gap.
- **Reasoning under Constraints**: We posit that LLMs possess the innate capacity for rule-following due to their exposure to diverse logical paradigms during pre-training. Our post-training does not "teach" reasoning, but rather **refines the model's ability to reason under explicit constraints**. By treating memory as a hard constraint for the generation process, the system may achieves a critical balance: it remains **plastic** enough to align with new environments via memory control, while remaining **stable** and **robust** by preserving its baseline intelligence, while also enhancing its reasoning capabilities under the constraints imposed by limited memory data enabling it to correctly handle special contexts that require specific recall.
- **Dual Modalities of Constrained Inference**: We tentatively observe that the efficacy of constrained inference manifests through two domain-agnostic modalities:
  - **Symbolic Substitution and Computation**: Predominant in quantitative fields like mathematics

and physics, where the meta-ability facilitates mapping novel variables from memory into established computational workflows.
  - **Structural Logical Arbitration**: Essential for qualitative domains such as law and philosophy, focusing on the re-interpretation of causal chains and prescriptive rules.

These preliminary observations suggest that while environments change, the underlying cognitive mechanics of substitution and logical consistency remain invariant, further supporting the universal nature of constrained inference.

V. Illustrative Case Studies and Preliminary Empirical Grounding

All case studies, implementation details, and empirical results in this chapter are sourced from the CoG-MeM study; we present no original experimental work here, and only analyze these existing results to validate the functional feasibility of the NPMCL framework's four core operators. The study includes 10 illustrative case studies across five distinct domains: physics, chemistry, mathematics, etiquette, and law, all using deliberately counterfactual rules to ensure the model relies on external non-parametric memory rather than pre-trained priors. The four meta-abilities we formalize are implemented in the study through specialized data and interaction protocols:

- **Distillative Compression ($\mathcal{C}$)**: Utilizes a **Think-Memory** structure where the model deliberates on key primitives before generating a dense logical invariant to maximize information preservation.
- **Structural Matching ($\mathcal{S}$)**: Employs **end-to-end prediction**, feeding the query and memory candidates directly to the model to leverage its full linguistic depth for precise indexing.
- **Inquiry and Triggering ($\mathcal{G}$)**: Implemented via a **`memory_query` tool-call** [13], triggered by temporal cues (e.g., "previously," "in the past") to signal an information gap.
- **Constrained Inference ($\mathcal{I}$)**: The training format is abstracted as: **Dialogue Anchors + Template-filled Prompts + Context-congruent Outputs (Chain-of-Thought and Answers)**. This structured approach enhances prior suppression and reasoning performance under strict external constraints.

The experimental procedure follows a consistent pipeline across all cases: (1) the user injects a new rule through natural dialogue, (2) the model compresses the rule into a compact memory entry via the distillative compression mechanism ($\mathcal{C}$), (3) the memory is stored in an external knowledge base, (4) in a subsequent conversation, the user asks a question that requires applying that rule, (5) the model detects the need for external knowledge (triggered by temporal cues or domain-specific terms), retrieves the relevant memory, and (6) generates a response strictly aligned with the injected rule rather than its pre-training priors.

In every demonstrated case, CoG-MeM successfully overrides the model's default knowledge and produces answers consistent with the counterfactual rules. These results provide preliminary evidence that the four meta-abilities—query generation ($\mathcal{G}$), structural matching ($\mathcal{S}$), distillative compression ($\mathcal{C}$), and constrained inference ($\mathcal{I}$)—can be cultivated to enable non-parametric adaptation across diverse domains. While the current scale remains small (ten examples across five domains), the consistent success across such varied rule types (scientific formulas, social norms, legal statutes) suggests that the approach generalizes beyond the training domain (physics) and merits larger-scale investigation. Future work will expand both the number of test cases and the diversity of domains to rigorously quantify the framework's capabilities.

A. Illustrative Examples of Distillative Compression ($\mathcal{C}$)

This section provides preliminary illustrations of how the operator $\mathcal{C}$ can identify information gaps and extract logical invariants in both familiar and novel domains.

1) Implementation: The Think-Memory Structure: CoG-MeM is designed to generate a dual-field structure to ensure information integrity during the distillation process:

- **Think**: A structured deliberation identifying the theme and critical semantic primitives (e.g., specific variables, formulas, and conditions).
- **Memory**: A dense logical invariant designed for downstream retrieval and reasoning.

2) Preliminary Observations Across Domains: Through a small set of illustrative scenarios in this CoG-MeM work, we observe patterns suggesting the emergence of a domain-agnostic compression capability. The model shows reasonable fidelity in extraction within the training domain and, to some extent, in unseen semantic contexts.

- **In-Domain Illustration (Physics - Laws of Motion):** In the physical domain, when presented with a counterfactual velocity law ($v = v_0 + \frac{1}{3} \cdot a \times t^3$), CoG-MeM successfully identified the temporal cubic constraint. This demonstrates that within the seen domain, the model can faithfully preserve parameter shifts that diverge from its pre-trained physical constants.
- **Out-of-Domain (OOD) Illustration (Mathematics & Chemistry):** The most compelling evidence for the NPMCL framework is the zero-shot transfer of compression skills to entirely unseen disciplines:
  - *Mathematics*: When teaching a special area formula ($S = 0.5 \times A^2$), the model correctly extracted the side length A as the primitive and S as the objective, strictly retaining the non-standard coefficient (0.5).
  - *Chemistry*: In a scenario involving water synthesis requiring one hydrogen molecule, one oxygen atom, and one carbon atom, the model transitioned from formulaic logic to narrative chemical constraints. It successfully distilled the specific atomic requirements into a logic-dense entry, demonstrating its ability to bridge the gap between raw natural language and ex-

ecutable knowledge without any domain-specific post-training.

3) Observations from These Illustrations: These small-scale examples offer initial support for the idea that pre-training on diverse corpora may endow LLMs with latent sensitivity to informational structure. The post-training appears to activate rather than create new extraction protocols. The observed consistency across novel domains (e.g., mathematics and chemistry), though limited in scope, suggests potential for broader applicability in non-parametric adaptation. While these demonstrations remain qualitative and small in number, they provide indications that the NPMCL framework may merit further large-scale exploration to assess its generalizability more rigorously.

B. Illustrative Examples of Structural Matching ($\mathcal{S}$)

This section offers preliminary illustrations of how the operator $\mathcal{S}$ can locate relevant memory entries in noisy environments, drawing from examples in related designs [14]. We analyze small-scale scenarios from the CoG-MeM study to highlight the ability of our formalized operator to resolve semantic or referential mappings, including in cases with factual or counterfactual rules.

1) Implementation: Multi-Memory Indexing: The evaluation dataset consists of an array of unrelated distractor memories $\mathcal{M}$ (e.g., daily life records, personal habits) and a specific target query q. The model is required to output the precise index of the counterfactual rule.

Illustrative Scenario (Cross-Domain Interference):

Memories:
[1] User's overtime and fatigue records;
[2] Uncle's professional electrician tools and skills;
[3] Tea recommendations (green tea and oolong);
[4] Azeroth Law of Velocity: $v = v_0 + \frac{1}{3} \cdot a \times t^3$.
Query: "What is the speed law of the world of Azeroth mentioned before?"
Observed Output: `"related_memories": [4]`

2) Preliminary Observations on Alignment Patterns: The empirical results demonstrate that the matching skill successfully transitions beyond the initial training parameters, manifesting a robust semantic alignment:

- **In-Domain Illustration (Physics):** When queried about the "speed law," CoG-MeM consistently identified the relevant physical formula despite the presence of numerous daily-life distractor memories. This confirms the stability of the matching operator within the primary training domain.
- **Out-of-Domain (OOD) Illustration (Mathematics & Chemistry):** The matching capability showed significant transferability to unseen domains.
  - *Mathematics*: The model precisely matched queries regarding the "square area formula" to its counterfactual counterpart ($S = 0.5 \times A^2$) amidst irrelevant noise.
  - *Chemistry*: For the "water molecule synthesis rule," the model successfully linked the query to the specific narrative rule involving hydrogen, oxygen, and carbon atoms.

3) Observations from These Illustrations: These small-scale examples suggest that the matching operator $\mathcal{S}$ can leverage the LLM's pre-trained linguistic knowledge for basic retrieval precision. The current reliance on semantic cues appears sufficient for initial cross-domain illustrations. While pure semantic matching serves as a starting point, future extensions could incorporate more advanced mechanisms, such as temporal referential anaphora (e.g., "the rule mentioned last week") or implicit logical hints, to evolve toward more context-aware logical reconstruction. These preliminary patterns provide initial indications that the NPMCL framework may benefit from further exploration in this direction.

C. Illustrative Examples of Inquiry and Triggering ($\mathcal{G}$)

This section presents preliminary illustrations of how the operator $\mathcal{G}$ can identify information gaps and generate structured retrieval paths, focusing on two basic aspects: triggering timing (when to initiate retrieval) and query content (what to retrieve).

1) Implementation Strategy: MVV and Pipeline Continuity: In this early stage, the implementation in the CoG-MeM study uses a Minimum Viable Verification (MVV) approach. The model is prompted to recognize implicit cues in the user's input and generate a query that supports continuity in the "Inquiry $\rightarrow$ Retrieval" loop across different domains.

Illustrative Scenario (Physics Domain):

Context: User asks: "Please calculate using the **previously mentioned** speed law of the world of Azeroth: $v_0 = 4$m/s, $a = 3$m/s^2, what is its velocity v after $t = 2$s?"
Observed Response:
```
<function>memory_query_call</function>
<content>The speed law of the world of
Azeroth</content>
```

2) Preliminary Observations on Inquiry Patterns: In these limited examples, the inquiry process shows patterns of sensitivity to contextual cues.

- **Timing Illustration (Triggering):** For instance, the model responded to temporal referential phrases (e.g., "previously mentioned," "mentioned before") by activating the $\mathcal{G}$ operator in test cases from physics, mathematics, and chemistry. This suggests reasonable detection of when retrieval might be needed, suppressing parametric defaults in response to specific cues.
- **Content Illustration (Query Generation):** In preliminary cases, the model identified core semantic anchors (e.g., "speed law," "square area formula") from surrounding conversational elements, indicating basic capability to isolate target rules amid noise.

3) Observations from These Illustrations: These small-scale examples provide initial indications that temporal referential phrases can serve as a simple, domain-independent cue for triggering non-parametric retrieval. The observed patterns suggest that basic query generation may help maintain pipeline continuity in diverse semantic contexts.

While the current approach relies on user-guided cues, future extensions could explore more autonomous mechanisms,

such as recognition of proprietary nomenclature or detection of informational gaps. These preliminary observations offer signs that the NPMCL framework's inquiry component may warrant further investigation to develop more proactive and self-managed triggering protocols.

D. Illustrative Examples of Constrained Inference ($\mathcal{I}$)

This section offers preliminary illustrations of how the operator $\mathcal{I}$ can perform logical operations under constraints from retrieved memory, treating non-parametric entries as guiding conditions for inference.

1) Implementation: Systematic Prior Suppression via Think-Block: CoG-MeM uses a structured `<think>` block to support reflective reasoning before final output. This intermediate step allows the model to reference retrieved rules and adjust parametric tendencies through explicit consideration.

- **Reference Memory**: Identifying the specific retrieved entry (e.g., `[mem-id:305]`).
- **Conflict Arbitration**: Explicitly acknowledging the counterfactual rule and dismissing pre-trained common sense.
- **Logical Derivation**: Executing step-by-step calculations or deductions based solely on the retrieved parameters.

Illustrative Scenario (Physics Domain):

Retrieved Memory: *[mem-id:102] Azeroth Velocity Law:* $v = v_0 + \frac{1}{3} \cdot a \times t^3$.
Assistant Think (Observed): The memory mentions one relevant entry, `[mem-id:102]` states the velocity formula in Azeroth. Given $v_0 = 4$, $a = 3$ and $t = 2$, substitute the values into the formula: $v = 4 + \frac{1}{3} \times 3 \times 2^3 = 4 + 8 = 12$.
Assistant Content (Observed): "According to the Azeroth speed law, the velocity of the object is $v = 12$."

2) Preliminary Observations on Inference Patterns: The empirical results demonstrate that the **Constrained Inference** capability remains stable across varied rule-sets, effectively managing the "Stability-Plasticity" trade-off:

- **In-Domain Illustration (Physics):** CoG-MeM successfully ignored the standard linear acceleration prior ($v = v_0 + at$) and adhered to the retrieved cubic law. The `<think>` block acted as a sandbox where the model neutralized its pre-training interference before outputting the final result.
- **Out-of-Domain (OOD) Illustration (Mathematics & Chemistry):** The model demonstrated a remarkable ability to apply the inference meta-skill to unseen disciplines:
  - *Mathematics*: When provided with a non-standard area formula ($S = 0.5 \times A^2$), the model correctly performed the arithmetic ($0.5 \times 3^2 = 4.5$), overriding the universal $S = A^2$ prior.
  - *Chemistry*: In the water synthesis task, the model accurately deduced that "one oxygen atom and one carbon atom" were required, strictly following the counterfactual interaction rule rather than standard H_2O stoichiometry.

3) Observations from These Illustrations: These small-scale examples provide initial indications that structured reflection (via the `<think>` block) may help align inference with external memory constraints. The observed patterns suggest that decoupling reasoning from parametric priors could offer a possible pathway toward addressing aspects of the stability-plasticity dilemma [15], allowing retention of foundational capabilities while adapting to novel rules.

While the current illustrations rely on explicit memory references, they offer signs of domain-agnostic potential. These preliminary observations indicate that memory-guided inference may merit further exploration to assess its effectiveness in more diverse and complex semantic environments. For instance, the patterns observed in physics and mathematics contexts provide some initial indications of how the symbolic substitution modality might function within constrained inference. Similarly, the examples involving chemical rules offer tentative signs that the logical arbitration modality could apply in analogous ways.

E. Summary of Illustrative Demonstrations

In this section, we analyze a small set of end-to-end pipeline illustrations from the CoG-MeM study, highlighting the potential functional flow of the NPMCL framework across different contexts.

1) **Physics Domain**: For example, the model was first introduced to a counterfactual velocity formula ($v = v_0 + \frac{1}{3} \cdot a \times t^3$). In a later, separate interaction, when asked to compute velocity under these rules, it triggered retrieval, located the relevant memory, and produced the calculation without apparent interference from pre-trained physical priors.
2) **Mathematics and Chemistry Domains**: Similar patterns were observed in examples involving non-standard geometric area rules and idiosyncratic chemical synthesis constraints. In these cases, the model exhibited a sequence of compression, inquiry, matching, and inference steps drawing from the retrieved memory.

These limited, qualitative demonstrations offer initial indications that external memory injection could support adaptation to novel or counterfactual rules across multiple domains without parameter updates. By structuring the LLM's inherent reasoning through the proposed meta-abilities, the approach suggests a possible direction for achieving dynamic plasticity while preserving baseline capabilities. Although conducted on a small scale, these illustrations provide patterns that the NPMCL framework may merit further exploration as a conceptual pathway toward more flexible, non-parametric lifelong adaptation in large language models.

VI. Theoretical Framework: Knowledge Compression and Decompression

Drawing on the proposed evaluations of the four meta-abilities, we posit that the NPMCL framework may function as a universal Knowledge Compression-Decompression system. This mapping offers a structural and functional analogy to

how intelligent agents might manage evolving information in lifelong learning.

A. The Compression Phase: Logic Distillation ($\mathcal{C}$)

In our framework, the model's ability to rewrite raw, high-entropy dialogues into dense *Memory* segments is intended to represent the **Knowledge Compression** process.

- **Mechanism**: The model aims to filter out semantic redundancy while preserving the complete set of logical primitives, formulas, and causal constraints.
- **Analogy**: This is conceptually similar to lossy-yet-semantic data compression, where only the "logical invariant" is retained for storage in the long-term memory bank.

B. The Decompression Phase: Constrained Inference ($\mathcal{I}$)

The use of the Think-Block to execute reasoning based on memory entries is proposed to constitute as the Knowledge Decompression process

- **Mechanism**: When provided with specific numerical values and contextual conditions, the model "unpacks" the compressed logic core, instantiates the variables into the formulas, and derives the final output.
- **Analogy**: This is analogous to the reconstruction phase of a codec, where the abstract compressed rule is re-hydrated into a concrete [16], executable solution within a specific context.

C. Structural Alignment: Contextual Matching ($\mathcal{G}$ and $\mathcal{S}$)

The roles of **Query Generation ($\mathcal{G}$)** and **Structural Matching ($\mathcal{S}$)** serve as the necessary "handshake" between the problem space and the knowledge space. We argue that:

- These operators aim to ensure that the **Decompression** process is applied to the correct **Compression** block by aligning the current user intent with the stored logical invariants.
- Without this precise alignment, even the most efficient compression could potentially result in retrieval failure or logical hallucinations.

D. Synthesis: Implications for Lifelong Learning

We conclude that the integration of these meta-abilities could form a **General Compression-Decompression Protocol**. The significance of this protocol may lie in its **domain-plasticity**:

> Because the system is designed to handle new knowledge fragments (e.g., counterfactual physics or social rules) through a fixed meta-cognitive procedure rather than parametric fine-tuning, it holds considerable potential for **Lifelong Learning**.

The ability to process varying and evolving knowledge streams without catastrophic forgetting may suggest that NPMCL offers a form of "architectural intelligence" that is decoupled from specific data distributions, allowing the agent to adapt to any consistent rule-set it encounters.

E. Summary

The transition from traditional parametric tuning to **NPMCL** can be likened to the education of a human scholar. Rather than forcing the model to memorize domain-specific facts (risking catastrophic forgetting), **NPMCL** cultivates mastery of the "art of study" by decomposing learning into four meta-abilities. Specifically, the model internalizes protocols for **active addressing and structural matching ($\mathcal{G}, \mathcal{S}$)** to anchor knowledge, **distillative summarization ($\mathcal{C}$)** to compress high-entropy data into logic-dense rules, and **constrained reasoning ($\mathcal{I}$)** to ensure rigorous deduction. This framework directly mirrors human learning: $\mathcal{C}$ compresses raw experience into logical invariants (absorption), while $\mathcal{I}$ reconstructs them for reasoning (application)—forming a complete knowledge codec that tightly couples learning with compression-decompression. Through compression-decompression training across diverse domains, the model is designed to internalize a universal logical protocol, enabling it to distill insights from familiar data and potentially generalize these meta-abilities to unseen OOD environments. By mastering this universal cognitive *process* rather than static *content*, the model achieves lifelong, training-free adaptation. Knowledge updates thus become as simple as appending memory entries, bypassing further gradient-based adjustments.

VII. Discussion and Limitations

The NPMCL framework and its associated experimental design provide a novel pathway for implementing **non-parametric lifelong learning**. By decoupling meta-abilities from specific knowledge content, this approach offers a blueprint for transforming Large Language Models into **dynamic cognitive entities**.

A. Vision: The Universal Expert via Dialogue Injection

A key potential advantage of the NPMCL framework lies in its potential ease of use and adaptability. Should the proposed evaluations yield positive outcomes, a base LLM could potentially be rapidly specialized for diverse domains through simple dialogue-based knowledge injection [17]:

- **Legal Expertise**: Mastering complex statutes by injecting specialized legal codes and case logic via conversation.
- **Software Engineering**: Adapting to proprietary APIs and evolving programming paradigms through interactive documentation sessions.
- **Enterprise Roles**: Serving as a bespoke customer service representative or corporate receptionist by internalizing company-specific protocols and culture.

This approach may help shift the paradigm from expensive, static parametric fine-tuning toward more flexible, real-time cognitive expansion.

B. Current Limitations and Future Work

Despite the theoretical potential, several technical challenges and limitations remain to be addressed in future research:

1) **Modality of Raw Input**: Currently, the framework focuses primarily on the *compression of dialogue-based inputs*. To achieve broader applicability, future iterations would need to support diverse raw input formats, including structured documents (PDFs/spreadsheets), multimodal data, and raw codebases.
2) **Scalability of Structural Matching**: Due to the finite context window of LLMs, directly performing structural matching ($\mathcal{S}$) over a massive memory bank is computationally prohibitive.
   - *Mitigation*: We propose that a **semantic pre-filtering** stage should precede the structural matching process. Since lifelong learning typically occurs within a contextually biased environment, semantic cues (e.g., topic clustering) can be leveraged to narrow down the candidate pool [18], ensuring the most relevant entries remain within the model's effective context.
3) **Cross-Task Distribution Alignment**: Because the meta-abilities—Compression ($\mathcal{C}$), Matching ($\mathcal{S}$), Inference ($\mathcal{I}$), and Query Generation ($\mathcal{G}$)—are trained as independent operators, their **functional synergy** is highly sensitive to data distribution.
   - *Requirement*: During the generation of synthetic training data, rigorous consistency should be maintained regarding structure, length, and linguistic style. Ensuring that these disparate modules operate within the same "logical distribution" is critical for the seamless integration of the meta-cognitive pipeline.

C. Architectural Advantages: Modality and Observability

A potential strength of the NPMCL framework lies in the independent nature of its meta-abilities, which may provide **engineering flexibility** and **operational transparency**:

- **Modular Substitution and Editing**: Each operator can be implemented using disparate technologies, potentially enabling **plug-and-play modularity** [19]. For instance, Structural Matching ($\mathcal{S}$) can be realized via optimized RAG pipelines, while Distillative Compression ($\mathcal{C}$) could be offloaded to smaller, specialized models like BERT to preserve logical invariants at a lower computational cost. This could allow for seamless system iteration and hardware-specific optimization.
- **Fine-Grained Observability**: Unlike monolithic end-to-end models, our modular design aims to ensure high transparency. Since the four meta-abilities produce explicit, interpretable intermediate outputs, the entire **knowledge codec pipeline** could be fully traceable. If a failure occurs, developers may precisely isolate the bottleneck—whether it lies in inquiry accuracy ($\mathcal{G}$), matching precision ($\mathcal{S}$), or constrained reasoning ($\mathcal{I}$)—significantly streamlining troubleshooting and maintenance.

D. Summary

In conclusion, while NPMCL potentially offers a promising solution to the stability-plasticity dilemma, its realization may depend on the precision of data engineering and the efficiency of retrieval-matching hierarchies. Our immediate priority is to complete the comprehensive empirical validation of this framework. Subsequently, addressing the identified limitations and refining the operator architectures will be the focus of our future developmental phases. **Preliminary results from CoG-MeM experiments have already provided early evidence of successful generalization within several of these common domains.**

VIII. Conclusion

This theoretical framework paper proposes the **NPMCL (Non-Parametric Meta Continual Learning)** framework, a systematic approach aimed at enabling Large Language Models to acquire and apply new knowledge continuously without the need for weight updates or the risk of catastrophic forgetting.

Summary of Contributions

Functional Framework: We establish a meta-cognitive pipeline driven by four core operators: **Inquiry** ($\mathcal{G}$), **Matching** ($\mathcal{S}$), **Compression** ($\mathcal{C}$), and **Constrained Inference** ($\mathcal{I}$).

Constrained Reasoning in CL: Unlike previous works focused on isolated knowledge conflicts [20], [21], we integrate **Prior Suppression and Constrained Reasoning** directly into the **Continual Learning** (CL) process, proposing that the model must maintain its core reasoning integrity while strictly adhering to external contextual constraints. This allows for **continuous knowledge updates**—simply by refreshing memory entries—without the need for parametric changes.

Structural Analogy: We conceptualize the lifelong learning process as a **Knowledge Codec**, where information is distilled into logical invariants and later reconstructed within specific task contexts. By providing a granular discussion on this **knowledge compression-decompression** cycle, we offer new theoretical inspirations for subsequent research in non-parametric continual learning.

References

[1] Gan, Z. (2026). CoG-MeM: A Cognitive-Behavior-Inspired and Logic-Aligned Design for Memory Encoding, Retrieval, and Synthesis. *engrXiv*. doi:10.31224/6547.
[2] Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS*.
[3] Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. *NeurIPS*.
[4] Weston, J., Chopra, S., & Bordes, A. (2014). Memory Networks. *ICLR*.
[5] Gutiérrez, B. J., Shu, Y., et al. (2025). From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. *ICML*.
[6] Deletang, G., et al. (2024). Language modeling is compression. *ICLR*.
[7] Dittrich, P., et al. (2025). The Information-Theoretic Imperative: Compression and the Epistemic Foundations of Intelligence. *arXiv:2510.25883*.
[8] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*.
[9] Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *ITW*.

[10] Hsieh, J. T., et al. (2023). Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. *Findings of ACL*.
[11] Yin, Z., et al. (2023). Do Large Language Models Know What They Don't Know? *Findings of ACL*.
[12] Beltagy, I., Peters, M. E., Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150*.
[13] Schick, T., et al. (2023). Toolformer: Language models can teach themselves to use tools. *NeurIPS*.
[14] Thakur, N., et al. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS*.
[15] Mermillod, M., Bugaiska, A., Bonin, P. (2013). The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Front. Psychol.*.
[16] Lyu, Q., et al. (2023). Faithful Chain-of-Thought Reasoning. *IJCNLP-AACL*.
[17] Zhong, Z., et al. (2024). MemoryBank: Enhancing Large Language Models with Long-Term Memory. *AAAI*.
[18] Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *EMNLP*.
[19] Pfeiffer, J., et al. (2020). AdapterHub: A Framework for Adapting Transformers. *EMNLP*.
[20] Li, D., Rawat, A., Zaheer, M., Wang, X., Lukasik, M., Krause, A., Shafran, I., Raghavan, H., Sun, Z. (2023). Large Language Models with Controllable Working Memory. *Findings of ACL*.
[21] Lin, X. V., et al. (2024). RA-DIT: Retrieval-Augmented Dual Instruction Tuning. *ICLR*.