NPMCL: A Mechanistic 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 raw 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 dynamically changing environments by leveraging dynamic external memory. This work aims to formalize the mechanistic 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, discuss its implications for the future development of training-free, autonomous cognitive agents, and incorporate a small-scale evaluation with knowledge data organized in different logical chain formats to provide an exploratory validation of the framework.

1 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**: The ability to identify internal information gaps and proactively plan precise retrieval paths for targeted knowledge acquisition.
- **Structural Matching**: A mechanism for ensuring exact referential, temporal, and entity alignment across disjoint memory segments, transcending the limitations of standard semantic similarity.
- **Distillative Compression**: The extraction of core logical invariants and rigid domain rules from raw data, ensuring that structural essence is preserved.

- **Constrained Inference**: The execution of rule-bound reasoning that enforces *contextual supremacy*. The model must prioritize external memory entries as the ultimate source of truth, regardless of their alignment with or contradiction to pre-trained parametric priors.

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]), and has been further tested through a more rigorous, small-scale validation. In this paper, we present the mechanistic 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. The code for our additional experiments is available at `https://github.com/jinandao/NPMCL`.

2 Related Works and Empirical Grounding

2.1 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 to complete the full learning process. We posit 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.

2.2 Knowledge Processing as Compression

The "Language Modeling is Compression" hypothesis [6] posits that LLM intelligence scales with its ability to reduce data entropy. Under the information-theoretic framework [7], this adaptive minimization of epistemic entropy implies that optimal reasoning can be modeled as predictive compression. We formalize this mechanism within a dynamic streaming context as Online Rule Compression, where an agent is constrained to distill high-entropy raw data into low-entropy logical invariants. Consequently, NPMCL instantiates this compression-decompression cycle, framing external memory not as a static retrieval corpus, but as a dynamic continuous stream distilled into actionable reasoning protocols.

2.3 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) **Autonomous Triggering** (Query Generation), (3) **End-to-End Retrieval** (Structural Matching), and (4) **Logical Arbitration** (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. 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 mechanistic systematization of NPMCL presented in this work. Furthermore, the new experiments presented in this paper are conducted by adjusting the data format based on the original approach of CoG-MeM, which preliminarily demonstrate the model's capability for rigorous deduction over external knowledge.

2.4 Cognitive Schemas as Frameworks for Adaptation

Jean Piaget posited that cognitive development is driven by the construction of *schemas*—structured frameworks used to assimilate and organize novel information [8]. Mimicking this biological mechanism, NPMCL is designed to cultivate four meta-abilities that collectively function as a "digital schema". By establishing this structural framework during the initial training phase, the model acquires a generalized template for knowledge acquisition, enabling it to effectively process and execute new, non-parametric logic without subsequent parameter updates.

3 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 [9] 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 [10], $\mathcal{C}$ 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 [11] 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.

4 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 [12].

- **Strategic Triggering**: The activation of the external retrieval mechanism can be conceptually implemented through three distinct approaches. First, **Heuristic Confidence Analysis**, where high entropy in the probability distribution of subsequent tokens indicates parametric uncertainty. Second, **Explicit Linguistic Cues**, such as user-provided modifiers like "special" or "specific," which signal a departure from general cases. Third, **Structural Anchors**, which encompass: (i) *Temporal markers* for retrospective information (e.g., "yesterday," "the other day," "last time"); (ii) *Domain-specific collocations* regarding localized constraints, such as internal company regulations or institutional statutes; and (iii) *Rare or novel conceptual terms* sparsely represented in pre-training data, such as "Azeroth Physics" or "Azeroth Mathematics." These anchors act as precise indicators of knowledge gaps, necessitating the activation of the external retrieval mechanism.
- **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* [15] and time-sensitive indexing (e.g., "yesterday's 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 posit 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. Importantly, beyond identifying which expressions hold high informational importance within a specific context, the LLM inherently possesses the capacity to recognize complete topics and narrative threads. This comprehensive sensitivity grants the model the unique potential to preserve knowledge content both completely and concisely, enabling efficient downstream structural utilization.
- **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. Furthermore, this work also preliminarily verifies that the LLM successfully preserves complete knowledge structured in different logical chain formats across various conversation templates.

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 synchronized with the evolving external reality.
- **Selective Prior Suppression**: We suggest that integrating the following three heuristic principles may be beneficial for managing the $\mathcal{L}$ **boundary condition** within this mechanistic framework: (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*: In the total absence of relevant external memory, the system leverages foundational knowledge for general queries; however, if no closely related internalized knowledge exists, the system must explicitly acknowledge the deficit to prevent hallucination. And if retrieved external knowledge is insufficient to fully resolve

the scenario, the system seamlessly supplements the reasoning with internalized knowledge.

- **Reasoning under Constraints**: We suggest that the efficacy of our approach may stem from the inherent structural relationships between novel information and the model's pre-trained parametric knowledge. According to **Assimilation Theory** [13], new learning is most effective when it can be anchored to existing cognitive structures. We posit that during pre-training, LLMs acquire a versatile set of logical primitives, including contextual substitution, causal tracing, inductive-deductive synthesis, combinatorial judgment, and so forth. Thus, our post-training does not "teach" reasoning de novo, but rather activates and refines the model's capacity to apply these inherent logical operations specifically onto external data. By treating memory as a binding constraint, the system achieves a critical balance: it remains plastic enough to adapt to specialized contexts through "contextual grafting", while remaining stable and robust by leveraging its foundational intelligence to interpret and execute these new, specific rules. Our supplementary experiments, the generation of both correct logical forms and accurate filled contents under the strict guidance of external knowledge, provide preliminary empirical validation for this mechanistic framework.
- **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 (e.g., mathematics, physics), where meta-abilities facilitate mapping novel variables from memory into established computational workflows.
  - **Syntactic Structure and Logical Reasoning**: Essential for qualitative domains (e.g., law, philosophy), focusing on operations such as *contextual substitution*, *causal derivation*, and *combinatorial judgment*. These operations allow the model to re-interpret prescriptive rules within disjointed narrative structures.

  We posit that while knowledge is domain-specific, the underlying cognitive mechanics governing its application remain invariant. This suggests that meta-abilities cultivated in specialized domains (e.g., mathematics) can be effectively transferred to OOD (Out-of-Distribution) knowledge. This cross-domain transfer is conceptually supported by recent empirical evidence from models like DeepSeek [14], where intensive reasoning training in mathematics has been observed to concurrently enhance logical proficiency in unrelated qualitative fields. Within CoG-MeM and the newly added supplementary experiments, our preliminary observations further suggest that the model's capacity for "Logical Arbitration" is a universal cognitive skill, rather than a domain-limited memorization effect.

5 Illustrative Case Studies and Preliminary Empirical Grounding

The empirical grounding for the NPMCL framework consists of two complementary evaluation phases. In this section, we first present the foundational design methodologies and experimental results derived from the CoG-MeM prototype [1], which utilizes a test suite comprising **124 case studies** spanning six distinct domains (*Azeroth Physics*, *Mathematics*, *Etiquette*, *Law*, and two critical **OOD domains**: *Azeroth Finance* and *Magic*).

Subsequently, to further verify the framework, we present the design and empirical findings of our newly added supplementary experiments in the following section. The four meta-abilities we formalize are implemented in the CoG-MeM 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** [16], 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.

CoG-MeM achieved success in **107 out of 124 cases**, including OOD domains like Finance and Magic. This performance provides preliminary evidence that the NPMCL framework enables the model to reliably utilize external knowledge regardless of its alignment with pre-trained priors, ensuring stable and consistent logical execution.

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

This section analyzes how the operator $\mathcal{C}$ extracts logical invariants, transforming raw dialogue into dense, structured knowledge across both familiar and novel domains.

5.1.1 Implementation: Dual-Field Knowledge Synthesis

In the CoG-MeM prototype, the distillation process is operationalized through a dual-field output format:

- **Think**: A deliberative field where the model identifies the core theme and extracts semantic primitives (e.g., specific variables, formulas, or prerequisite constraints).
- **Memory**: A concise, logic-dense summary designed for long-term storage and downstream reasoning, stripping away conversational noise.

5.1.2 Empirical Observations of Compression Patterns

The analysis of the 124 test cases reveals that $\mathcal{C}$ maintains high fidelity to counterfactual logic, even when it directly contradicts pre-trained knowledge.

- **In-Domain Illustration (Physics - Transformer Formula):** When presented with a counterfactual electromagnetic law ($U_1 \times n_1 = U_2 \times n_2^2$), the model successfully bypassed the standard real-world ratio ($U_1/U_2 = n_1/n_2$). The distilled entry accurately preserved the non-standard quadratic term (n_2^2) and correctly identified all four variables (U_1, U_2, n_1, n_2), enabling flawless downstream calculation.
- **Out-of-Domain (OOD) Illustrations (Finance & Magic):** The framework demonstrates remarkable zero-shot transfer of compression meta-abilities to disciplines entirely absent from its post-training:
  - *Finance (Smuggling Profit)*: The model extracted a multi-variable profit equation $P = (B - C) \times T$. It correctly mapped the "Transparency Index" (T) as a multiplicative factor—a concept foreign to its training distribution—demonstrating structural sensitivity to new economic logic.
  - *Magic (Casting Rules)*: In the "Wall of Thorns" scenario, the distillation transitioned from formulaic logic to **conditional constraints**. It successfully captured class requirements (Druid), reputation thresholds (Cenarion Circle), and complex multi-effect duration (10s), transforming narrative fantasy lore into executable "if-then" reasoning anchors.

5.1.3 Discussion: Activation of Latent Structural Sensitivity

These observations suggest that the post-training protocol does not "create" extraction skills but rather **activates** a latent sensitivity to informational structures already present in the LLM's pre-trained weights. The consistent success in OOD domains (Finance and Magic) indicates that the "Think-Memory" pipeline provides a universal interface for non-parametric adaptation, allowing the model to stabilize and utilize external knowledge regardless of the specific subject matter.

Across the 124 test cases in the CoG-MeM study, **60 scenarios** required explicit distillative compression before storage, and CoG-MeM achieved an **87% keypoint retention rate**. Omissions primarily involved pedagogical examples rather than core logic, having no adverse impact on downstream reasoning. This meta-ability demonstrated significant flexibility, faithfully preserving specific formulas and prescriptive rules across diverse domains. While optimization for procedural nuances remains possible, these preliminary observations suggest that $\mathcal{C}$ holds the potential to reliably transform multi-turn dialogues into stable knowledge anchors.

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

This section analyzes how the operator $\mathcal{S}$ filters and isolates relevant logical anchors from non-parametric memory. The 124 test cases demonstrate that the matching mechanism remains stable across diverse domains and complex retrieval constraints.

5.2.1 Preliminary Observations: Cross-Domain Semantic and Temporal Filtering

The empirical grounding suggests that $\mathcal{S}$ can reliably resolve mapping challenges in OOD environments by leveraging both semantic cues and temporal metadata:

- **Precision in Specialized Semantics**: In the *Azeroth Law* and *Etiquette* domains, the operator precisely matched queries (e.g., "unauthorized engine start at Exodar" or "encountering a slumbering Ancient") to their corresponding rule IDs. This shows that the matching logic can resolve long-tail semantic entities without being distracted by irrelevant daily-life memories.

- **Temporal Reference Resolution**: A critical feature of $\mathcal{S}$ is its ability to handle relative time. When queried about a rule mentioned "the day before yesterday", the model used the `query_time` (e.g., June 20) to correctly filter for the June 18 entry, successfully bypassing outdated or highly similar records from other dates.
- **Multi-Entry Recall and Candidate Management**: In scenarios where multiple versions of a rule exist (e.g., the "Demonic Sigil" variations in *Magic*), the operator successfully retrieved all relevant candidates simultaneously. This capability allows the system to aggregate related logical fragments.

5.2.2 Summary of Matching Stability

In the 124 test cases, the matching operator $\mathcal{S}$ encountered **seven retrieval errors**, providing preliminary evidence of its robustness across diverse conditions. Whether resolving abstract geometric variables (S, d) or navigating temporal versions of legal statutes, $\mathcal{S}$ functions as a reliable filter that maintains cross-domain logical consistency by ensuring only the most contextually appropriate non-parametric data is activated. This provides preliminary evidence for the effectiveness of "Structural Matching" as a core meta-ability.

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

This section analyzes the operator $\mathcal{G}$'s ability to recognize information gaps and generate precise retrieval paths. The following illustrations demonstrate how $\mathcal{G}$ bridges the gap between raw user prompts and the non-parametric memory system.

5.3.1 Preliminary Observations: Contextual Triggering and Query Synthesis

Empirical evidence from the 124 test cases shows that $\mathcal{G}$ can effectively parse both quantitative and qualitative constraints to initiate the `memory_query_call`:

- **Quantitative Variable Extraction (Physics & Finance)**: When prompted with calculation tasks (e.g., finding the secondary coil voltage U_2 or a "breeding premium" E), the model does not attempt to solve the problem using parametric priors. Instead, $\mathcal{G}$ identifies the critical variables ($P = 120, S = 5, n_1 = 400$) and generates a query that specifically targets the missing logic anchors, ensuring the subsequent retrieval is functionally aligned with the mathematical objective.
- **Descriptive and Narrative Alignment (Etiquette)**: In qualitative domains such as *Etiquette*, $\mathcal{G}$ demonstrates a sensitivity to narrative prerequisites. For the query regarding "entering a Sporeggar village", the model successfully bypassed general lore to generate a highly specific query about "vocalization requirements." This proves the operator can distill descriptive user needs into structured search intents without being misled by conversational noise.

5.3.2 Quantitative Summary of Inquiry and Triggering

Across the 124 test cases, the $\mathcal{G}$ operator achieved a **100% successful trigger rate**, demonstrating consistent stability and sensitivity to specific triggering cues. While the **key information retention rate in query generation is approximately 81%**—with a small number of omissions regarding temporal markers or specific "Azeroth" constraints—the framework maintained functional stability. These omissions did not impede the pipeline, as the core semantic primitives were preserved sufficiently to enable successful matching in the subsequent $\mathcal{S}$ stage. These results validate the initial effectiveness of the inquiry meta-ability while highlighting specific directions for optimizing the precision of generated retrieval prompts.

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

This section examines how the operator $\mathcal{I}$ executes logical operations by treating non-parametric memory as a set of governing constraints, effectively suppressing pre-trained parametric priors.

5.4.1 Implementation: Reflective Reasoning via Think-Block

In the CoG-MeM framework, $\mathcal{I}$ is operationalized through a structured `<think>` block. This intermediate reasoning space allows the model to perform "logical arbitration"—explicitly referencing retrieved memory IDs and substituting counterfactual rules into its derivation process before generating a final response.

5.4.2 Preliminary Observations: Domain-Agnostic Deduction and Conflict Resolution

The evaluation across 124 cases highlights $\mathcal{I}$'s flexibility in handling diverse symbolic and narrative constraints:

- **Mathematical and Physical Substitution**: In the *Transformer* and *Casino Chip* scenarios, the model successfully ignored standard physical laws and financial logic. It accurately applied counterfactual formulas (e.g., $U_1 \times n_1 = U_2 \times n_2^2$ and $G = C \times (100 - N)/100$), demonstrating that the inference engine can perform rigid algebraic substitution even when the logic contradicts universal constants.
- **Complex Rule Application (Law & Magic)**: The operator proved capable of multi-step logical derivation. In the *Exodar Law* case, the model correctly identified the "doubling" condition for airway chaos, transforming a base fine of 250 into a final 500 gold coins. Similarly, for *Molten Mantle*, it accurately mapped specific status effects to relevant targets (allies vs. enemies) as dictated by the retrieved ritual rules.
- **Conflict Arbitration via Temporal Priority**: A significant observation involves the model's ability to resolve contradictory memories. When presented with two competing probability formulas ($P(A \cap B)$), the model utilized the `time` metadata to override the older entry ([mem-id: 20]) in favor of the more recent specification ([mem-id: 23]), ensuring that the non-parametric knowledge remains plastic and up-to-date.
- **Graceful Fallback to Parametric Priors**: A crucial observation of $\mathcal{I}$ is its ability to maintain task continuity when memory retrieval yields no results. The model demonstrates a bifurcated fallback strategy:
  1. **Domain-Specific Restoration**: For universal fields like physics and mathematics, when special Azerothian rules are missing, the model correctly reverts to its parametric "baseline" knowledge (e.g., using $W = F \times s$ or standard differentiation) while explicitly notifying the user of the absence of external constraints.
  2. **Collaborative Logic Augmentation**: In scenarios where retrieved external knowledge is identified as insufficient to fully resolve the user's query, the system facilitates the measured use of internalized knowledge as a supplement.
  3. **Strict Failure for Fictional Constructs**: In purely counterfactual domains without parametric counterparts, such as the *Frost Grip* magic effect, the model avoids hallucination by admitting a lack of information.

This illustrates that $\mathcal{I}$ does not merely "copy" memory, but performs active logical weighing between retrieved context and internal priors.

5.4.3 Summary of Inference Stability

Across the 124 test cases, **ten reasoning errors** were identified during the Constrained Inference stage, primarily concentrated in two categories: (i) **Computational Inaccuracies**, where the system successfully identified and retrieved the correct external formulas but failed to execute the precise arithmetic steps within the "Think" field; and (ii) **Hallucination under Deficit**, where the model failed to trigger the "I don't know" safety mechanism and instead generated hallucinated content when both external memory and internal priors were insufficient to resolve the query. Despite these edge cases, this error distribution provides preliminary evidence for the effectiveness of "Constrained Inference" as a core meta-ability. It suggests that the decoupling of reasoning from parametric weights via structured reflection enables the model to reliably utilize external knowledge, maintaining logical consistency across both formulaic and descriptive task environments even in the presence of conflicting information.

5.5 Overall Performance and the Criticality of Meta-Abilities

This section synthesizes the end-to-end pipeline performance across the 124 test cases in the CoG-MeM study. The results provide an empirical foundation for the NPMCL framework, illustrating how the four operators—$\mathcal{C}, \mathcal{G}, \mathcal{S}$, and $\mathcal{I}$—function as an interdependent logical loop.

5.5.1 Error Analysis and Loop Integrity

Out of the 124 diverse scenarios spanning physics, finance, law, and magic, the system recorded **17 total loop failures**. A granular analysis of these failures highlights the mechanical necessity of each meta-ability:

- **Matching Failures (7 cases)**: Errors in $\mathcal{S}$ led to the retrieval of irrelevant or incomplete memory indices, breaking the contextual anchor required for reasoning.
- **Inference Failures (10 cases)**: Even with successful memory retrieval, failures in $\mathcal{I}$ occurred when the model failed to suppress its pre-trained weights or misapplied the retrieved logical constraints during derivation.

Although no **Triggering ($\mathcal{G}$)** or **Compression ($\mathcal{C}$)** failures were directly observed in this specific proof-of-concept, their mechanistic necessity remains absolute within the NPMCL framework. Due to the strictly sequential architecture of the four operators, any failure in the initial stage—such as a failure to initiate the inquiry ($\mathcal{G}$)—would render all downstream processes (matching and inference) impossible. Therefore, the triggering meta-ability functions as the "ignition" for the cognitive loop, ensuring that the system can autonomously recognize the need for external knowledge.

Furthermore, Distillative Compression ($\mathcal{C}$) serves as a vital safeguard against the "Lost in the Middle" phenomenon [21]. Any loss of core semantic primitives during the compression stage would inevitably lead to the downstream failure of the $\mathcal{I}$ operator due to the absence of critical logical pillars. Thus, the integrity of the entire non-parametric adaptation cycle relies on the seamless execution of all four interdependent operators.

5.5.2 Concluding Remarks

The preliminary evaluation of CoG-MeM validates the importance of meta-abilities within the NPMCL architecture. By structuring the model's inherent linguistic sensitivity into a

formal loop of inquiry, retrieval, and constrained Inference, the framework demonstrates a viable pathway for achieving dynamic plasticity.

6 Experiments on Logic-Chain Knowledge Assimilation and Application

6.1 Experimental Setup and Data Organization Strategy

To rigorously evaluate the framework's capacity for structural preservation and constrained reasoning, we establish a formal evaluation methodology by decoupling abstract logical inference paths from their semantic environments. We construct our compression training and testing sets, as well as our constrained inference training and testing sets, based on the following multi-tiered data organization strategy.

6.1.1 Design of Logical Chain Skeletons

We first define a diverse set of abstract structural skeletons that represent core rules of formal logic. It is critical to note that the forms listed below constitute only a representative subset of the extensive logical skeletons utilized in this work:

- *Sequential Execution Chains*: Step 1: $A \Rightarrow$ Step 2: $B \Rightarrow$ Step 3: $C \Rightarrow$ Finally: D (representing strict step-by-step sequential operations).
- *Disjunctive Conditional Anchors*: If $A \vee B \rightarrow C$.
- *Parallel Condition Mappings*: If $A_1 \rightarrow B_1$, If $A_2 \rightarrow B_2$, If $A_3 \rightarrow B_3$.

6.1.2 Training Set Generation via Semantic Instantiation

To generate the training instances, we perform semantic instantiation by populating the aforementioned foundational skeletons with rich, domain-specific text. To prevent the model from exploiting its parametric memory shortcuts, we deliberately inject content from specialized and fictional environments:

- **Fictional Military and Legal Domain**: Illustrated by the *Stormwind City Defense Codex* in Azeroth, such as:
  1. If the Horde zeppelin conducts reconnaissance without firing $\rightarrow$ intercept and drive away via Gryphon Knights.
  2. If the zeppelin drops incendiary devices or bombs $\rightarrow$ activate anti-air mana cannons for elimination **AND** close the harbor magic shield to block splash damage.
  3. If the zeppelin deploys a spell-shield and attempts breaching the walls $\rightarrow$ trigger the Holy Light annihilation array.
- **Fictional Corporate Management and Factory Production Regulations**: Rules governing standard operational procedures, safety boundaries, and administrative workflows in simulated industrial settings.

6.1.3 Test Set Expansion: Novel Skeletons and Unseen Domains

To strictly evaluate the framework's Out-of-Distribution (OOD) generalization and structural flexibility, the test set introduces both unseen logical structures and an entirely new semantic domain:

- **Newly Added Logical Skeletons**: The testing phase incorporates more complex, extended logic configurations, including but not limited to:
  - *Extended Parallel Mappings*: If $A_1 \to B_1$, If $A_2 \to B_2$, If $A_3 \to B_3$, If $A_4 \to B_4$.
  - *Conditional Aggregation Chains*: If $A_1 \to B_1$; If $A_2 \to B_2$; If $A_3 \to B_3$, and if any of the above conditions are satisfied $\to B_4$.
- **Newly Added Semantic Domain**: We inject raw corpora from a completely unencountered sector—*School Campus and Academic Regulations*—forcing the model to execute non-parametric adaptation under novel environmental constraints.

Through this systematic construction pipeline, we successfully developed a comprehensive evaluation suite. Based on these datasets, we ultimately compiled an empirical demonstration demo consisting of **75 high-quality data instances** to verify the whole lifecycle of knowledge assimilation and application under the NPMCL framework.

6.2 Knowledge Compression Pipeline and Distillation Evaluation

The primary objective of the Knowledge Compression operator ($\mathcal{C}$) is to ingest noisy, multi-turn conversational data streams and distill them into highly structured, dense, and complete logical representations. To train and evaluate this meta-ability, we implement a structured pipeline utilizing intermediate reasoning behaviors.

6.2.1 Data Formatting and Prompt Templates

We structure our raw knowledge corpus into standard multi-turn dialogue blocks between a user and an assistant to simulate decentralized, non-structured real-world info streams. A representative instance from the simulated industrial setting is illustrated below:

6.2.2 Intermediate Reasoning Chains (Think CoT)

To cultivate the model's precise localization and alignment abilities, we mandate the generation of an internal *Think Chain-of-Thought (CoT)* trajectory prior to emitting the final compressed output. For the instance above, the target training output is formulated as:

Model Trace (Think CoT):
Parsing safety protocols for Heimai Mining Ordinance. Base condition: Mine car derailment → warn personnel + initiate buffer/deceleration. Exceptional constraint: Sliding + brakes jammed + heading for pillars → signals malicious sabotage/structural collapse hazard. Strategy: Bypass conventional deceleration; directly trigger on-board detonators for pre-emptive destruction in open section.
Final Compressed Memory:

```
<memory>
Heimai Mining Underground Accident Handling Ordinance:  In the event of
a deep mine car derailment and sliding, personnel below must be immediately
warned to seek cover, and buffer-deceleration mechanisms must be engaged.
However, if the brakes are discovered to be jammed during the slide while
the vehicle is heading directly toward load-bearing pillars, the incident
is classified as malicious sabotage threatening a structural mine collapse.
Under this state, the system must immediately bypass all deceleration procedures
and directly detonate the on-board detonators to destroy the mine car beforehand
within an open section.
</memory>
```

Table 1: Simplified Multi-turn Dialogue Example for Compression Input

System Prompt: Extract and condense the knowledge tokens from the dialogue block into a structural summary enclosed within `<memory></memory>` tags.
Condensed Dialogue Stream: **User:** Next, explaining the Heimai Mining Underground Accident Handling Ordinance. If a deep mine car derails and slides, you need to immediately warn personnel below to take cover and attempt to buffer and decelerate. **Assistant:** Understood. Mine car derailing and sliding → yell for people to hide, then try to brake and slow down. **User:** But if, during the slide, the brakes are found to be jammed and it is heading straight for the load-bearing pillars, it is judged as malicious sabotage triggering a mine collapse. **Assistant:** Brakes jammed deliberately while heading for load-bearing pillars? Trying to collapse the whole mine! Too vicious! **User:** At this point, must immediately bypass the deceleration procedure, and directly detonate the detonators on the mine car to destroy it in an open section beforehand. **Assistant:** Received. If it's heading for load-bearing pillars, deceleration is hopeless. Must detonate the detonators first, destroy the car in an open area to save the mine.

By pairing the dialogue inputs with these dual-target outputs (CoT + Memory), we train the model to systematically map conversation fragments into clean, non-parametric logic rules. In this process, the model is guided to first systematically organize and analyze the underlying knowledge structure within the Chain-of-Thought (CoT) prior to synthesizing the final memory item, thereby ensuring the completeness and correctness of the extracted knowledge. For this stage, we compile a training set of **300 instances** and an independent test set of **170 instances**.

6.2.3 Empirical Performance and Comparative Analysis

To evaluate the compression quality with high stringency, we perform a comprehensive human-in-the-loop audit on the test set. An output is marked as strictly correct only if it achieves a full match with the human-annotated ground truth in both its structural logic skeleton and its granular semantic node content (allowing only minor non-material paraphrasing).

Table 2: Compression Efficacy Comparison

Method	**Accuracy**	**Primary Failure Modes**
NPMCL ($\mathcal{C}$-operator)	**95%**	Minor stylistic variations, minor structural omissions.
Base Model + Zero-Shot Prompt	85%	Hallucinations, factual omission, inclusion of irrelevant dialogue noise.

As indicated in Table 2, the performance gap highlights the value of the fine-tuned $\mathcal{C}$-operator. Without explicit meta-training, the baseline model suffers from degraded extraction accuracy and a propensity for cognitive hallucinations. We recognize that the current setup focuses on formal, template-driven instructional dialogues; however, scaling the $\mathcal{C}$-operator to distill volatile knowledge from complex, high-entropy daily conversations remains a valuable open challenge for future research.

6.3 Constrained Inference Engine and Knowledge Execution

The Constrained Inference operator ($\mathcal{I}$) aims to guarantee that the system can faithfully execute downstream multi-step reasoning under the absolute boundary constraints of the externalized, non-parametric memory fragments.

6.3.1 Data Schema and Execution Protocol

To evaluate this reasoning rigor, we formalize the execution protocol into a structured multi-role transaction pattern. When faced with an environmental query, the model must explicitly trigger a memory query call, retrieve the relevant logic segment, generate an internal thinking trajectory to parse the branches, and subsequently produce the final response.

Table 3: Data Instance for Constrained Inference Phase

[**User Question**] We detected unauthorized invisible barrier fluctuations at Stormwind Harbor, and the approaching vessels are flying Bloodsail Pirate flags and refusing to halt! According to the Harbor Defense Regulations, how should we proceed?

[**Assistant (Tool Call)**]
`<function>memory_query_call</function>`
`<content>Stormwind Harbor Defense Regulations</content>`

[**Retrieved Non-Parametric Memory**] **Relevant Fragment [mem-id:18001]:** Stormwind Harbor Defense Regulations: If unauthorized invisible barrier fluctuations are detected → immediately activate Towers of Counter-Stealth Vision **AND** lock down the portal network. Subsequently evaluate intrusion status: If approaching vessels fly pirate flags and refuse to halt → sink immediately via coastal batteries; If vessels are unarmed and signal distress → escort to the quarantine dock.
[**Assistant (Think CoT)**] *Memory [mem-id:18001] specifies the defense regulations. The user describes invisible barrier fluctuations, which mandates activating the anti-stealth towers and locking the portals. Furthermore, the vessel flies a pirate flag and refuses to halt, matching the first conditional branch. Under this protocol, it must be directly sunk via coastal batteries.*
[**Assistant (Final Content)**] According to the Stormwind Harbor Defense Regulations, you must immediately activate the Towers of Counter-Stealth Vision and lock down the portal network! Subsequently, the vessels must be directly sunk via coastal batteries!

Under this paradigm, the intermediate *Think* trajectory forces the system to execute formal variable assignment and rule tracking, preventing it from skipping operational dependencies. Specifically, the Chain-of-Thought (CoT) systematically constrains the model to perform step-by-step logical derivation governed strictly by the provided external rules, ultimately ensuring that both the final structural logical forms and the instantiated contents are simultaneously accurate.

We assemble a training set consisting of **300 instances** and a testing set consisting of **200 instances** across varying logic configurations.

6.3.2 Empirical Verification and Baseline Contrast

The verification metric adheres to the strict dual-criterion framework: a prediction is flagged as correct only if both its sequential execution skeleton and its instantiated granular

actions are completely aligned with the human-curated ground truth (tolerating minor non-material phrasing).

The fine-tuned inference operator ($\mathcal{I}$) under the NPMCL framework achieves an execution accuracy of **88%** on the test set. In comparison, utilizing the frozen base model via purely zero-shot prompt engineering presents critical operational vulnerabilities:

- **High Error Propensity:** The baseline model exhibits an approximate **20% error rate** directly attributed to either logical skeleton errors or filled content errors within the valid nodes.
- **Severe Cognitive Hallucination:** Over **60% of the evaluation items** under the zero-shot baseline suffer from pure cognitive hallucinations, where the model independently fabricates novel regulations or unmentioned conditional paths entirely beyond the boundaries of the injected knowledge text.

This distinct empirical margin confirms that relying purely on textual in-context steering is insufficient for complex procedural reasoning. The dedicated meta-training for the $\mathcal{I}$-operator is indispensable for binding the LLM's autoregressive generation tightly within the strict deductive boundaries of non-parametric knowledge.

6.4 End-to-End Pipeline Demonstration and Framework Synergy

To validate the seamless orchestration of the non-parametric meta-abilities under the NPMCL framework, we construct a comprehensive, end-to-end evaluation pipeline. This stage serves as a macro-level demonstration to verify whether the decentralized operators can collaboratively process and utilize knowledge when subjected to entirely unseen logical configurations.

6.4.1 Leveraging Legacy Assets and Data Co-adaptation

To endow the frozen foundation model with standardized conversational routines and rigorous tool-triggering behaviors, we repurpose the foundational fine-tuning assets from the CoG-MeM prototype [1]. Specifically, we reuse the specialized training data designed for the *Query Generation Operator* ($\mathcal{G}$) alongside historic multi-turn interaction dialogues to alignment-tune our model's basic communication boundaries. Furthermore, the *Structural Matching Operator* ($\mathcal{S}$) inherited from CoG-MeM is directly integrated into our inference pipeline to act as a rigorous semantic gatekeeper.

6.4.2 Demo Protocol under Zero-Shot Skeletons

To introduce a high-difficulty evaluation threshold, we curate a brand-new demonstration dataset consisting of **75 paired dialogue instances**. Crucially, all 75 instances are instantiated using complex *logical skeletons completely excluded from the training sets* of both the $\mathcal{C}$ and $\mathcal{I}$ operators, representing a strict zero-shot generalization test for abstract structure tracking.

The empirical workflow for the demonstration is executed through a three-phase cascade:

1. **Knowledge Assimilation Phase:** The system first ingests 75 extensive, raw tutorial dialogues. Utilizing the fine-tuned $\mathcal{C}$-operator, it compresses them into 75 high-density, structure-preserving core knowledge tokens stored in the external memory bank.

2. **Coarse Semantic Retrieval:** When a user poses a downstream reasoning problem, the system executes an initial coarse selection using embedding-based semantic similarity, fetching the **Top-30 most relevant** knowledge candidates from the repository.

3. **Fine Algorithmic Filtering and Inference:** The structural matching operator ($\mathcal{S}$) is then deployed to scrub the 30 candidates, isolating the exact, hard-constrained logical rules matching the problem context. Finally, the $\mathcal{I}$-operator ingests these exact rules to formulate the final answer.

6.4.3 Empirical Outcomes

Out of the 75 highly complex, unseen logical evaluation trials, the unified NPMCL pipeline successfully delivers structure-congruent and semantically correct reasoning chains in **63 instances** (achieving an overall success rate of **84.0%**). Given that the logical skeletons utilized in this test suite were entirely novel to the underlying model, this robust performance demonstrates the compelling potential of the NPMCL architecture. It empirically confirms that decoupled, training-free meta-abilities can successfully internalize and execute highly sophisticated cognitive protocols without relying on massive parametric update cycles.

To ensure a fair and controlled evaluation, we conducted comparative experiments against two prominent memory-centric frameworks, MemGPT [17] and Mem0 [18], all standardized on the identical Qwen3-8B [19] backbone. Due to the lack of specialized optimization within these baselines for structured knowledge retention and constrained reasoning, both frameworks exhibited clear performance bottlenecks. Specifically, MemGPT yielded an overall accuracy rate below **40%**, which can likely be attributed to its overly complex prompt overhead; we speculate that this heavy text burden induces control-flow disruption, frequently leading to a failure to trigger the memory retrieval primitives during active dialogue. By contrast, Mem0 achieved a accuracy rate of approximately **65%**. Since Mem0 does not natively provide a memory-integrated chat pipeline, we implemented an equivalent standard chat interface utilizing a conservative similarity threshold to screen memories. The empirical failures captured in the Mem0 pipeline revealed systematic informational distortions across different stages: a significant portion of errors stems from cognitive hallucinations introduced early during the unconstrained key-value extraction stage, while other instances suffer from reasoning-induced deviations during the final response generation phase, ultimately leading to critical structural skeleton omissions and incorrect content payloads when dealing with complex logical rules.

7 Mechanistic 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.

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

In our framework, the model's ability to rewrite raw 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.

7.2 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 [20], executable solution within a specific context.

7.3 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 posit 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.

7.4 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.

7.5 Summary

Within an information-theoretic framework, **Distillative Compression** ($\mathcal{C}$) and **Constrained Inference** ($\mathcal{I}$) function as a knowledge codec: $\mathcal{C}$ represents the efficient compression of high-entropy data into logic-dense invariants, while $\mathcal{I}$ serves as the decompression and reconstruction process for reasoning. This framework mirrors the human cognitive cycle, where compression equates to the **absorption** of experience into rules, and inference signifies the **application** of that knowledge to novel tasks.

The essence of training $\mathcal{I}$ within the **NPMCL** framework lies in cultivating the model's ability to operationalize knowledge. While standard RAG systems typically supplement models with factual data—thereby reinforcing the retrieval and integration of

information—the training for Constrained Inference involves external data containing explicit logical chains. This specialized regime reinforces the model's capacity to apply foundational logical primitives—such as contextual substitution, causal tracing, combinatorial judgment, inductive-deductive synthesis, and so forth—across diverse and dynamic logic environments.

Although existing RAG systems exhibit these capabilities to some extent, recent work has highlighted their inherent limitations, particularly when facing persistent knowledge conflicts between external context and internal priors [23]. Consequently, we maintain that a dedicated training phase focused on reasoning under external constraints is indispensable for empowering models to navigate the ever-evolving logical landscapes of real-world applications.

8 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**.

8.1 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 [22]:

- **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.

8.2 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), multi-modal 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 [24], 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.

8.3 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** [25]. For instance, Structural Matching ($\mathcal{S}$) can be realized via optimized RAG pipelines. Although semantic retrieval in RAG may introduce partial noise, the overall system performance may remain stable through improved noise suppression and logical filtering of the Constrained Inference ($\mathcal{I}$) operator. This modular flexibility grants the framework preliminary potential for real-world production deployment, adapting to hardware constraints and existing legacy infrastructures.

- **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 Inference ($\mathcal{I}$)—significantly streamlining troubleshooting and maintenance.

9 Conclusion

This mechanistic 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 Inference in CL: Unlike previous works focused on isolated knowledge conflicts [26, 27], we integrate **Prior Suppression and Constrained Inference** 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.

Empirical Circuit: We compiled a specialized knowledge dataset formalized around structural logical skeletons and instantiated content payloads. Completing the loop of knowledge ingestion and application on these test suites provides an opportunity to observe the system's ability to reason within external factual boundaries (i.e., constrained inference), thereby offering a preliminary indication of the empirical feasibility of the NPMCL framework.

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 mechanistic 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., Pascanu, R., Rabinowitz, N., 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. (2015). Memory Networks. *ICLR*.

[5] Gutiérrez, B. J., Shu, Y., Qi, W., et al. (2025). From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. *ICML*.

[6] Deletang, G., Ruoss, A., Duquenne, P., et al. (2024). Language modeling is compression. *ICLR*.

[7] Dittrich, C., & Flygare Kinne, J. (2025). The Information-Theoretic Imperative: Compression and the Epistemic Foundations of Intelligence. *arXiv:2510.25883*.

[8] Piaget, J. (1952). *The Origins of Intelligence in Children*. New York: International Universities Press.

[9] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*.

[10] Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *ITW*.

[11] Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., et al. (2023). Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. *Findings of ACL*.

[12] Yin, Z., Sun, Q., Guo, Q., et al. (2023). Do Large Language Models Know What They Don't Know? *Findings of ACL*.

[13] Ausubel, D. P. (1968). *Educational Psychology: A Cognitive View*. New York: Holt, Rinehart and Winston.

[14] Shao, Z., Wang, P., Zhu, Q., et al. (2024). DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300*.

[15] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

[16] Schick, T., Dwivedi-Yu, J., Dessì, R., et al. (2023). Toolformer: Language models can teach themselves to use tools. *NeurIPS*.

[17] Packer, C., Wooders, S., Lin, K., et al. (2023). MemGPT: Towards LLMs as Operating Systems. *arXiv:2310.08560*.

[18] Chhikara, P., Khant, D., Aryan, S., et al. (2025). Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *arXiv:2504.19413*.

[19] Yang, A., Yang, B., Zhang, B., et al. (2025). Qwen3 Technical Report. *arXiv:2505.09388*.

[20] Lyu, Q., Havaldar, S., Stein, A., et al. (2023). Faithful Chain-of-Thought Reasoning. *IJCNLP-AACL*.

[21] Liu, N. F., Lin, K., Hewitt, J., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *TACL*.

[22] Zhong, Z., Guo, L., Gao, Q., et al. (2024). MemoryBank: Enhancing Large Language Models with Long-Term Memory. *AAAI*.

[23] Li, G., Chen, Y., & Tong, H. (2025). Taming Knowledge Conflicts in Language Models. *ICML*.

[24] Karpukhin, V., Oğuz, B., Min, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *EMNLP*.

[25] Pfeiffer, J., Rücklé, A., Poth, C., et al. (2020). AdapterHub: A Framework for Adapting Transformers. *EMNLP*.

[26] Li, D., Rawat, A., Zaheer, M., et al. (2023). Large Language Models with Controllable Working Memory. *Findings of ACL*.

[27] Lin, X. V., Chen, X., Chen, M., et al. (2024). RA-DIT: Retrieval-Augmented Dual Instruction Tuning. *ICLR*.