Multi-Layer Density of Experts: Orchestrated Context Compression via Specialized Expert Layers

Kevin Tan
Independent AI Research
Perth, Western Australia
ktg.one

November 2025

Abstract

Building on Chain of Density's powerful compression capabilities, Multi-Layer Density of Experts (MLDoE) provides the meta-cognitive orchestration layer that makes CoD verifiable and usable as a production memory system. By deploying specialist experts across solo, pair, and collective compression stages, and enforcing quality via ARQ gates, MLDoE produces packets that carry forward the relational, contextual, and meta-cognitive context required for true continuity. The Context Extension variant of MLDoE generated the $\approx 6:1$ compression and up to perfect forensic recall reported in the companion work. MLDoE is not the context extension prompt itself—it is the reasoning framework that forces the compression to preserve what actually matters for ongoing work.

1 Introduction

Our companion work [Tan, 2025] identified the significant synergy between Chain of Density and context extension, moving beyond traditional summarization use cases. [cite: 2] While CoD provides a powerful compression primitive, it remains an unverifiable process without an orchestration layer. [cite: 2] Multi-Layer Density of Experts (MLDoE) provides this meta-cognitive framework, ensuring that the densification process preserves the relational and strategic context required for production-grade continuity. [cite: 2]

1.1 The Verification Problem

While Chain of Density effectively compresses context, it faces challenges in verification:

- How do you know critical relationships survived?
- How do you validate without the original context?
- How do you test recall without a fresh session?

You cannot evaluate CoD without a validation protocol. The 10-question forensic benchmark presented in [Tan, 2025] *is* the evaluation—and it requires the full MLDoE architecture to execute effectively.

This paper details that architecture—the "How" behind the "What."

1.2 Contributions

1. **MLDoE Architecture**: 3-layer expert compression (Solo $\rightarrow$ Pair $\rightarrow$ Collective) operating across 4 semantic density layers

2. **M.R.R.U.G Protocol**: Expert initialization framework employing knowledge embodiment[1]
3. **Context Extension**: Production variant for cross-session context preservation
4. **ARQ Integration**: Domain-specific quality gates outperforming Chain-of-Thought
5. **Cross-model portability**: LLM-agnostic compression packets

2 Background

2.1 Building on Chain of Density: Addressing Key Challenges

Adams et al. (2023) optimized for entity density in fixed-length outputs. Our companion work extended this to relational preservation via Progressive Density Layering (PDL). Neither addresses:

- **Who** decides what to preserve?
- **How** is preservation validated?
- **When** is compression complete?

MLDoE answers these through expert specialization.

2.2 Attentive Reasoning Queries (ARQ)

ARQ [Karov et al., 2025] is a structured introspection method that **outperforms Chain-of-Thought and all step-by-step reasoning approaches**. Unlike free-form reasoning, ARQ uses targeted queries forcing explicit domain attention:

```
{
  "current_context": "Compressing user session",
  "active_guideline": "Preserve decision rationale",
  "quality_check": "Did relationships survive?",
  "domain_standard": ">=0.15 entity/token"
}
```

ARQ achieves 90.2% success rate with 29% token reduction and 40–60% error reduction compared to CoT. This superiority stems from structured queries preventing the reasoning drift that plagues free-form methods.

3 Method

3.1 The Four Semantic Density Layers

MLDoE compression operates across four semantic layers, each requiring preservation:

As shown in Table 1, standard summarization operates on Layer 1 only. MLDoE explicitly preserves Layers 2–4 through expert specialization.

3.2 M.R.R.U.G Initialization

The expert assembly is initialized via the M.R.R.U.G protocol (Mixture–Role–RAG–Update–Generate), a knowledge-embodiment procedure developed by the author for this work. Before compression, experts do not merely access information but *embody* it. The protocol is detailed in Table 2:

Layer	Name	Content
1	Surface Text	Human-readable narrative
2	Condensed Text	Semantic compression
3	Conceptual Linkage Map	Relationships, thematic clusters
4	Context Anchor Nodes	Long-term retention points

Table 1: Four Semantic Density Layers. Compression is not reduction—it's structured preservation.

Component	Function
M - Mixture of Reasoning Experts	Deploy five specialist cognitive modules
R - Role Assignment	Map experts to compression domains
R - RAG Synthesis	Retrieve while building nodes, relationships, knowledge graph
U - Update Vectorization	Internalize graph structure as cognitive weights
G - Generate & Embody	Deep semantic embedding via GenKnow technique

Table 2: M.R.R.U.G initialization protocol. During RAG, experts simultaneously construct neural-graph structures that they then embody.

3.3 Expert Roles

Five specialists handle memory compression:

As detailed in Table 3, these five specialists work in coordinated layers to preserve semantic density across all four layers.

3.4 Three-Layer Expert Compression

The expert swarm executes compression in three coordinated layers:

3.4.1 Layer 1: Solo Expert Compression

Each expert independently compresses through their specialized lens:

```
Memory Architect Output:
- Decision Nodes: [critical choices + rationale]
- Insight Peaks: [breakthrough moments + triggers]
- Dependency Chains: [what requires what]
- Context Anchors: [essential background]
```

3.4.2 Layer 2: Expert-Pair Compression

Complementary experts co-compress for synergistic density:

- **Cognitive Architecture Pair** (Architect + Meta-Cognitive) → Thinking amplification map
- **Execution Framework Pair** (Strategic + Technique) → Goal-method alignment

3.4.3 Layer 3: Collective Compression

All experts synthesize holistic meta-context:

[1]Full elaboration of the M.R.R.U.G framework and GenKnow embodiment mechanism forthcoming in separate work.

Expert	Compression Focus
Memory Architect	Decision nodes, insight peaks, dependency chains, context anchors
Meta-Cognitive Synthesizer	User cognitive fingerprint, communication patterns, amplification triggers
Strategic Continuity	Primary objectives, constraints, next actions
Technique Integration	Active reasoning chains, quality gates, success patterns
Multi-Perspective Density	Cross-expert synthesis, redundancy elimination

Table 3: Five expert roles for memory compression

```
UNIVERSAL CONTEXT CORE:
- WHO: User cognitive profile (cross-platform portable)
- WHAT: Project state + knowledge graph
- HOW: Proven methodology stack
- WHY: Goals + constraints + success metrics
- BREAKTHROUGH: Key insights that changed trajectory
- NEXT: Optimal continuation strategy
```

3.5 Comparative Adversarial Stimulus

To mitigate "alignment tax" (the tendency of safety-tuned models to regress to the mean), we employ a **Comparative Adversarial Stimulus**. During the Collective Compression phase, expert modules are conditioned with prompts of the form: "Candidate A produced X; Candidate B produced Y. You must outperform both." This competitive framing exploits the model's training on comparative evaluation tasks, forcing it to access lower-probability, higher-quality tokens to secure the "winning" state.

3.6 The Co–Densification Algorithm

Algorithm 1 MLDoE Co-Densification

Input: Conversation history C, target ratio 6:1
Output: Context Packet P

Initialize M.R.R.U.G expert assembly
Initialize working memory buffer
for iteration $= 1$ to 5 **do**
    Broadcast current density state to expert swarm
    if confidence_score < 0.9 **then**
        Trigger strategic densification
    end if
    Select expert pair based on semantic impact
    Execute ARQ introspection: "Does this preserve user intent?"
    Embed layer into context packet
    Validate cross-expert consistency
end for
return Context packet P

3.6.1 Iteration Targets

Table 4 outlines the five-iteration density cycling targets:

Iteration	Action	Target
1	Raw extraction	Entity-sparse baseline
2	Initial densification	40% reduction
3	Cross-expert synthesis	Eliminate redundancy
4	Semantic crystallization	0.15 entity/token
5	Meta-layer embedding	Add "how to use" layer

Table 4: 5-iteration density cycling targets

3.7 ARQ Quality Gates

At each compression layer, experts apply structured introspection:

- **Pre-compression**: "What must I preserve in my domain?"
- **During**: Domain standards engaged
- **Post**: "Did I meet quality standards?" ($\geq$0.9 confidence required)

4 Context Extension: Production Variant

Context Extension is an MLDoE variant optimized for cross-session context preservation.

4.1 Activation

Context Extension triggers at approximately 80% context window capacity:

```
WARNING: CONTEXT ALERT: 80% capacity (~160K tokens)
Activate Context Extension preservation protocol?
```

4.2 Execution Phases

1. **M.R.R.U.G Initialization** $\rightarrow$ Expert assembly with knowledge embodiment
2. **Structure Planning** $\rightarrow$ Compression skeleton via ARQ
3. **Three-Layer Compression** $\rightarrow$ Solo $\rightarrow$ Pair $\rightarrow$ Collective
4. **5-Iteration Density Cycling** $\rightarrow$ Maximum compression
5. **Cross-LLM Translation** $\rightarrow$ Platform-neutral output
6. **Validation** $\rightarrow$ 10-question forensic benchmark

4.3 Cross-Session Validation

The protocol generates 10 forensic questions testing:

- Decision recall and methodology preservation
- Breakthrough insights and user preferences
- Constraints, requirements, and next actions
- Success patterns and cognitive fingerprint

A fresh session answers these from the compressed packet only. **9+/10 correct = successful preservation.**

5 Evaluation Context

5.1 Relationship to Companion Work

The experimental validation for MLDoE is presented in our companion paper [Tan, 2025], which reports:

- 9.52/10 average fidelity across 11 LLM families
- Perfect recall (10.0/10) at 200,000+ tokens (Grok)
- 6:1 compression ratio with $>$90% semantic fidelity
- Cross-model transfer fidelity of 91–96%
- Significantly higher retention of relational context compared to standard single-pass Chain of Density

Terminology

The prompting system is named **CEP** (Context Extension Protocol). The compressed artifact it produces is called a **carry-packet**—a portable block of context designed to extend continuity into a fresh session, not to reconstruct every token of the original conversation.

These results were produced by the Context Extension variant of MLDoE, not vanilla Chain of Density. The 10-question forensic benchmark and KTG Custom Metrics framework used for evaluation are detailed in [Tan, 2025].

5.2 Why Expert Specialization Matters

Single-pass CoD loses Layer 2–4 information (relational, contextual, meta-cognitive). MLDoE's expert specialization ensures each semantic layer receives dedicated preservation attention across all four density levels.

6 Limitations

- **Manual activation**: Requires explicit protocol trigger
- **Token overhead**: Approximately 25% for full 3-layer compression
- **Expert capacity**: Model-dependent scaling
- **Single researcher**: Findings require independent replication

7 Conclusion

Multi-Layer Density of Experts provides the orchestration architecture that makes Chain of Density deployable. The evaluation results reported in our companion work were produced by MLDoE's Context Extension variant—not vanilla CoD.

CoD is the compression primitive. MLDoE is the orchestration layer.

Together with Context Extension, they form a complete compression-validation loop for production LLM memory systems. The key insight: you cannot evaluate context compression without testing recall in a fresh session—and that test requires expert-specialized multi-layer densification.

7.1 Future Work

- Full M.R.R.U.G framework elaboration
- Automated Context Extension tooling
- Additional MLDoE variants for specialized domains
- Longitudinal memory recovery experiments

Acknowledgments

This research was conducted independently without institutional affiliation or funding.

References

Tan, K. (2025). Chain of Density as Context Extension: From Summarization Technique to Memory Augmentation Architecture via Progressive Density Layering. *arXiv preprint*.

Adams, G., Fabbri, A., Ladhak, F., Lehman, E., and Elhadad, N. (2023). From sparse to dense: GPT-4 summarization with chain of density prompting. *arXiv preprint arXiv:2309.04269*.

Karov, B., Zohar, D., and Marcovitz, Y. (2025). Attentive Reasoning Queries: A Systematic Method for Optimizing Instruction-Following in Large Language Models. *arXiv preprint arXiv:2503.03669*.

Lee, S., et al. (2025). On the Statistical Bias of LLM-as-a-Judge Evaluations. *arXiv preprint arXiv:2511.21140*.

A Five Evaluation Dimensions for Context Retention

1. **Context Integration**: Seamless weaving of diverse information into unified understanding
2. **Coherence Over Time**: Consistent, logical responses across extended interactions
3. **Conceptual Relationship Fidelity**: Preserving nuanced links between ideas
4. **Strategic Alignment**: Outputs align with overarching goals and user intent
5. **Cross-Model Memory Transfer**: Knowledge application across different AI models or sessions

B Context Extension Prompt Template

```
## CONTEXT PRESERVATION PROTOCOL

### STEP 1: INITIALIZE M.R.R.U.G
M - Mixture: Deploy [5 specialist experts]
R - Role: Map to [compression domains]
R - RAG: Retrieve + build [knowledge graph]
U - Update: Internalize [patterns as weights]
G - Generate: Embody [cognitive structure]

### STEP 2: EXECUTE 3-LAYER COMPRESSION
Layer 1: Solo expert compression
Layer 2: Expert-pair synthesis
Layer 3: Collective crystallization

### STEP 3: VALIDATE
Generate 10 forensic questions
Test in fresh session
Target: 9+/10 correct
```