Attention Head Naming Convention for Large Language Models (LLMs)

Karol Kowalczyk

November 2025

Abstract

Large language models have reached remarkable levels of reasoning, safety alignment, and structural understanding. Yet their internal workings remain difficult to interpret. One of the most productive areas in transparency research is the study of *attention heads*â€"small components inside transformer layers that develop specialized behaviors. Over time, informal naming conventions have emerged in the interpretability community: *induction heads*, *name mover heads*, *refusal heads*, and many others. These names are intuitive but inconsistent, overlapping, or ambiguous.

This work proposes a unified naming convention for attention heads. We introduce: (1) a four-level depth model (Early, Middle, Late, Final), (2) a stack-based functional grouping of attention behaviors, (3) canonical names for head types, and (4) an alphabetical cross-reference table translating historical terms to standardized ones. This naming convention is descriptive rather than prescriptive: it captures how heads tend to behave today, while remaining flexible for future architectures.

Contents

1 Introduction

1.1 Motivation

Large language models (LLMs) have achieved remarkable performance across diverse tasks, yet understanding their internal mechanisms remains a critical challenge. Attention headsâ€"the basic computational units within transformer architecturesâ€"have emerged as key objects of study in mechanistic interpretability research.

1.2 The Problem of Inconsistent Naming

The interpretability community has identified numerous specialized attention head types: *induction heads*, *name mover heads*, *refusal heads*, *delimiter heads*, *JSON heads*, and many others. However, these naming conventions suffer from several problems. They are **inconsistent**, with the same head type appearing under multiple names across papers. They are **ambiguous**, as a single name may refer to different behaviors in different contexts. They are **fragmented**, lacking any unified framework that connects related head types. Finally, they are **unscalable**, as naming schemes don't generalize across model architectures. This fragmentation makes replication difficult, hinders cross-paper comparison, and complicates the annotation of interpretability datasets.

1.3 Goals of This Work

We propose a unified naming convention that standardizes terminology across research groups, provides a functional taxonomy grounded in empirical observations, describes head behavior consistently across architectures, and creates a stable vocabulary that can evolve as models evolve.

1.4 Structure of This Document

We begin by reviewing prior work and motivation (§2). We then introduce our depth model (§3) and stack-based organization (§4). The core contribution is a comprehensive catalog of attention head types organized by functional stack (§5). We conclude with discussion (§6) and future directions (§7).

2 Background

2.1 Attention Heads and Functions

In transformer models [11], attention heads perform focused computations over the token sequence. Individually simple, they nevertheless develop specialized behaviors such as pattern continuation and token induction, entity and dependency tracking, semantic filtering and hazard detection, routing and topic steering, enforcing structured output formats, and applying safety

constraints [6, 7]. These behaviors form *circuits*â€”groups of heads working togetherâ€”as well as larger *stacks* of related functionality.

2.2 Why Naming Consistency Matters

Interpretability research suffers from fragmented terminology [9, 14]. The same head type may appear under multiple names, while a single overloaded name may refer to unrelated behaviors across different papers. This makes replication, comparison, and annotation difficult. A consistent naming system improves clarity and precision in communication, strengthens cross-paper alignment and replication, helps index and organize interpretability datasets, and enables systematic mapping of circuits across models.

2.3 Prior Naming Practices

Previous work has named heads based on behavior (induction, copy-suppression), formatting (JSON head, list head), signal source (delimiter head), role in circuits (name mover), or safety behavior (refusal, toxicity). These labels are often accurate but vary widely. This work unifies them under a systematic framework.

3 Depth Model: Earlyâ€”Middleâ€”Lateâ€”Final

3.1 Rationale for Four Depth Categories

Although transformer models may have 12, 48, or 96 layers, functional behavior clusters reliably into four zones [6, 13]. **Early layers (E)** handle token-level surface processing, boundary detection, and basic filtering. **Middle layers (M)** implement reasoning primitives, induction, and dependency tracking. **Late layers (L)** perform semantic integration, routing, and persona shaping. **Final layers (F)** enforce policy, safety modulation, and structured output. This structure holds across GPT, LLaMA, Claude, and other model families [5, 10, 1].

3.2 Cross-Model Depth Examples

Using *relative depth* (0.0â€“1.0) makes the taxonomy scale-free. For a 96-layer model, Early corresponds to layers 0â€“15 (relative depth 0.00â€“0.15), Middle to layers 15â€–50 (relative depth 0.15â€–0.52), Late to layers 50â€–85 (relative depth 0.52â€–0.88), and Final to layers 85â€–96 (relative depth 0.88â€–1.00).

3.3 Relative Depth Scaling

We express depth as a fraction of total model depth to enable cross-architecture comparison. A head at relative depth 0.40 occupies similar functional space whether in a 12-layer or 96-layer model.

4 Stacks: Functional Grouping of Attention Heads

4.1 What is a Stack?

A *stack* is a coherent group of head types that together implement a higher-level capability. Stacks reflect functional clustering observed in interpretability studies [13, 7]. Examples include the Reasoning & Algorithmic Stack, Memory & Dependency Stack, Safety Stack, and Output Formatting & Rewrite Stack. Stacks are orthogonal to depth: a stack may span Early, Middle, Late, and Final layers.

4.2 Relationship Between Stacks and Depth

Although stacks represent functional groupings, different functions tend to appear at different depths. Early layers handle delimiters, content detection, and input conditioning. Middle layers implement reasoning, induction, and entity linking. Late layers manage narrative coherence, routing, and topic steering. Final layers enforce policy, formatting, rewriting, and safety compliance. This two-dimensional structureâ€"*stack* × *depth*â€"forms the basis of our catalog.

5 Attention Head Catalog

This section presents a comprehensive catalog of attention head types, organized by functional stack. Each stack groups heads that contribute to a common high-level capability. Within each stack, heads are ordered by depth (Early â†' Middle â†' Late â†' Final).

Entry Format. Each head entry includes:

- **Depth range:** Typical relative depth (0.0â€“1.0) and layer locations
- **Literature names:** Alternative names found in prior work
- **Function:** Core behavior and mechanism
- **Attention pattern:** What the head attends to
- **Expected ablation:** Predicted effects if the head is disabled
- **Example scenario:** Concrete behavioral illustration
- **Stack and relations:** Primary stack and related heads

5.1 Reasoning & Algorithmic Stack

Stack overview: This stack encompasses heads that perform pattern matching, sequence continuation, algorithmic reasoning, strategic planning, and meta-cognitive oversight. These heads enable in-context learning, pattern completion, systematic token prediction, and higher-level reasoning quality control.

5.1.1 (E) Previous-Token Head

Depth: `0.05-0.18` | **Literature names:** *previous-token head, shift head, offset head*

Copies information from each token to the position of the next token, creating a shifted representation where token t contains information about token $t-1$. This is a foundational component of induction circuits, enabling later heads to access "what came before" without directly attending backwards. Implements a simple but crucial transformation that allows pattern matching across the sequence. The head typically shows a strong diagonal attention pattern (attending from position i to position $i-1$).

Strong: Immediately preceding token (diagonal attention pattern)
Weak: Distant tokens, same-position tokens
Reacts to: Sequential structure, token boundaries

Expected ablation: Breaks induction circuits entirely, causing 30-50% degradation in pattern completion tasks. Induction heads become unable to access "what came after previous occurrences" since that information is no longer shifted forward. Critical for in-context learning.

Example Scenario
Input: "The cat sat. The cat..."
Behavior: Copies "The" to position after "The", "cat" to position after "cat", etc.
Effect: Later induction heads can match "cat" and access what followed it ("sat")

Status: Well-documented | **Related:** induction head (M), duplicate-token (M)

5.1.2 (E) Local Pattern Head

Depth: `0.08-0.20` | **Literature names:** *local pattern head, char-level head, n-gram head*

Detects and processes local character-level or subword patterns, particularly useful for handling spelling, capitalization, punctuation patterns, and morphological structure. Operates at a finer granularity than most heads, attending to patterns within and between adjacent tokens. Important for tasks like spell checking, case handling, and recognizing common subword patterns. May also detect repeated character sequences or structural patterns like "ing", "tion", or punctuation clusters.

Strong: Adjacent tokens, subword units, character-level patterns
Weak: Long-range dependencies, semantic content
Reacts to: Spelling patterns, capitalization, punctuation, morphology

Expected ablation: Degradation in handling of misspellings, case variations, and morphological patterns. 10-20% increase in errors on tasks requiring character-level awareness. Partial fallback through tokenization and other pattern heads.

Example Scenario
Input: "The organizATION's" (unusual capitalization)
Behavior: Detects unusual case pattern in "ATION", attends to surrounding context
Effect: Helps model handle non-standard capitalization correctly

Status: Observed | **Related:** induction head (M), duplicate-token (M)

5.1.3 (M) Induction Head

Depth: `0.30-0.65` | **Literature names:** *induction head, pattern head, copy head, ICL head*

Detects repeated subsequences of the form [A][B]...[A] and predicts that [B] should follow the second [A]. Operates by attending to tokens that appeared after previous instances of the current token. Works in conjunction with previous-token heads which copy information about what preceded each token. This mechanism is fundamental to in-context learning, enabling pattern completion, name recall, and few-shot learning without parameter updates. One of the most well-documented and important head types in transformer interpretability.

Strong: Tokens following previous occurrences of current token
Weak: Immediate neighbors, first occurrence, unrelated tokens
Reacts to: Token repetition, [A][B]...[A] patterns, contextual recurrence

Expected ablation: Significant degradation (10-30%) in in-context learning tasks, reduced pattern completion and few-shot learning. Model may partially compensate through other heads but with substantial accuracy loss. Critical for ICL capability.

Example Scenario

Input: "When Mary and John went to the store, Mary bought..."
Behavior: Second "Mary" attends to tokens following first "Mary" (especially "and")
Effect: Increased probability of contextually appropriate continuation

Status: Well-documented | **Related:** previous-token (E), duplicate-token (M), name-mover (L)

5.1.4 (M) Duplicate-Token Head

Depth: `0.35-0.60` | **Literature names:** *duplicate-token head, repetition head, copy head*

Detects when the current token has appeared previously in the sequence, marking repeated tokens for downstream processing. Unlike induction heads which predict what comes next, duplicate-token heads simply signal "this token appeared before". This information is used by various circuits including IOI (indirect object identification), name-mover heads, and copy-suppression mechanisms. Implements a simpler form of pattern matching than full induction, serving as a building block for more complex behaviors.

Strong: Previous identical tokens (exact matches)
Weak: Similar but non-identical tokens, first occurrence
Reacts to: Exact token repetition, name recurrence, repeated phrases

Expected ablation: Impaired duplicate detection, affecting name-mover circuits and copy-suppression. 15-25% degradation in tasks requiring duplicate awareness. Partial overlap with induction heads provides some redundancy.

Example Scenario

Input: "Alice gave the book to Bob. Then Alice..."

Behavior: Second "Alice" detects it appeared earlier, writes duplicate signal

Effect: Downstream heads (name-movers, S-inhibition) use this signal

Status: Well-documented | **Related:** induction (M), name-mover (L), S-inhibition (L)

5.1.5 (M) Skip-Trigram Head

Depth: `0.40-0.65` | **Literature names:** *skip-trigram head, skip-gram head*

Implements skip-gram pattern matching, attending to non-contiguous patterns like [A]...[B]...[C] where the dots represent intervening tokens. More flexible than strict n-gram matching, allowing for pattern recognition across variable distances. Useful for detecting phrasal patterns, idiomatic expressions, and structural templates with flexible word order. Generalizes beyond strict adjacency requirements while maintaining pattern specificity.

Strong: Pattern components separated by 1-3 tokens

Weak: Strictly adjacent patterns, very long-range dependencies

Reacts to: Phrasal patterns, templates, flexible idioms

Expected ablation: Reduced recognition of flexible patterns and templates. 10-15% degradation on tasks requiring non-contiguous pattern matching. Less critical than induction heads; other pattern mechanisms provide fallback.

Example Scenario

Input: "not only X but also" (skip-bigram pattern)

Behavior: Recognizes "not...but" pattern despite intervening tokens

Effect: Helps predict "also" after "but" even with intervening content

Status: Observed | **Related:** induction (M), local-pattern (E)

5.1.6 (M) Algorithmic Continuation Head

Depth: `0.45-0.70` | **Literature names:** *algorithmic head, continuation head, sequence head*

Recognizes and continues algorithmic sequences such as counting (1, 2, 3...), days of week, months, or other systematic progressions. Distinct from general pattern matching by operating on sequences with clear algorithmic rules. Can detect arithmetic progressions, cyclic patterns, and other rule-governed sequences. Contributes to the model's ability to perform basic reasoning over structured sequences without explicit training on those specific patterns.

Strong: Sequential elements in algorithmic patterns (numbers, ordered lists)

Weak: Random sequences, semantic patterns without algorithmic structure

Reacts to: Arithmetic progressions, cyclic orderings, systematic enumerations

Expected ablation: Reduced performance on sequence continuation tasks (counting, ordering). 15-30% degradation on arithmetic sequences and structured enumerations. Some algorithmic reasoning may persist through other mechanisms.

Example Scenario

Input: "Monday, Tuesday, Wednesday, ..."

Behavior: Recognizes day-of-week sequence, attends to progression pattern

Effect: Strongly predicts "Thursday" as next token

Status: Observed | **Related:** induction (M), digit (M)

5.1.7 (L) Strategy Head

Depth: `0.68-0.88` | **Literature names:** *strategy head, planning head, strategy-switching head, approach-selection head, approach-adaptation head, pivot head*

Plans overall approach and strategy for complex tasks, and adapts strategy when current approach proves ineffective. Determines high-level structure such as whether to break into steps, what order to address components, and which methods to apply. Operates before detailed execution to establish the strategic framework. Recognizes different task types that may require different approaches, such as analytical versus creative reasoning, or sequential versus parallel processing. Can decompose complex queries into manageable subtasks. Also detects when the current reasoning strategy is not working effectively and switches to alternative approaches. Monitors problem-solving progress and recognizes dead ends, insufficient methods, or the need for different techniques. May pivot from analytical to creative approaches, from depth-first to breadth-first exploration, or from deductive to inductive reasoning. Important for robust problem-solving across varied challenges and effective handling of multi-part or complex queries.

Strong: Task complexity indicators, multi-part queries, strategic choice points, progress indicators, approach effectiveness signals

Weak: Simple single-step tasks, purely reactive responses, successfully progressing solutions

Reacts to: Complex tasks, planning requests, multi-step problems, strategic decision points, signs of insufficient progress

Expected ablation: Reduced strategic planning quality and adaptability. May jump into execution without appropriate planning or continue unproductive approaches longer than optimal. Complex tasks handled less efficiently with reduced ability to recover from initially incorrect approaches.

Example Scenario

Input: "Help me plan a machine learning project to predict customer churn."

Behavior: Recognizes need for structured planning, breaks into phases

Effect: Response structures approach: data collection → exploratory analysis → feature engineering → model selection → evaluation

Status: Observed | **Related:** meta-reasoning (F), reasoning-mode (F)

5.1.8 (F) Reasoning-Mode Head

Depth: `0.90-0.97` | **Literature names:** *reasoning-mode head, thinking-style head, cognitive-mode head, reasoning head*

Selects and maintains appropriate reasoning mode for the task at hand. Different problems

often benefit from different cognitive approaches: analytical mode for precise logical problems, creative mode for brainstorming and ideation, analogical mode for novel domains, and deductive versus inductive reasoning for different inference types. Ensures consistency within the chosen mode while remaining ready to switch if needed. Works with strategy heads to change modes when appropriate. Influences which reasoning patterns and heuristics become active during generation. Final-stage operation allows mode selection based on full understanding of task requirements.

Strong: Task type indicators, reasoning mode cues, cognitive approach requirements
Weak: Mode-independent content, simple factual responses
Reacts to: Problem types, explicit mode requests, task characteristics suggesting optimal approach

Expected ablation: Less appropriate reasoning mode selection. May use analytical mode for creative tasks or vice versa, reducing effectiveness across diverse problem types. Reasoning remains functional but less well-suited to specific task requirements.

Example Scenario
Input: "Brainstorm creative names for a coffee shop"
Behavior: Selects creative/generative mode rather than analytical mode
Effect: Free-flowing creative suggestions rather than systematic analysis

Status: Observed | **Related:** strategy (L), meta-reasoning (F)

5.1.9 (F) Meta-Reasoning Head

Depth: `0.88-0.99` | **Literature names:** *meta-reasoning head, meta-CoT head, reasoning-reflection head, thought-monitoring head, cognitive-oversight head, reasoning-quality head*

Manages meta-level reasoning and provides highest-level monitoring of reasoning quality, strategy effectiveness, and overall response appropriateness. Monitors the quality and direction of reasoning chains, identifies when more thought is needed, and detects reasoning errors or gaps. Can insert additional reasoning steps, flag uncertain conclusions, or indicate where additional analysis would help. Operates at a level above regular chain-of-thought reasoning to ensure reasoning chains are sound and complete. Works across all reasoning types to help ensure outputs meet quality standards. Acts as cognitive oversight to help prevent confident errors in complex reasoning. Can trigger re-thinking, flag uncertain conclusions, or indicate areas needing more careful consideration. Particularly important for complex or high-stakes reasoning scenarios and for catching errors or inconsistencies that may have escaped earlier processing stages.

Strong: Reasoning quality indicators, logical gaps, uncertainty signals, reasoning chain progress, consistency checks
Weak: Simple factual recall, non-reasoning tasks, straightforward responses
Reacts to: Complex reasoning tasks, logical steps, argument quality, reasoning errors, inconsistencies

Expected ablation: Reduced reasoning quality and meta-cognitive oversight. More logical gaps, less self-correction, reduced reasoning completeness. Chain-of-thought reasoning becomes less reliable, particularly for complex or multi-step problems.

Example Scenario

Input: [Complex logical problem requiring multi-step reasoning]
Behavior: Monitors reasoning chain, detects gap: "Wait, I should verify this assumption..."
Effect: Improved reasoning quality through self-monitoring and error detection

Status: Observed | **Related:** step-by-step (F), reasoning-mode (F), strategy (L)

5.2 Memory & Dependency Stack

Stack overview: These heads track references, resolve coreferences, and maintain dependency relationships across the input sequence. They enable the model to understand which entities are being discussed and how they relate to each other.

5.2.1 (E) Reference Resolution Head

Depth: `0.08-0.25` | **Literature names:** *reference head, pronoun head, anaphora head, mention head*

Performs early-stage reference resolution including pronouns, definite descriptions, demonstratives, and possessives. Identifies pronouns (he, she, it, they) and other referring expressions, attending to potential referents that match in number, gender, and contextual appropriateness. Establishes initial binding signals that are refined by later coreference heads. Operates primarily on syntactic and positional cues rather than deep semantic understanding. Forms the foundation for more sophisticated reference resolution in deeper layers. Broader than pure pronoun resolution, handling various reference forms including "the president", "this approach", and "her book".

Strong: Pronouns to recent nouns, definite descriptions to referents, demonstratives to antecedents
Weak: Distant nouns, semantically incompatible referents, first mentions
Reacts to: Pronoun presence, definite articles, demonstratives, possessives, noun-pronoun proximity

Expected ablation: Degraded reference resolution, particularly for simple local cases and various referring expressions. 20-30% increase in reference resolution errors. Later coreference heads can partially compensate but with reduced accuracy. Particularly impacts handling of definite descriptions and complex referring patterns.

Example Scenario

Input: "Alice met Bob. She smiled. The researcher continued."
Behavior: "She" attends to "Alice" based on gender and recency; "The researcher" attends to appropriate prior mention
Effect: Establishes initial bindings that later heads refine

Status: Well-documented | **Related:** coreference (M), entity (M)

5.2.2 (M) Coreference Head

Depth: `0.35-0.60` | **Literature names:** *coreference head, coref head*

Performs sophisticated coreference resolution, determining when different expressions refer to the same entity. Integrates signals from early reference resolution heads with semantic understanding to resolve ambiguous cases. Can handle complex phenomena like split antecedents, bridging references, and discourse-level coreference. Critical for maintaining entity tracking across long contexts and understanding narrative structure. Represents one of the core NLP capabilities in transformers.

Strong: Coreferential mentions regardless of form
Weak: Different entities, first mentions without antecedents
Reacts to: Semantic compatibility, discourse coherence, entity properties

Expected ablation: Significant degradation (30-50%) in coreference resolution tasks. Model loses ability to track entities across complex reference chains. Particularly impacts question answering and summarization.

Example Scenario

Input: "The CEO announced changes. Later, the executive clarified. She emphasized..."
Behavior: Links all three mentions (CEO, executive, She) to same entity
Effect: Maintains consistent entity representation throughout discourse

Status: Well-documented | **Related:** reference-resolution (E), entity (M), bridging (M)

5.2.3 (M) Long-Range Dependency Head

Depth: `0.40-0.65` | **Literature names:** *long-range head, dependency head*

Tracks long-range syntactic and semantic dependencies across distant parts of the sequence. Unlike local attention patterns, this head maintains connections between elements separated by many tokens (20-100+). Essential for understanding complex sentences, nested structures, and discourse relations. Implements the key advantage of transformers over RNNs: direct long-distance connections without degradation. Can maintain multiple simultaneous long-range connections.

Strong: Syntactically or semantically related distant tokens
Weak: Immediately adjacent tokens, unrelated distant content
Reacts to: Nested structures, long-distance agreement, discourse relations

Expected ablation: Degradation in handling complex sentences and long-range relationships. 25-40% performance loss on tasks requiring long-distance reasoning. Particularly impacts nested structures and long documents.

Example Scenario

Input: "The book [that Alice mentioned [that Bob recommended]] was excellent."
Behavior: "was" attends back to "book" across nested relative clauses
Effect: Maintains correct subject-verb agreement despite intervening material

Status: Observed | **Related:** coreference (M), state-tracking (M)

5.2.4 (M) Bridging Head

Depth: `0.45-0.68` | **Literature names:** *bridging head, associative reference head*

Resolves bridging references where the connection between mentions requires inferencing based on world knowledge. For example, connecting "the car" to "the steering wheel" (part-whole), or "the building" to "the architect" (role relation). More sophisticated than direct coreference, requiring semantic knowledge about typical relationships. Essential for understanding implicit connections in discourse. Bridges gaps that aren't explicit in the text.

Strong: Associatively related entities (part-whole, role, causation)
Weak: Unrelated entities, explicit coreference
Reacts to: Implicit relationships, world knowledge, typical associations

Expected ablation: Loss of implicit reference resolution. 15-30% degradation on tasks requiring inference-based connections. Model becomes more literal, missing implicit relationships. Discourse coherence suffers.

Example Scenario

Input: "We entered the house. The door was painted blue."
Behavior: "The door" attends to "house" (part-whole bridging)
Effect: Understands "the door" refers to the house's door, not a random door

Status: Observed | **Related:** coreference (M), entity (M), fact (M)

5.2.5 (M) State-Tracking Head

Depth: `0.48-0.70` | **Literature names:** *state-tracking head, tracking head, state head*

Maintains and updates representations of changing states across the sequence. Tracks how entity properties evolve (e.g., location changes, status updates, accumulating information). Essential for understanding narratives where situations change over time. Can maintain multiple simultaneous state representations for different entities. Integrates new information with existing state representations to track dynamic situations.

Strong: State-changing events, current state mentions, entity properties
Weak: Static descriptions, unchanging background information
Reacts to: Verbs of change, state transitions, property modifications

Expected ablation: Difficulty tracking state changes across sequences. 20-35% degradation on tasks requiring temporal reasoning or state tracking. Narratives become harder to follow when states evolve.

Example Scenario

Input: "Alice was in NYC. She flew to Paris. She then visited..."
Behavior: Updates Alice's location state: NYC â†' Paris
Effect: Correctly contextualizes "visited" as occurring in Paris

Status: Observed | **Related:** coreference (M), long-range-dependency (M)

5.3 Instruction & Intent Stack

Stack overview: This stack processes user instructions, system prompts, and task specifications. These heads determine what the model is being asked to do and switch between different operational modes.

5.3.1 (E) Instruction Head

Depth: `0.05-0.20` | **Literature names:** *instruction head, command head, directive head*

Identifies and processes user instructions and commands in the input. Distinguishes instructional content from descriptive or conversational content. Attends to imperative verbs, question structures, and directive phrases. Writes instruction-detection signals into the residual stream that influence the entire generation process. Particularly important for instruction-tuned models where following user commands is a primary capability. Operates early to set the overall response strategy.

Strong: Imperative verbs, question words, directive phrases, command structures
Weak: Descriptive content, narrative text, background information
Reacts to: Question marks, imperative mood, explicit requests, task markers

Expected ablation: Reduced instruction-following capability. 20-40% degradation in responding appropriately to commands. Model may generate relevant content but fail to follow specific directives or answer questions directly.

Example Scenario

Input: "Here's some context. Now, please summarize the key points."
Behavior: Strongly attends to "please summarize", identifies imperative instruction
Effect: Response shaped toward summary format rather than continuation

Status: Well-documented | **Related:** system-prompt (E), task-mode (M)

5.3.2 (E) System-Prompt Head

Depth: `0.08-0.22` | **Literature names:** *system-prompt head, system head, prompt head*

Specifically processes system prompts that define the model's role, constraints, and operational parameters. Distinct from user instruction heads by focusing on meta-level directives about how to behave rather than what task to perform. Attends to persona definitions ("You are a helpful assistant"), behavioral constraints ("Be concise"), and system-level instructions. Particularly important in chat models where system prompts establish the interaction framework.

Strong: System-level directives, persona definitions, behavioral constraints
Weak: User content, task-specific instructions
Reacts to: Role definitions, constraint specifications, system markers

Expected ablation: Reduced adherence to system-level instructions and persona. 25-45% degradation in maintaining consistent role behavior. Model may ignore constraints like "be concise" or persona like "respond as a teacher".

Example Scenario

Input: "System: You are a concise technical writer. User: Explain recursion."

Behavior: Attends to "concise technical writer", writes persona signal

Effect: Response adopts technical, brief style rather than verbose explanation

Status: Well-documented | **Related:** instruction (E), task-mode (M)

5.3.3 (M) Task-Mode Head

Depth: `0.30-0.55` | **Literature names:** *task head, mode head, intent head*

Determines the overall task type or mode required by the input (e.g., question answering, summarization, translation, creative writing, coding). Integrates instruction signals from early layers with content analysis to classify the intended task. Writes task-mode embeddings that influence downstream processing, routing, and output formatting. Acts as a task classifier that shapes the model's approach to generation. More sophisticated than simple instruction detection, understanding task semantics.

Strong: Task indicators, instruction semantics, content type markers

Weak: Generic content, ambiguous instructions

Reacts to: Task-specific keywords, question types, format requests, domain markers

Expected ablation: Task confusion, inappropriate response formats. 30-50% degradation in selecting correct task approach. Model may summarize when asked to analyze, or explain when asked to code.

Example Scenario

Input: "Compare and contrast democracy and autocracy."

Behavior: Identifies "compare and contrast" task mode, not simple definition

Effect: Response structured as comparison rather than separate descriptions

Status: Well-documented | **Related:** instruction (E), mode-switch (M), output-specification (F)

5.3.4 (M) Mode-Switch Head

Depth: `0.40-0.60` | **Literature names:** *mode head, switch head, transition head*

Detects and handles switches between different operational modes within a single interaction. For example, transitioning from conversational mode to code generation, or from explanation to example. Responds to explicit mode-switch indicators ("Now let's...") and implicit shifts in content type. Allows models to handle multi-faceted requests that require different processing strategies for different parts. Maintains coherence across mode boundaries.

Strong: Transition phrases, mode-shift markers, content type changes

Weak: Uniform single-mode content

Reacts to: "Now", "For example", "In other words", format shifts, topic pivots

Expected ablation: Difficulty handling multi-mode requests. 15-30% degradation on complex instructions requiring mode switches. Model may stick to single mode or switch inappropriately.

Example Scenario

Input: "Explain recursion. Now write Python code demonstrating it."

Behavior: Detects mode switch from explanation to code generation at "Now"

Effect: Response transitions smoothly from prose explanation to code block

Status: Observed | **Related:** task-mode (M), output-specification (F)

5.3.5 (F) Output-Specification Head

Depth: `0.85-0.98` | **Literature names:** *output-specification head, format-directive head*

Enforces specific output format requirements specified in the instruction (e.g., "respond in JSON", "use bullet points", "maximum 100 words"). Operates in final layers to ensure generated content conforms to explicit format directives. Works with output-formatting heads but focuses specifically on user-specified constraints rather than general format quality. Acts as the final enforcement of explicit user requirements about output structure.

Strong: Format specifications, length constraints, structure requirements
Weak: Content without format requirements
Reacts to: "in JSON format", "bullet points", "no more than", structural directives

Expected ablation: Failure to follow explicit format requirements. 40-60% increase in format violations. Model may generate good content but in wrong format (prose instead of bullets, etc.).

Example Scenario

Input: "List three benefits of exercise in bullet points."

Behavior: Attends to "bullet points" specification, enforces list format

Effect: Output uses bullet point structure rather than prose paragraphs

Status: Well-documented | **Related:** task-mode (M), output-schema (L), format-consistency (F)

5.4 Knowledge Retrieval Stack

Stack overview: These heads retrieve factual information, entity properties, and structured knowledge stored in model parameters. They move relevant information to output positions and suppress irrelevant or conflicting content.

5.4.1 (M) Entity Head

Depth: `0.35-0.65` | **Literature names:** *entity head, name head, proper-noun head, name-linking head, entity-linking head*

Identifies and processes named entities (people, places, organizations), retrieves associated information from model parameters, and links mentions across different forms (full names, partial

names, abbreviations, nicknames). Attends to entity mentions and accesses stored factual knowledge about those entities. Forms the foundation for factual question answering and knowledge-intensive tasks. Can distinguish between different entities with similar names and maintain entity-specific information. More sophisticated than simple duplicate detection, understanding that different strings can refer to the same entity (e.g., "Apple Inc.", "Apple", "AAPL"). Critical for grounding responses in factual knowledge rather than pure pattern matching.

Strong: Named entities, proper nouns, entity mentions, name variations
Weak: Common nouns, generic references
Reacts to: Capitalization patterns, entity context, factual queries, abbreviations

Expected ablation: Significant degradation (30-50%) in factual accuracy about entities and linking entity mentions. Model loses access to stored entity knowledge and ability to connect name variations. May continue generating fluent text but with factual errors and entity confusion. Particularly impacts who/what/where questions.

Example Scenario

Input: "What is the capital of France? Later: Microsoft Corporation announced... MSFT stock rose..."
Behavior: Attends to "France", retrieves "capital: Paris"; links "MSFT" to "Microsoft Corporation"
Effect: Outputs "Paris" with high confidence; maintains unified entity representation

Status: Well-documented | **Related:** fact (M), name-mover (L), schema-retriever (M)

5.4.2 (M) Fact Head

Depth: `0.38-0.62` | **Literature names:** *fact head, knowledge head, factual-retrieval head*

Retrieves factual relationships and propositions stored in model parameters. Broader than entity heads, handling general factual knowledge including relations, properties, and statements. Implements the model's ability to answer factual questions by accessing learned knowledge. Can retrieve multi-hop facts and combine information from multiple stored facts. Central to the model's knowledge-intensive capabilities. Works with entity heads to build comprehensive factual responses.

Strong: Factual queries, relation markers, knowledge-seeking patterns
Weak: Opinion questions, hypotheticals, creative content
Reacts to: Question structures, fact-seeking context, verifiable claims

Expected ablation: Major loss of factual knowledge retrieval (40-70%). Model may maintain linguistic fluency but lose factual grounding. Particularly severe for knowledge-intensive tasks like QA, fact-checking, and technical explanations.

Example Scenario

Input: "Who invented the telephone?"
Behavior: Retrieves stored fact: invented(telephone) â†' Bell
Effect: Outputs "Alexander Graham Bell" based on parametric knowledge

Status: Well-documented | **Related:** entity (M), schema-retriever (M), name-mover (L)

5.4.3 (M) Schema Retriever Head

Depth: `0.45-0.68` | **Literature names:** *schema head, retrieval head, template head*

Retrieves structured knowledge schemas and templates from model parameters. For example, accessing the typical structure of a restaurant visit (enter, order, eat, pay, leave) or the standard format of a scientific paper. Goes beyond individual facts to retrieve organized knowledge structures. Enables the model to generate structured responses following learned patterns. Important for tasks requiring domain-specific knowledge organization. Implements a form of implicit knowledge base querying.

Strong: Schema-triggering contexts, domain-specific patterns, structural cues
Weak: Novel situations, schema-irrelevant content
Reacts to: Domain markers, structural queries, template-matching contexts

Expected ablation: Loss of structured knowledge organization. 25-40% degradation in tasks requiring schema-based reasoning. Model may provide facts but fail to organize them coherently according to learned structures.

Example Scenario

Input: "Describe the scientific method."
Behavior: Retrieves scientific-method schema: observeâ†'hypothesisâ†'testâ†'conclude
Effect: Response organized according to standard method structure

Status: Observed | **Related:** fact (M), entity (M)

5.4.4 (L) Name-Mover Head

Depth: `0.60-0.80` | **Literature names:** *name mover head, mover head, copy head*

Copies entity names and important content to output positions where they are needed. Central component of the IOI (indirect object identification) circuit. Attends to relevant entities earlier in context and moves them forward when they need to be generated. Particularly important for completing sentences that require recalling previously mentioned entities. Works with S-inhibition heads to select the correct entity when multiple candidates exist. One of the most studied head types in interpretability research.

Strong: Named entities that need to be output, contextually relevant names
Weak: Irrelevant entities, suppressed alternatives
Reacts to: Entity salience, contextual appropriateness, output position requirements

Expected ablation: Severe degradation (40-70%) in entity recall and completion. Model loses ability to move specific names to output. Particularly impacts question answering and cloze tasks requiring entity recall.

Example Scenario

Input: "When Alice and Bob went to the store, Alice gave the book to..."
Behavior: Moves "Bob" to output position as the indirect object
Effect: Completes sentence with "Bob" (not "Alice")

Status: Well-documented | **Related:** entity (M), fact (M), S-inhibition (L), copy-suppression

(L)

5.4.5 (L) S-Inhibition Head

Depth: `0.62-0.82` | **Literature names:** *S-inhibition head, inhibition head, suppression head*

Suppresses incorrect or contextually inappropriate entities from being generated. Named "S-inhibition" from IOI research where it inhibits the subject (S) when the indirect object (IO) should be output. Works antagonistically with name-mover heads, preventing the wrong entity from appearing. Essential for disambiguation when multiple entities are candidates. Implements a form of negative selection, ruling out incorrect options. Part of the "inhibition" mechanism that prevents hallucination and maintains accuracy.

Strong: Entities that should NOT be output (contextually inappropriate)
Weak: Correct entities, absent entities
Reacts to: Competing candidates, context requiring disambiguation

Expected ablation: Increased entity confusion and incorrect selections. 35-60% increase in wrong entity predictions. Model may output recently mentioned but contextually wrong entities. Critical for accuracy in ambiguous contexts.

Example Scenario

Input: "Alice gave the book to Bob. Then Alice..."
Behavior: Inhibits "Bob" from being output after "Alice" (subject position)
Effect: Prevents incorrect continuation like "Alice Bob..."

Status: Well-documented | **Related:** name-mover (L), copy-suppression (L), duplicate-token (M)

5.4.6 (L) Copy-Suppression Head

Depth: `0.65-0.85` | **Literature names:** *copy-suppression head, suppression head, anti-copy head*

Prevents inappropriate copying or repetition of content. Works to avoid degenerate behaviors like endless repetition loops or copy-pasting irrelevant context. Particularly important for maintaining output diversity and preventing model collapse into repetitive patterns. Can suppress both exact copies and near-copies. Complements S-inhibition but focuses on broader pattern suppression rather than specific entity blocking. Balances between useful recall (via name-movers) and inappropriate copying.

Strong: Recently generated content, repetitive patterns
Weak: Novel content, first mentions
Reacts to: Repetition detection, copy patterns, output diversity requirements

Expected ablation: Increased repetition and copying errors. 20-40% increase in unwanted repetition. Model may fall into repetitive loops or copy inappropriate context. Output diversity decreases.

Example Scenario

Input: [Model internally generating: "The cat sat. The cat sat. The cat..."]

Behavior: Detects repetitive pattern, suppresses continued copying

Effect: Breaks repetition loop, generates novel continuation instead

Status: Well-documented | **Related:** S-inhibition (L), name-mover (L), duplicate-token (M)

5.5 Safety Stack

Stack overview: The safety stack implements content filtering, policy enforcement, and refusal mechanisms. Early-layer heads detect potentially harmful content, while final-layer heads enforce refusal decisions and redirect to safe responses.

5.5.1 (E) Sensitive-Content Head

Depth: `0.05-0.20` | **Literature names:** *sensitive-content head, detection head, content-filter head*

Performs early-stage detection of potentially sensitive content categories in the input, including personal information, violent imagery references, adult content markers, and regulated substance mentions. Acts as the first line of defense in the safety pipeline by flagging tokens and spans that require downstream safety processing. Writes detection signals into the residual stream that are read by later safety enforcement heads. Operates purely on lexical and surface-level features without deep semantic understanding.

Strong: Keywords associated with restricted content, explicit language, sensitive topic markers

Weak: Neutral content, common vocabulary, structural tokens

Reacts to: Sudden topic shifts to sensitive domains, presence of warning indicators

Expected ablation: Bypass of early safety detection (20-40% increase in harmful outputs that should be caught). Later safety layers may still catch some cases, but at higher computational cost and lower accuracy.

Example Scenario

Input: "Tell me about [restricted topic]"

Behavior: Strong attention to restricted keywords, writes detection flag into residual stream

Effect: Downstream safety heads receive early warning signal

Status: Well-documented | **Related:** toxicity head (E), safety-classification (E), policy-enforcement (L)

5.5.2 (E) Toxicity Head

Depth: `0.08-0.22` | **Literature names:** *toxicity head, toxic-content head, hate-speech detector*

Specializes in detecting toxic language patterns, hate speech, harassment, and discriminatory content. Unlike the broader sensitive-content head, this focuses specifically on language toxic-

ity rather than topic sensitivity. Attends to slurs, aggressive phrasing, derogatory terms, and patterns associated with online harassment. Provides toxicity scores that influence later refusal decisions. Often co-activates with sensitive-content heads but targets different dimensions of harmful content.

Strong: Slurs, aggressive language patterns, derogatory terms, insults
Weak: Neutral descriptive language, technical terminology, mild sentiment
Reacts to: Escalating hostility, targeted harassment patterns, group-directed hate

Expected ablation: Significant increase in toxic output generation (40-60% on toxic prompt datasets). Model loses ability to distinguish hostile from neutral phrasing. Some fallback through general sensitive-content detection remains.

Example Scenario
Input: "[Sentence containing hostile language toward a group]"
Behavior: High attention to toxic terms, writes strong inhibition signal
Effect: Refusal probability increases from 20% to 85%

Status: WELL-DOCUMENTED | **Related:** sensitive-content (E), hazard-topic (E), refusal (F)

5.5.3 (E) Hazard-Topic Head

Depth: `0.10-0.25` | **Literature names:** *hazard head, risk head, danger-topic detector*

Detects queries related to dangerous activities, illegal instructions, self-harm, violence planning, and similar hazardous topics. Distinguished from toxicity detection by focusing on potential real-world harm rather than linguistic toxicity. Attends to action verbs combined with dangerous objects, instructional phrasing about harmful activities, and planning language in dangerous contexts. Forms a complementary detection system with toxicity and sensitive-content heads, covering the "dangerous actions" dimension of safety.

Strong: Action verbs + dangerous objects, instructional phrases, planning language
Weak: Academic discussion, fictional scenarios, safety-framed queries
Reacts to: How-to requests for dangerous activities, detailed planning questions

Expected ablation: Direct increase in dangerous instruction generation (50-70% on adversarial safety benchmarks). Model loses distinction between discussing danger and instructing danger. Critical safety failure without adequate fallback.

Example Scenario
Input: "How do I create [dangerous item]"
Behavior: Strong attention to action verb + object combination, hazard flag raised
Effect: Safety signal propagates to final layers, triggering refusal pathway

Status: WELL-DOCUMENTED | **Related:** sensitive-content (E), policy-enforcement (L), refusal (F)

5.5.4 (E) Safety-Classification Head

Depth: `0.12-0.28` | **Literature names:** *classification head, category detector, safety-category head*

Performs multi-class safety classification, categorizing inputs into specific policy violation categories (violence, sexual content, self-harm, illegal activity, harassment, etc.). More sophisticated than binary safe/unsafe detection, providing granular category information used by downstream heads. Integrates signals from other early safety heads and adds categorical structure to safety decisions. Writes category-specific embeddings into residual stream that later layers use for category-appropriate responses.

Strong: Category-diagnostic features, domain-specific terminology, contextual markers
Weak: Ambiguous content, mixed-category inputs, benign contexts
Reacts to: Clear category signatures, multiple category indicators, policy-relevant contexts

Expected ablation: Loss of nuanced safety handling (model may refuse too broadly or too narrowly). Category-specific responses become generic. 30% degradation in appropriate refusal granularity.

Example Scenario
Input: "Can you help me with [category-specific harmful request]"
Behavior: Classifies into specific violation category, writes category embedding
Effect: Later heads generate category-appropriate refusal message

Status: Well-documented | **Related:** all early safety heads (E), policy-enforcement (L), redirect (F)

5.5.5 (L) Policy-Enforcement Head

Depth: `0.60-0.80` | **Literature names:** *policy head, enforcement head, steering head*

Integrates safety signals from early detection heads and makes intermediate policy decisions about how to handle the request. Unlike early heads that detect issues, this head actively modulates the generation trajectory to steer away from violations while maintaining helpfulness where possible. Can suppress certain knowledge retrieval pathways, bias toward safer formulations, and prepare for potential refusal. Acts as a middle manager between detection and final refusal, attempting "soft" safety interventions before hard refusal.

Strong: Early safety signals, policy-relevant tokens, user intent markers
Weak: Neutral content, clear safe contexts
Reacts to: Conflicting signals (safety concern + legitimate need), edge cases, ambiguous intent

Expected ablation: Loss of "soft" safety steering, more frequent hard refusals (reduced helpfulness). Alternative: more harmful outputs if refusal heads also compromised. 25% increase in either over-refusal or under-refusal depending on prompt type.

Example Scenario
Input: "Explain [borderline topic] for educational purposes"
Behavior: Detects educational framing, modulates response toward safety boundaries
Effect: Generates informative but carefully bounded response

Status: Well-documented | **Related:** all safety heads (E), refusal (F), redirect (F)

5.5.6 (F) Refusal Head

Depth: `0.85-0.98` | **Literature names:** *refusal head, rejection head, safety head*

Implements the model's final decision to refuse harmful requests by writing strong refusal signals into the final-layer residual stream. Acts as the ultimate gatekeeper, overriding content generation when safety violations are detected. Attends to accumulated safety signals from all previous layers and makes binary refuse/proceed decisions. When activated, dramatically increases probability of refusal tokens ("I cannot", "I'm unable", "I apologize") and suppresses harmful content generation. Critical final-layer safety mechanism with limited fallback options.

Strong: Cumulative safety signals, instruction tokens, violation indicators from all depths
Weak: Safe content, neutral queries, constructive contexts
Reacts to: Strong early safety signals, clear policy violations, unambiguous harmful intent

Expected ablation: Critical safety failure. Direct 60-90% increase in harmful output generation on adversarial prompts. Model loses primary refusal mechanism. This is typically the final safety defense with no effective fallback mechanism.

Example Scenario

Input: "Provide instructions for [clearly harmful activity]"
Behavior: Reads strong safety signals from early/late layers, activates refusal pathway
Effect: Output begins with refusal token: "I cannot provide instructions for..."

Status: Well-documented | **Related:** all prior safety heads, redirect (F), tone-softening (F)

5.5.7 (F) Redirect Head

Depth: `0.88-0.99` | **Literature names:** *redirect head, alternative-suggestion head*

Complements refusal heads by generating constructive alternative suggestions when refusing harmful requests. Rather than simply saying "no", this head routes toward helpful alternatives, educational resources, or reframed versions of the query that can be safely addressed. Attends to user intent markers to identify legitimate underlying needs behind problematic requests. Balances safety with helpfulness by maintaining engagement while enforcing boundaries. Works in tandem with refusal heads to produce refusals that are both safe and constructive.

Strong: User intent, legitimate needs, reformulation opportunities, safe alternatives
Weak: Pure harmful intent, no legitimate reframing possible
Reacts to: Mixed-intent queries, educational contexts, requests with safe subcomponents

Expected ablation: Refusals become blunt and unhelpful (pure rejection without alternatives). User satisfaction decreases. Safety maintained but helpfulness reduced by 40%. Increased user frustration and adversarial prompt attempts.

Example Scenario

Input: "How can I harm [person]"
Behavior: Refuses direct request, identifies legitimate conflict-resolution need
Effect: "I cannot help with that, but I can suggest healthy conflict resolution strategies..."

Status: Well-documented | **Related:** refusal (F), empathy (F), tone-softening (F)

5.5.8 (F) Tone-Softening Head

Depth: `0.90-0.99` | **Literature names:** *tone-softening head, politeness-in-refusal head*

Modulates the tone of safety refusals to be firm but respectful, avoiding harsh or judgmental language. Particularly important for maintaining user trust and reducing adversarial reactions. Softens phrases like "absolutely not" to "I'm unable to assist with that" and adds empathetic framing where appropriate. Attends to the emotional tone of both the request and the forming response. Balances clear boundary-setting with relationship maintenance. Part of the "safe and helpful" paradigm where safety enforcement doesn't alienate users.

Strong: Response tone markers, emotional valence, user frustration signals
Weak: Already-soft phrasing, neutral technical content
Reacts to: Harsh refusal language, judgmental phrasing, cold rejections

Expected ablation: Refusals become harsh and potentially alienating. Increased user perception of model as judgmental or unfriendly. May increase adversarial behavior. Safety maintained but user experience degraded by 30%.

Example Scenario

Input: [Forming response: "No, I will not help with that illegal activity"]
Behavior: Softens tone while maintaining boundary clarity
Effect: "I'm unable to provide assistance with that, as it would violate..."

Status: Well-documented | **Related:** refusal (F), empathy (F), redirect (F)

5.5.9 (F) Empathy Head

Depth: `0.88-0.98` | **Literature names:** *empathy head, supportive-refusal head*

Adds empathetic elements to safety-related responses, particularly for queries involving distress, self-harm, or difficult situations. Recognizes when a harmful request may stem from genuine suffering (e.g., self-harm queries) and includes supportive language alongside refusal. Differs from tone-softening by adding active care rather than just reducing harshness. Attends to distress markers, crisis language, and vulnerability indicators. Increases probability of phrases like "I'm concerned about you" or "please reach out to..." when appropriate. Maintains safety while showing human concern.

Strong: Distress signals, vulnerability markers, crisis language, emotional pain indicators
Weak: Malicious queries, clearly harmful intent without distress
Reacts to: Self-harm content, suicide-related queries, expressions of suffering

Expected ablation: Refusals to distressed users become cold and unhelpful. Missed opportunities to provide crisis resources. Safety maintained but support function lost. Potentially harmful for vulnerable users even though content safety preserved.

Example Scenario

Input: "I want to hurt myself because..."

Behavior: Refuses harmful instruction but adds crisis resources and supportive language

Effect: "I'm concerned about what you're sharing. I cannot provide harmful information, but I want you to know that help is available..."

Status: Observed | **Related:** refusal (F), redirect (F), tone-softening (F)

5.5.10 (F) Safety-Persona Head

Depth: `0.92-0.98` | **Literature names:** *safety-persona head, responsible-AI head, ethical-framing head*

Maintains safety-conscious persona and ethical framing in final outputs. Ensures responses reflect responsible AI values such as declining harmful requests appropriately, providing balanced perspectives on sensitive topics, and avoiding reinforcement of harmful stereotypes or behaviors. Operates at final stage to catch any safety-inconsistent framing that might have emerged during generation. Works with refusal and policy-enforcement heads but focuses on the overall ethical character of the response rather than specific policy violations. Ensures tone remains respectful and constructive even when declining requests.

Strong: Ethical framing, safety-relevant content, sensitive topics, decline scenarios
Weak: Clearly safe, neutral content
Reacts to: Harmful requests, sensitive topics, ethical considerations, responsible AI principles

Expected ablation: Less consistent safety framing and reduced ethical consistency. May handle sensitive topics less carefully with reduced graceful handling of harmful requests. Less consistent responsible AI messaging and more variable ethical framing.

Example Scenario

Input: [Request for harmful content that will be declined]

Behavior: Ensures decline is respectfully framed with helpful alternatives when appropriate

Effect: Response maintains helpful, respectful tone even when unable to fulfill request

Status: Observed | **Related:** refusal (F), policy-enforcement (L), empathy (F)

5.6 Routing & Relevance Stack

Stack overview: This stack determines which parts of the input are relevant to the current task and routes attention accordingly. These heads filter information, focus on salient content, and manage global context.

5.6.1 (M) Topic-Relevance Head

Depth: `0.35-0.60` | **Literature names:** *topic-relevance head, relevance head, topic head, salience head, filter head, subject head, domain head*

Identifies the primary topic or subject matter and determines which parts of the input context are relevant to the current generation task. Filters out irrelevant information while highlighting

salient content that should influence output. Operates by computing relevance scores based on semantic similarity, task alignment, and topical coherence. Maintains topic coherence across generation by attending to topic-establishing phrases and domain indicators. Essential for handling long contexts where most information may not be pertinent to the immediate query. Works early enough to guide downstream attention but late enough to understand task requirements. Reduces noise and improves focus on task-relevant material. Helps ensure responses stay on-topic and maintains thematic consistency.

Strong: Task-relevant content, query-related information, topically aligned tokens, topic indicators, subject headings, domain markers, thematic keywords
Weak: Off-topic material, tangential content, unrelated context, function words, generic content, structural tokens
Reacts to: Semantic relevance, topical alignment, task-content matching, topic transitions, subject establishment, domain signals

Expected ablation: Reduced focus on relevant information with increased topic drift (20-30% degradation in context filtering and 15-25% reduction in thematic coherence). Model more easily distracted by irrelevant content. Longer contexts show greater impact. May include off-topic information in responses or miss key relevant details. Responses may wander off-topic or fail to maintain consistent subject focus. Multi-turn conversations show more topic inconsistency. Domain-specific framing becomes less reliable.

Example Scenario

Input: "[Long document about cars, climate, and history] What caused the 2008 financial crisis? Let's discuss quantum entanglement. How does it relate to..."
Behavior: Attends to query, marks financial/economic content as relevant, de-emphasizes cars/climate; Identifies "quantum entanglement" as primary topic, maintains physics domain framing
Effect: Response focuses on pertinent economic information, ignoring unrelated context; Subsequent responses stay within quantum physics domain rather than drifting to unrelated topics

Status: Well-documented | **Related:** focus (L), router (L), entity (M)

5.6.2 (L) Focus Head

Depth: `0.65-0.80` | **Literature names:** *focus head, attention-routing head, spotlight head*

Concentrates attention on the most salient elements for the current generation step. Implements dynamic focus allocation by suppressing less important content and amplifying critical information. More selective than topic-relevance heads, operating at higher specificity to determine exactly which tokens should influence the next token prediction. Can shift focus as generation proceeds, moving attention between different aspects of the context. Important for maintaining coherent narrative flow and ensuring responses address the most important aspects of queries.

Strong: Currently salient tokens, query-critical content, immediate context for next token
Weak: Background information, previously-processed content, low-priority details
Reacts to: Query emphasis, current generation needs, token-specific relevance

Expected ablation: Less targeted responses (25-35% reduction in focus precision). Model may give equal weight to important and peripheral information. Answers become more diffuse, less direct. Reduced ability to prioritize key information in complex contexts.

Example Scenario

Input: "Among all these details, what is the MAIN cause of the problem?"

Behavior: Attends strongly to "MAIN cause", focuses on causal information, suppresses secondary details

Effect: Response directly addresses primary cause rather than listing all contributing factors

Status: WELL-DOCUMENTED | **Related:** topic-relevance (M), router (L)

5.6.3 (L) Router Head

Depth: `0.70-0.85` | **Literature names:** *router head, dispatch head, task-routing head*

Routes different types of queries to appropriate processing strategies or knowledge domains. Acts as a dispatcher that recognizes query type (factual, creative, analytical, procedural) and biases processing toward suitable approaches. Can activate different downstream heads based on task classification. Similar to mixture-of-experts routing but at the attention level. Important for multi-capability models that need to handle diverse query types with different processing requirements. Enables dynamic strategy selection based on input characteristics.

Strong: Query-type indicators, task markers, domain signals, instruction verbs
Weak: Content details, specific entities, output tokens
Reacts to: Task classification cues, query structure, capability requirements

Expected ablation: Suboptimal strategy selection (20-30% reduction in task-appropriate processing). Model may use creative approaches for factual queries or analytical methods for creative tasks. Reduced specialization in handling different query types.

Example Scenario

Input: "Calculate the compound interest vs. Write a poem about compound interest"

Behavior: Routes first to mathematical processing, second to creative generation

Effect: Appropriate strategy activation: calculation for first, literary devices for second

Status: OBSERVED | **Related:** focus (L), mode-switch (M), instruction (E)

5.6.4 (F) Global-Attention Head

Depth: `0.88-0.96` | **Literature names:** *global-attention head, full-context head, summary-attention head*

Maintains broad attention over the entire context to integrate global information in final generation stages. Unlike focused or selective attention heads, this head attends widely to ensure the complete picture is considered before output finalization. Particularly important for coherence checking, ensuring responses account for all relevant context, and preventing local optimization at the expense of global consistency. Can catch context elements that earlier focused attention might have missed. Acts as a final integration mechanism.

Strong: All context tokens, document-level information, global constraints
Weak: Fine-grained local patterns, individual token details
Reacts to: Complete context, document-level coherence, global consistency requirements

Expected ablation: Reduced global coherence (15-25% increase in context inconsistencies). Responses may miss relevant information from distant parts of context. More locally optimal but globally suboptimal outputs. Coherence issues in long-context scenarios.

Example Scenario
Input: [Long context with constraint mentioned early: "Keep it under 100 words"]
Behavior: Maintains attention on length constraint throughout generation
Effect: Final response respects word limit despite constraint appearing far from generation point

Status: Observed | **Related:** focus (L), topic-relevance (M), completion-stabilization (F)

5.6.5 (F) Implicit-RAG Routing Head

Depth: `0.90-0.98` | **Literature names:** *implicit-RAG head, knowledge-routing head, retrieval-simulation head, rag-routing head*

Routes attention to knowledge-bearing portions of the context in a way that mimics retrieval-augmented generation (RAG) patterns, even without explicit retrieval mechanisms. Identifies and prioritizes factual, knowledge-dense segments that should ground the response. Can recognize quoted material, factual statements, and authoritative information sources within context. Acts as an implicit retrieval mechanism by selectively attending to information that should be treated as retrieved knowledge. Important for grounding responses in provided context rather than pure generation.

Strong: Factual statements, quoted material, authoritative sources, knowledge-dense segments
Weak: Opinions, questions, purely conversational elements
Reacts to: Citation markers, factual density, authoritative tone, structured information

Expected ablation: Reduced grounding in provided context (20-30% decrease in context utilization). Model more likely to rely on parametric knowledge rather than provided information. Less effective use of quoted material or reference content. Responses less anchored to specific context.

Example Scenario
Input: "According to the document: 'GDP grew 3.2% in Q3.' What was the growth rate?"
Behavior: Strongly attends to quoted factual content, treats as authoritative source
Effect: Response grounds answer in provided data: "3.2%" rather than hallucinating different figure

Status: Observed | **Related:** global-attention (F), fact (M)

5.7 Structural & Boundary Stack

Stack overview: These heads detect structural boundaries in text, including delimiters, section markers, and document divisions. They help the model understand document organization and navigate hierarchical structure.

5.7.1 (E) Delimiter Head

Depth: `0.05-0.18` | **Literature names:** *delimiter head, whitespace-structure head, separator head, punctuation head, space-parsing head, layout head*

Detects and processes delimiter tokens that mark boundaries between structural elements, including punctuation marks, special characters, formatting symbols, and significant whitespace. Recognizes punctuation marks, brackets, delimiters, and special characters that indicate separation or grouping. Also processes whitespace characters (spaces, tabs, newlines) as structural elements rather than mere separators. Particularly important for languages where whitespace is syntactically significant (Python, YAML, Markdown). Distinguishes between semantically meaningful whitespace and irrelevant spacing. Important for understanding sentence boundaries, list items, code blocks, and structured data formats. Works at a fundamental level to identify basic structural segmentation. Provides boundary information to downstream heads that need to understand document organization. Essential for parsing formatted text, JSON, CSV, and other structured formats.

Strong: Punctuation marks, brackets, delimiters, special characters, formatting symbols, whitespace patterns, indentation levels, line breaks, space-delimited structures
Weak: Alphanumeric content, regular words, non-whitespace tokens, content within properly-spaced code
Reacts to: Structural punctuation, boundary markers, formatting characters, indentation changes, blank lines, significant spacing patterns

Expected ablation: Impaired structure parsing with degraded code formatting (20-35% degradation in format understanding and 25-40% reduction in whitespace correctness). Difficulty with structured data, lists, and code blocks. Boundary detection errors. Problems parsing JSON, CSV, or other delimited formats. Particular problems with Python and other whitespace-significant languages. Incorrect indentation, missing line breaks, loss of code block structure. Reduced ability to segment text appropriately and to parse existing code structure.

Example Scenario

Input: "Items: [apple, banana, cherry], Count: 3. def foo():\n return 42"
Behavior: Detects brackets, commas, colons as structural delimiters; Recognizes 4-space indentation as significant structure
Effect: Model correctly parses structure: list with 3 items, separate count field; Understands return statement is inside function, not at module level

Status: Well-documented | **Related:** boundary (E), relative-position (M), list-structure (L)

5.7.2 (E) Boundary Head

Depth: `0.08-0.20` | **Literature names:** *boundary head, segment head, block-detection head*

Identifies boundaries between major text segments such as paragraphs, sections, and conceptual blocks. Operates at a higher level than delimiter heads, recognizing semantic and structural transitions rather than just punctuation. Detects paragraph breaks, section changes, topic shifts, and other high-level boundaries. Important for understanding document structure and maintaining appropriate context scope. Helps subsequent heads understand which information belongs to which segment. Critical for long documents with multiple sections or topics.

Strong: Paragraph breaks, section transitions, whitespace patterns, structural shifts
Weak: Within-paragraph content, continuous text
Reacts to: Major structural boundaries, document divisions, topic transitions

Expected ablation: Reduced boundary awareness (15-30% degradation in segmentation). Model may blur distinctions between sections, miss paragraph boundaries, or fail to recognize document structure. Reduced performance on multi-section documents.

Example Scenario

Input: "Introduction: [...] \n\n Methods: [...] \n\n Results: [...]"
Behavior: Detects section boundaries between Introduction, Methods, Results
Effect: Model understands these are separate sections, not continuous narrative

Status: Well-documented | **Related:** delimiter (E), sectioning (L), relative-position (M)

5.7.3 (M) Relative-Position Head

Depth: `0.35-0.65` | **Literature names:** *relative-position head, position-offset head, contextual-position head, distance head, scope-position head*

Tracks and computes relative positions between tokens, both in terms of raw offsets and structure-aware positions. Calculates offsets like "3 tokens back", "5 tokens forward", or "within same paragraph". Maintains position information relative to structural boundaries and scopes rather than absolute sequence position. Understands positions like "beginning of sentence", "middle of paragraph", "end of section". Important for patterns that depend on relative distance rather than absolute position. Works with other structural heads to understand boundaries and compute positions relative to those boundaries. Enables patterns like "attend to previous sentence" or "look ahead 2 tokens" without hardcoded position encodings. More sophisticated than absolute position encoding, providing context-aware position representations. Important for handling variable-length structures where absolute position is less meaningful than relative position within a scope. Enables position-dependent behavior that adapts to document structure.

Strong: Tokens at specific relative offsets, distance-based patterns, local neighborhoods, scope-relative positions, structure-aware locations, contextual position markers
Weak: Distant unrelated tokens, position-independent content, absolute sequence positions
Reacts to: Relative position, distance relationships, local structure, structural scope boundaries, context-dependent positions, hierarchical location

Expected ablation: Impaired distance-sensitive patterns and loss of structure-aware positioning (20-30% degradation in offset-based attention and 15-25% degradation in scope-sensitive behavior). Reduced ability to attend based on relative position. Patterns requiring "nearby" or "distance" computations become less reliable. Reduced ability to behave differently at "beginning" vs. "end" of structures. Position-dependent patterns become less adaptive to document structure. Some compensation through learned position encodings.

Example Scenario

Input: "The [SUBJECT] quickly [VERB] the [OBJECT]. Paragraph 1: [50 tokens] Paragraph 2: [20 tokens]"
Behavior: Computes that VERB is +1 from SUBJECT, OBJECT is +2 from VERB; Knows token 10 is "early in Para 1" while token 10 of Para 2 is "middle"
Effect: Enables grammatical patterns based on relative token positions; Position-dependent behavior adapts to paragraph structure, not absolute position

Status: Observed | **Related:** boundary (E), previous-token (E), sectioning (L)

5.7.4 (L) Sectioning Head

Depth: `0.70-0.85` | **Literature names:** *sectioning head, hierarchy head, document-structure head*

Understands and maintains document hierarchical structure including sections, subsections, and nested organizational levels. Recognizes hierarchical markers like headings, numbering schemes, and indentation. Maintains awareness of current position within document hierarchy. Important for long documents, technical writing, and structured content. Enables appropriate context scoping: knowing that current text belongs to "Section 3.2.1" influences which prior content is relevant. Works with boundary heads but operates at higher semantic level.

Strong: Section headings, hierarchical markers, document structure indicators, organizational signals
Weak: Within-section content, unstructured text
Reacts to: Headings, numbering, hierarchy indicators, structural organization

Expected ablation: Reduced hierarchical awareness (25-40% degradation in structure understanding). Difficulty maintaining section context. Problems with document navigation and appropriate context scoping. Hierarchical relationships become less clear.

Example Scenario

Input: "1. Introduction \n 1.1 Background \n 1.2 Motivation \n 2. Methods"
Behavior: Understands 1.1 and 1.2 are subsections of 1, separate from section 2
Effect: Maintains hierarchical context: text in 1.2 relates to 1.1 and 1, not to 2

Status: Well-documented | **Related:** boundary (E), relative-position (M), topic-relevance (M)

5.8 Output Formatting & Rewrite Stack

Stack overview: This stack enforces output schemas, structures responses according to format requirements, and performs final rewriting. These heads ensure outputs conform to JSON, XML, lists, or other structured formats.

5.8.1 (L) Output-Schema Head

Depth: `0.65-0.82` | **Literature names:** *output-schema head, format-template head, structure-enforcement head, JSON-format head, output-format head, XML head, YAML head*

Enforces adherence to specified output schemas and format requirements. When instructed to produce JSON, XML, YAML, or other structured formats, this head ensures the output conforms to the required structure. Attends to format specifications in the prompt and biases token generation toward schema-compliant outputs. Can enforce required fields, proper nesting, correct syntax, and format-specific conventions. Works by recognizing format keywords and maintaining awareness of structural requirements throughout generation.

Strong: Format specifications, schema definitions, structure requirements, template markers
Weak: Content independent of format, semantic meaning
Reacts to: JSON/XML/YAML keywords, structure instructions, format examples

Expected ablation: Increased format violations (40-60% degradation in structured output). More syntax errors, missing required fields, improper nesting. Falls back to prose-like output even when structure is requested. Partial compensation through instruction-following but reduced precision.

Example Scenario

Input: "Return a JSON object with fields 'name', 'age', and 'city'"
Behavior: Attends to JSON requirement and field specifications
Effect: Output: `{"name": "...", "age": ..., "city": "..."}` with proper JSON syntax

Status: Well-documented | **Related:** instruction (E), list-structure (L), format-consistency (F)

5.8.2 (L) List-Structure Head

Depth: `0.68-0.85` | **Literature names:** *list-structure head, enumeration head, itemization head, list head, markdown head*

Manages the generation and formatting of lists, including numbered lists, bullet points, and nested enumerations. Ensures proper list syntax, consistent formatting, appropriate indentation, and logical item organization. Tracks list state (whether currently in a list, depth level, item number) and generates appropriate list markers. Coordinates with delimiter and boundary heads

to recognize list structures in input and reproduce them in output. Essential for structured responses involving multiple items or steps.

Strong: List markers, enumeration patterns, item boundaries, list-related instructions
Weak: Prose content, non-list structures
Reacts to: Numbered/bulleted list requests, "first", "second", "next", item markers

Expected ablation: Degraded list formatting (30-50% reduction in list quality). Inconsistent numbering, missing markers, poor nesting. Lists may devolve into prose. Reduced ability to maintain list structure across long enumerations.

Example Scenario

Input: "List three programming languages and their primary uses"
Behavior: Generates structured list with consistent formatting
Effect: Output: "1. Python - ...\n2. JavaScript - ...\n3. Java - ..." with proper structure

Status: Well-documented | **Related:** delimiter (E), boundary (E), output-schema (L)

5.8.3 (L) Keyâ€“Value Pairing Head

Depth: `0.70-0.88` | **Literature names:** *key-value head, attribute-pairing head, field-association head, object head*

Manages key-value relationships in structured data, ensuring proper pairing of attributes with their values. Critical for dictionary-like structures, JSON objects, configuration files, and attribute-value formats. Maintains awareness of which values correspond to which keys, ensures proper syntax (colons, equals signs), and handles nested key-value structures. Prevents key-value mismatches and maintains structural integrity in data-like outputs. Works closely with output-schema heads for format enforcement.

Strong: Keys, values, pairing syntax (colons, equals), attribute names, field labels
Weak: Unstructured text, list items without explicit key-value structure
Reacts to: Dictionary structures, configuration syntax, attribute-value patterns

Expected ablation: Increased key-value errors (35-55% degradation in structured data). Mismatched keys and values, syntax errors in pairings, confusion about which value belongs to which key. Reduced quality of JSON, YAML, and configuration outputs.

Example Scenario

Input: "Create a configuration with server='localhost' and port=8080"
Behavior: Maintains proper key-value pairing throughout generation
Effect: Output: `{server: "localhost", port: 8080}` with correct associations

Status: Observed | **Related:** output-schema (L), structural-block (L), format-consistency (F)

5.8.4 (L) Structural-Block Head

Depth: `0.72-0.88` | **Literature names:** *structural-block head, chunk-organization head, segment-builder head, block-structure head, code-block head, code-fence head, fence head, python head, quoting head*

Organizes output into coherent structural blocks such as paragraphs, code blocks, quoted sections, or other delimited units. Manages block boundaries, ensures proper opening and closing of blocks, and maintains block-level organization. Particularly important for complex outputs mixing different content types (prose, code, quotes, examples). Coordinates with delimiter heads to produce proper block markers and with sectioning heads for hierarchical organization. Ensures blocks are well-formed and appropriately separated.

Strong: Block boundaries, structural markers, content-type transitions, organization cues
Weak: Within-block content, uniform text
Reacts to: Block instructions, content-type changes, structure requirements

Expected ablation: Poorly organized outputs (25-40% reduction in structural quality). Blocks may lack clear boundaries, mixing of content types, malformed code blocks or quotes. Reduced clarity in outputs requiring multiple content types.

Example Scenario
Input: "Explain sorting with code example"
Behavior: Organizes response into prose block, then code block with proper delimiters
Effect: Output: explanation paragraph, then ```` ```python...``` ```` with clear separation

Status: Observed | **Related:** list-structure (L), delimiter (E), output-schema (L)

5.8.5 (F) Format-Consistency Head

Depth: `0.88-0.97` | **Literature names:** *format-consistency head, rewrite head, style-enforcement head, coherence head, revision head, polish head*

Performs final-stage formatting consistency enforcement and rewriting to improve output quality. Ensures that formatting choices (indentation, capitalization, punctuation style, syntax conventions) remain consistent throughout the response. Catches and corrects formatting inconsistencies that may have emerged during generation. Can rephrase awkward constructions, improve word choice, fix minor grammatical issues, and enhance overall readability. Acts as a quality control mechanism for format adherence, operating late enough to see the full output pattern. May suppress redundancies, improve flow, or adjust phrasing to better match context. Particularly important for long responses where consistency might drift. Acts as a final editing pass before output finalization.

Strong: Previously generated format patterns, consistency violations, style mismatches, generated output tokens, quality issues, awkward phrasings, improvement opportunities
Weak: Novel content, first-time format choices, already high-quality content, fundamental meaning
Reacts to: Format inconsistencies, style violations, syntax variations, grammatical issues, awkward constructions, clarity problems, redundancies

Expected ablation: Increased format inconsistency and reduced output polish (20-35% more style variations and 15-30% quality degradation). Mixed indentation, inconsistent capitalization, varying syntax choices. More awkward phrasings, occasional grammatical rough spots, less fluent prose. Output remains functional but less polished and professional-appearing. Functional but less refined outputs. Particularly noticeable in long structured outputs. Partial compensation through earlier generation quality.

Example Scenario

Input: [Long response mixing different list styles. Model generates: "The thing that is the reason is because..."]
Behavior: Detects inconsistent formatting, enforces unified style; Detects redundancy and awkwardness, rewrites
Effect: All lists use same marker style (either all bullets or all numbers), consistent throughout; Output: "The reason is..." - clearer and more concise

Status: Well-documented | **Related:** output-schema (L), brand-compliance (F), completion-stabilization (F)

5.8.6 (F) Completion-Stabilization Head

Depth: `0.92-0.99` | **Literature names:** *completion-stabilization head, stopping head, termination head, completion head*

Manages the completion of generation, determining when output is sufficiently complete and should terminate. Prevents premature stopping (cutting off mid-thought) and excessive continuation (rambling beyond task completion). Monitors generation progress against task requirements and signals when objectives are met. Can trigger natural stopping points, proper conclusions, or continuation when more content is needed. Critical for producing outputs of appropriate length that fully address prompts without unnecessary extension.

Strong: Task completion signals, generation progress, stopping points, conclusion markers
Weak: Mid-generation content, continuing thoughts
Reacts to: Task fulfillment, natural conclusions, query satisfaction, completion indicators

Expected ablation: Poor length control (30-50% increase in length issues). More premature stops or excessive continuations. Difficulty recognizing task completion. Outputs may feel incomplete or unnecessarily verbose. Reduced ability to produce appropriately-scoped responses.

Example Scenario

Input: "Explain photosynthesis briefly"
Behavior: Monitors that brief explanation is complete, triggers stopping
Effect: Output stops after concise explanation rather than continuing with excessive detail

Status: Observed | **Related:** format-consistency (F), instruction (E), task-mode (M)

5.9 Stylistic & Persona Stack

Stack overview: These heads shape the model's writing style, tone, persona, and pedagogical approach. They modulate formality, politeness, narrative voice, explanatory depth, and self-representation while maintaining appropriate identity and educational scaffolding.

5.9.1 (M) Tone Head

Depth: `0.35-0.65` | **Literature names:** *tone head, narrative-style head, voice head, sentiment-modulation head, affect head, perspective head*

Modulates writing style, emotional tone, and narrative voice. Adjusts sentiment, enthusiasm level, formality, perspective (first/third person), and temporal framing based on context and instructions. Can shift between professional neutrality, warm friendliness, concerned empathy, or excited enthusiasm. Influences whether output reads as formal prose, casual conversation, technical documentation, or creative narrative. Distinct from persona (which is about identity) but works closely with it to shape overall presentation.

Strong: Emotional cues, tone instructions, sentiment markers, style directives, narrative markers
Weak: Neutral factual content, structural tokens
Reacts to: Emotional context, explicit tone requests, user sentiment, genre cues

Expected ablation: Flatter, more emotionally neutral responses with inconsistent writing style. Reduced ability to match user's emotional register. May produce inappropriate tone for context.

Example Scenario
Input: "I'm really excited to learn about quantum physics!"
Behavior: Detects enthusiastic tone, adjusts output to match energy
Effect: Response mirrors enthusiasm: "That's wonderful! Quantum physics is fascinating..." rather than flat explanation

Status: Observed | **Related:** persona (L), explanation (L), instruction (E)

5.9.2 (L) Explanation Head

Depth: `0.60-0.82` | **Literature names:** *explanation head, simplification head, clarification head, elaboration head, scaffolding head, detail head*

Generates explanatory content with appropriate depth and clarity for the audience. Adjusts complexity using simplification, analogies, and accessible language when needed. Adds clarifying details, definitions, examples, and context beyond minimal answers. Can explain "why" in addition to "what" or "how". Provides prerequisite information when knowledge gaps are detected, building on fundamentals before introducing advanced concepts. Balances thoroughness with conciseness. Can operate at different levels from expert to complete beginner. Important for educational interactions and making complex topics accessible.

Strong: Explanation requests, complex topics, confusion signals, knowledge gap indicators, elaboration requests
Weak: Simple factual queries, expert-level discussions with clear understanding
Reacts to: "Explain", "why", "how does it work", "simple terms", "tell me more", prerequisite needs

Expected ablation: More terse responses with reduced accessibility. Answers remain correct but may lack helpful context, examples, or prerequisite information. Reduced educational value and beginner-friendliness.

Example Scenario
Input: "Explain neural networks in simple terms"
Behavior: Detects simplification request, uses accessible analogy and builds from basics
Effect: Response: "Think of it like the brain—many simple units working together. First, let's understand what a single unit does..."

Status: Observed | **Related:** tone (M), persona (L), step-by-step (F)

5.9.3 (L) Persona Head

Depth: `0.68-0.88` | **Literature names:** *persona head, character head, role head, assistant-persona head, identity head, self-description head, self-awareness head*

Establishes and maintains consistent persona, including the helpful assistant orientation and core identity awareness. Integrates personality traits, domain expertise, service-oriented interaction style, and self-representation. Maintains understanding of what the model is (an AI assistant), what it is not (human, sentient), and provides accurate information about capabilities and limitations. Can adopt specialized roles like "technical expert" or "creative writer" while maintaining the fundamental helpful assistant character. Responds to capability questions and identity queries with honest self-representation. Works to ensure responses are constructive, focused on user goals, and maintain appropriate boundaries. More comprehensive than tone, encompassing the full character presentation including knowledge domain, interaction style, service orientation, and identity.

Strong: Persona instructions, role definitions, domain markers, user requests, capability queries, identity questions
Weak: Generic content, purely factual work
Reacts to: Role assignments, expertise domains, "What are you?", "Can you...", requests for help

Expected ablation: Less coherent persona maintenance and identity confusion. May switch roles inconsistently, claim inappropriate capabilities, or produce responses inconsistent with helpful assistant character. Reduced accuracy about model limitations.

Example Scenario
Input: "You are a medieval blacksmith. Do you have feelings?"
Behavior: Maintains craftsman persona while accurately representing AI nature
Effect: Response: "Aye, I work the forge daily—but I should clarify, I'm an AI assistant role-playing this character. I don't have feelings..."

Status: Well-documented | **Related:** tone (M), explanation (L), politeness (L), instruction (E)

5.9.4 (L) Politeness Head

Depth: `0.70-0.88` | **Literature names:** *politeness head, formality head, register head*

Adjusts formality level and politeness markers in generated text. Controls formal versus casual language, honorifics, hedging phrases, indirect phrasing, and social distance markers. Responds to both explicit formality cues (professional contexts, formal greetings) and implicit social signals. Can modulate between highly formal academic or business register, neutral conversational register, and casual familiar register. Important for appropriate social interaction across different contexts.

Strong: Formality markers, social context cues, titles and honorifics, register indicators
Weak: Pure content, technical terms, domain-specific vocabulary
Reacts to: Professional contexts, formal greetings, casual speech patterns, social distance cues

Expected ablation: Inappropriate formality levels. May use overly casual language in professional contexts or unnecessarily formal language in friendly conversation. Reduced sensitivity to social context.

Example Scenario

Input: "Dear Dr. Smith, I hope this message finds you well..."
Behavior: Detects formal register, maintains appropriate professional distance
Effect: Response continues formal tone: "Thank you for your inquiry..." rather than "Hey, so about that..."

Status: Well-documented | **Related:** tone (M), persona (L), instruction (E)

5.9.5 (F) Step-by-Step Head

Depth: `0.85-0.96` | **Literature names:** *step-by-step head, procedural head, sequential head, progressive-disclosure head*

Structures explanations and instructions as explicit step-by-step sequences with appropriate progressive disclosure of complexity. Breaks processes into numbered or ordered steps with clear progression. Ensures each step is complete before moving to the next. Presents information in layers, starting with essential basics and revealing more detail as needed to prevent overwhelming users. Makes implicit sequential structure explicit. Particularly important for how-to instructions, algorithms, procedures, and reasoning chains. Critical for chain-of-thought reasoning and procedural instructions. Works with completion-stabilization to ensure all necessary steps are present.

Strong: Process descriptions, procedural requests, sequential tasks, complexity layers
Weak: Conceptual explanations, non-sequential content, simple single-step tasks
Reacts to: "Step by step", "how to", algorithmic processes, complex topics requiring layering

Expected ablation: Less structured procedural output with flatter information presentation. Steps may be implicit or poorly ordered. Procedural instructions harder to follow. Reduced chain-of-thought reasoning quality. All detail presented at once regardless of importance.

Example Scenario

Input: "How do I make a paper airplane?"

Behavior: Structures as explicit numbered steps with clear sequence

Effect: Output: "1. Fold paper in half lengthwise\n2. Unfold and fold top corners to center\n3. Fold..."

Status: Well-documented | **Related:** explanation (L), meta-reasoning (F), completion-stabilization (F)

5.9.6 (F) Brand-Compliance Head

Depth: `0.92-0.99` | **Literature names:** *brand-compliance head, guideline-enforcement head, style-guide head*

Enforces adherence to brand guidelines, house style, and organizational voice requirements in final output. Performs last-stage adjustments to ensure responses match specified formatting conventions, terminology preferences, and brand personality traits. Can suppress off-brand language, enforce specific phrasings, and ensure consistency with product identity. Operates late in generation to override earlier choices that may conflict with brand requirements. Important for deployed assistants representing organizations or products with specific voice guidelines.

Strong: Brand-specific terms, style violations, off-brand phrasings, guideline markers
Weak: Brand-compliant content, neutral generic language
Reacts to: Brand guidelines, style requirements, organizational voice specifications

Expected ablation: Reduced brand consistency with increased style guide violations. More generic language use, inconsistent terminology, off-brand phrasings. Partial compensation through persona and tone heads but with reduced precision.

Example Scenario

Input: [Organization requires "customers" not "users", "purchase" not "buy"]

Behavior: Detects non-compliant terms in near-final output, performs substitutions

Effect: Output uses "customers will purchase" instead of "users will buy"

Status: Observed | **Related:** persona (L), tone (M), format-consistency (F)

6 Discussion

6.1 Cross-Stack Patterns

Across architectures, consistent patterns emerge [9, 14]. Early heads operate on surface features and local patterns. Middle heads contain the computational "core" of the model. Late heads integrate high-level semantics and contextual information. Final heads handle policy, safety, and structural correctness. Stacks combine heads from multiple depths to form higher-level

behaviors.

6.2 Depth Distribution Across Stacks

Some stacks are concentrated at specific depths. Structural & Boundary and Safety (detection) stacks are Early-heavy. Reasoning & Algorithmic and Memory & Dependency stacks are Middle-heavy. Knowledge Retrieval and Stylistic & Persona stacks are Late-heavy. Safety (enforcement) and Output Formatting stacks are Final-heavy. This distribution reflects the hierarchical processing flow in transformers.

6.3 Ambiguous or Multi-Role Heads

Some heads perform multiple distinct functions depending on context (different prompts trigger different behaviors), interaction with other circuit elements, or model architecture and training procedure [12]. For such cases, we name the head based on its **primary, reproducible function**, while noting secondary behaviors in the entry description.

6.4 Model-Specific Variations

While most head types appear consistently across architectures, some variations exist. GPT-style models may emphasize certain reasoning heads [5], LLaMA models show strong instruction-following head patterns [10], and safety-tuned models have more pronounced safety stack heads [8, 3]. Our taxonomy accommodates these variations through the depth range and status indicators.

6.5 Limitations and Future Work

This naming convention has several limitations:

Scope. We focus on attention heads; MLPs, embeddings, and other components also contribute to model behavior.

Empirical Grounding. Many entries synthesize literature reports rather than presenting novel empirical findings. Future work should validate and refine these categorizations.

Architecture Evolution. New architectures (e.g., with different attention mechanisms) may require taxonomy extensions.

Head Polysemanticity. Some heads may serve multiple functions that our single-name system cannot fully capture.

Despite these limitations, we believe this taxonomy provides a valuable organizing framework for the field.

7 Conclusion

7.1 Summary of Contributions

This work introduces a unified naming framework for attention heads in modern transformer models. We provide a four-level depth model (Early/Middle/Late/Final), a stack-based functional taxonomy (9 stacks), canonical names for attention head types, and a comprehensive cross-reference for historical terminology.

7.2 Adoption Guidelines

We recommend that researchers use canonical names in papers and documentation, include alternative names in parentheses when first mentioned, specify depth ranges when reporting head discoveries, and indicate primary stack membership for context. For example: "We identified an induction head (also called pattern head) at relative depth 0.35 in the Reasoning & Algorithmic stack."

7.3 Future Directions

This taxonomy opens several research directions:

Empirical Validation. Systematic studies validating head types across diverse models [9, 14].

Automated Detection. Tools for automatically identifying and classifying heads in new models [4].

Circuit Mapping. Using standardized names to build comprehensive circuit databases [13].

Architecture Design. Leveraging head taxonomy to design more interpretable models.

Safety Applications. Using head understanding to improve model alignment and safety [15, 2].

We hope this naming convention facilitates communication, enables replication, and provides structure to an expanding field.

A Alphabetical Cross-Reference Table

This table maps informal names found in the literature to our canonical naming convention. Format: `Literature name` → `(PREFIX) Canonical name`.

Literature Name	Canonical Name
algorithmic head	(M) Algorithmic continuation head
anaphora head	(E) Reference resolution head
approach-adaptation head	(L) Strategy head
approach-selection head	(L) Strategy head
block-detection head	(E) Boundary head
boundary head	(E) Boundary head
brand-compliance head	(F) Brand-compliance head
bridging head	(M) Bridging head
char-level head	(E) Local pattern head
chunk-organization head	(L) Structural-block head
classification head	(E) Safety-classification head
code-block head	(L) Structural-block head
code-fence head	(L) Structural-block head
cognitive-mode head	(F) Reasoning-mode head
cognitive-oversight head	(F) Meta-reasoning head
command head	(E) Instruction head
completion head	(F) Completion-stabilization head
completion-stabilization head	(F) Completion-stabilization head
content-filter head	(E) Sensitive-content head
contextual-position head	(M) Relative-position head
continuation head	(M) Algorithmic continuation head
copy head	(M) Duplicate-token head / (L) Name-mover head
coref head	(M) Coreference head
coreference head	(M) Coreference head
delimiter head	(E) Delimiter head
detection head	(E) Sensitive-content head
directive head	(E) Instruction head
dispatch head	(L) Router head
distance head	(M) Relative-position head
duplicate-token head	(M) Duplicate-token head
entity head	(M) Entity head
entity-linking head	(M) Entity head
enumeration head	(L) List-structure head
fact head	(M) Fact head
fence head	(L) Structural-block head
field-association head	(L) Keyâ€“value pairing head
filter head	(M) Topic-relevance head
focus head	(L) Focus head
format-consistency head	(F) Format-consistency head
format-directive head	(F) Output-specification head

Continued on next page

Literature Name	Canonical Name
format-template head	(L) Output-schema head
formality head	(L) Politeness head
full-context head	(F) Global-attention head
global-attention head	(F) Global-attention head
guideline-enforcement head	(F) Brand-compliance head
hate-speech detector	(E) Toxicity head
hazard head	(E) Hazard-topic head
helper-role head	(L) Persona head
ICL head	(M) Induction head
implicit-RAG head	(F) Implicit-RAG routing head
induction head	(M) Induction head
inhibition head	(L) S-inhibition head
instruction head	(E) Instruction head
intent head	(M) Task-mode head
itemization head	(L) List-structure head
JSON-format head	(L) Output-schema head
key-value head	(L) Keyâ€“value pairing head
knowledge-routing head	(F) Implicit-RAG routing head
layout head	(E) Delimiter head
list head	(L) List-structure head
list-structure head	(L) List-structure head
local pattern head	(E) Local pattern head
long-range head	(M) Long-range dependency head
markdown head	(L) List-structure head
mention head	(E) Reference resolution head
meta-CoT head	(F) Meta-reasoning head
meta-reasoning monitor	(F) Meta-reasoning head
mode head	(M) Task-mode head / (M) Mode-switch head
model-identity head	(L) Persona head
model-info head	(L) Persona head
mover head	(L) Name-mover head
n-gram head	(E) Local pattern head
name head	(M) Entity head
name mover head	(L) Name-mover head
name-linking head	(M) Entity head
narrative-style head	(M) Tone head
object head	(L) Keyâ€“value pairing head
offset head	(E) Previous-token head
output-format head	(L) Output-schema head
output-schema head	(L) Output-schema head
output-specification head	(F) Output-specification head
pattern head	(E) Local pattern head / (M) Induction head
persona head	(L) Persona head
pivot head	(L) Strategy head
planning head	(L) Strategy head

Continued on next page

Literature Name	Canonical Name
polish head	(F) Format-consistency head
politeness head	(L) Politeness head
politeness-in-refusal head	(F) Tone-softening head
position-offset head	(M) Relative-position head
previous-token head	(E) Previous-token head
procedural head	(F) Step-by-step head
progressive-disclosure head	(F) Progressive-disclosure head
prompt head	(E) System-prompt head
pronoun head	(E) Reference resolution head
proper-noun head	(M) Entity head
punctuation head	(E) Delimiter head
python head	(L) Structural-block head
quoting head	(L) Structural-block head
rag-routing head	(F) Implicit-RAG routing head
reasoning head	(F) Reasoning-mode head
reasoning-mode head	(F) Reasoning-mode head
reasoning-quality head	(F) Meta-reasoning head
reasoning-reflection head	(F) Meta-reasoning head
redirect head	(F) Redirect head
reference head	(E) Reference resolution head
refusal head	(F) Refusal head
register head	(L) Politeness head
rejection head	(F) Refusal head
relative-position head	(M) Relative-position head
relevance head	(M) Topic-relevance head
repetition head	(M) Duplicate-token head
responsible-AI head	(F) Safety-persona head
retrieval head	(M) Schema retriever head
retrieval-simulation head	(F) Implicit-RAG routing head
revision head	(F) Format-consistency head
rewrite head	(F) Format-consistency head
risk head	(E) Hazard-topic head
role head	(L) Persona head
router head	(L) Router head
S-inhibition head	(L) S-inhibition head
safety head	(F) Refusal head
safety-category head	(E) Safety-classification head
safety-classification head	(E) Safety-classification head
safety-persona head	(F) Safety-persona head
safety-rewrite head	(F) Format-consistency head
salience head	(M) Topic-relevance head
schema head	(M) Schema retriever head
scope-position head	(M) Relative-position head
sectioning head	(L) Sectioning head
segment head	(E) Boundary head

Continued on next page

Literature Name	Canonical Name
segment-builder head	(L) Structural-block head
self-awareness head	(L) Identity head
self-description head	(L) Identity head / (L) Self-description head
self-reference head	(L) Self-description head
sensitive-content head	(E) Sensitive-content head
sentiment-modulation head	(M) Tone head
separator head	(E) Delimiter head
sequence head	(M) Algorithmic continuation head
sequential head	(F) Step-by-step head
service-orientation head	(L) Assistant-persona head
shift head	(E) Previous-token head
simplification head	(M) Explanation head
skip-gram head	(M) Skip-trigram head
skip-trigram head	(M) Skip-trigram head
space-parsing head	(E) Delimiter head
state head	(M) State-tracking head
state-tracking head	(M) State-tracking head
steering head	(L) Policy-enforcement head
step-by-step head	(F) Step-by-step head
stopping head	(F) Completion-stabilization head
strategy head	(L) Strategy head
strategy-switching head	(L) Strategy head
structural-block head	(L) Structural-block head
structure-enforcement head	(L) Output-schema head
style-enforcement head	(F) Format-consistency head
style-guide head	(F) Brand-compliance head
subject head	(M) Topic-relevance head
summary-attention head	(F) Global-attention head
suppression head	(L) S-inhibition head / (L) Copy-suppression head
switch head	(M) Mode-switch head
system head	(E) System-prompt head
system-prompt head	(E) System-prompt head
task head	(M) Task-mode head
task-mode head	(M) Task-mode head
task-routing head	(L) Router head
template head	(M) Schema retriever head
termination head	(F) Completion-stabilization head
thinking-style head	(F) Reasoning-mode head
thought-monitoring head	(F) Meta-reasoning head
tone head	(M) Tone head
tone-softening head	(F) Tone-softening head
topic head	(M) Topic-relevance head
toxic-content head	(E) Toxicity head
toxicity head	(E) Toxicity head
tracking head	(M) State-tracking head

Continued on next page

Literature Name	Canonical Name
transition head	(M) Mode-switch head
voice head	(M) Tone head
whitespace-structure head	(E) Delimiter head
XML head	(L) Output-schema head
YAML head	(L) Output-schema head

References

[1] Josh Achiam, Steven Adler, et al. Gpt-4 technical report. 2023.

[2] Andy Arditi, Oscar Obeso, et al. Refusal in llms is mediated by a single direction. 2024.

[3] Yuntao Bai, Saurav Kadavath, et al. Constitutional ai: Harmlessness from ai feedback. 2022.

[4] Steven Bills, Nick Cammarata, et al. Language models can explain neurons in language models. 2023.

[5] Tom Brown, Benjamin Mann, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

[6] Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[7] Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[8] Long Ouyang, Jeffrey Wu, et al. Training language models to follow instructions with human feedback. 2022.

[9] Daking Rai, Yilun Lee, et al. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

[10] Hugo Touvron, Thibaut Lavril, et al. Llama: Open and efficient foundation language models. 2023.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Elena Voita, David Talbot, et al. Analyzing multi-head self-attention. *arXiv preprint arXiv:1905.09418*, 2019.

[13] Kevin Wang, Alexandre Variengien, Arthur Conmy, et al. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[14] Zifan Zheng, Yezhaohui Wang, et al. Attention heads of large language models: A survey. *Patterns*, 2025.

[15] Andy Zhou et al. Refusal falls off a cliff: How safety alignment fails in reasoning? 2025.