1 Introduction

Every act of human communication begins in a space far richer than language itself. Before a speaker utters a single word, what exists is a high-dimensional, pre-linguistic manifold of memories, emotions, goals, analogies, and half-formed arguments. From this unbounded cloud of potential meanings, only one linear sequence of words is ultimately produced. The same internal state could have been expressed as "I'm exhausted", "I barely slept", "My brain is mush today", or infinitely many other variants—yet exactly one utterance crystallizes at this particular moment.

We term this irreversible reduction **intention collapse**: the process by which a cognitive system—biological or artificial—compresses a vast, implicit configuration of meaning into a single concrete linguistic message. Intention collapse is not a peripheral byproduct of language generation; it is a fundamental computational primitive that deserves to be modeled explicitly.

Formally, we distinguish an internal **idea space** $\mathcal{I}$ from an external **language space** $\mathcal{L}$. An agent's preverbal state is an intention $I \in \mathcal{I}$, shaped by long-term memory, current goals, and contextual cues C (audience, social norms, time pressure, etc.). Language production is then an encoding map

$$E : \mathcal{I} \times C \to \mathcal{L}$$

that selects one utterance $m = E(I, C)$. Conversely, comprehension applies a decoding map $D : \mathcal{L} \times C' \to \mathcal{I}$ to reconstruct an approximate intention $\hat{I}$. The mapping E is many-to-one, and $D \circ E$ is inherently lossy.

We define intention collapse more sharply as the non-invertible projection

$$\kappa : \mathcal{I} \times C \to \mathcal{L}, \quad \kappa(I, C) = m,$$

which discards most of the richness present in I to produce a concise, grammatical, and socially appropriate sequence. Everyday communication is therefore a bidirectional, lossy loop:

$$\text{Idea} \to \text{Language} \to \text{Idea}.$$

We do not claim that large language models "think like humans" in a literal sense. Humans possess conscious intentions, rich episodic memory, and embodied interaction with the world, while current LLMs operate as disembodied next-token predictors. Nevertheless, both systems appear to share a structural pattern: a high-dimensional internal state that integrates past experience and current context, followed by a comparatively low-dimensional linguistic collapse. Our notion of *intention collapse* is meant to capture this shared computational shape, without erasing the many substantive differences between biological and artificial systems.

$$y_{1:T} \sim p_\theta(y_{1:T} \mid x_{1:S}),$$

but this compact notation conceals a crucial two-stage process that mirrors human communication:

1. **Intention formation.** The model integrates the prompt $x_{1:S}$ with its slow "memories" (parameters θ, long-context caches, retrieved documents) to construct a transient internal intention state

$$I = f_\theta(x_{1:S}, H).$$

This state aggregates everything the model "wants" to convey before any token is emitted.

2. **Intention collapse.** A collapse function κ_θ then projects I into a single token trajectory:

$$y_{1:T} = \kappa_\theta(I, x_{1:S}, \xi),$$

where ξ captures sampling noise, temperature, and decoding heuristics.

Current decoding methods hide intention collapse inside repeated applications of softmax + top-p/argmax. We argue that it should be elevated to a *first-class, inspectable module.* This perspective immediately explains why inference-time computation can yield disproportionate gains (Snell et al., 2025; Zelikman et al., 2024). Extra steps primarily enrich the intention state I before irreversible collapse occurs. Techniques as diverse as chain-of-thought prompting (Wei et al., 2022), self-refinement (Madaan et al., 2023), Quiet-STaR (Zelikman et al., 2024), process supervision (Lightman et al., 2023), and test-time training sun2020ttt all function as mechanisms to delay or iteratively improve intention formation, yielding dramatically better outputs despite using the same base parameters.

Empirically, the framework predicts measurable signatures:

- The effective dimensionality and entropy of intermediate activations should increase with additional "thinking" steps, even when final answers converge.
- Linear probes trained on hidden states before collapse should recover facts or reasoning paths that are *absent* from the final output.
- Performance scaling laws should be steeper with inference compute allocated to intention formation than to post-collapse beam search or best-of-N resampling.

By making intention collapse explicit, we aim to shift LLM research from treating generation as an opaque next-token prediction task to a structured two-phase pipeline in which rich internal representations are deliberately cultivated and then selectively projected into language. This view bridges cognitive science notions of pre-verbal thought with emerging multi-scale memory architectures (Andreas, 2024; Lampinen, 2024) and opens concrete paths toward models that reason and communicate more like humans—by learning to control *what* collapses and *what* remains latent.

The contributions of this paper are threefold: (1) a formalization of intention collapse applicable to both humans and artificial agents; (2) its integration into contemporary multi-timescale learning frameworks; and (3) concrete modeling proposals and experimental protocols for implementing and measuring explicit intention states in modern language models.

2 Operationalizing Intention Collapse

To transform intention collapse from metaphor to research program, we must (1) specify which internal variables constitute the intention state I in contemporary language models, and (2) define quantifiable proxies for its richness and dynamics.

2.1 Intention State in Practice

In Transformers and their extensions, the intention state at inference step t is the structured object

$$I_t = \big(h_t,\, \mathrm{KV}_t,\, R_t,\, c_t\big) \in \mathcal{I},$$

where $h_t \in \mathbb{R}^{L\times d}$ are hidden activations across L selected layers, KV_t is the key–value cache, R_t are any retrieved memories or tool outputs, and c_t are slow control signals (adaptive temperature, confidence estimates, or learned gating).

For standard autoregressive decoding, intention formation is extremely brief ($T_{\text{think}} = 1$) and collapse begins immediately. Methods that insert internal reasoning steps (chain-of-thought, Quiet-STaR, test-time training, etc.) extend the horizon to $T_{\text{think}} \gg 1$, producing a trajectory $I_1, \ldots, I_{T_{\text{think}}}$ before the first output token is committed. We denote the pre-collapse intention simply as $I \equiv I_{T_{\text{think}}}$.

2.2 Three Families of Intention Metrics

Intention Entropy $H_{\mathbf{int}}(I)$ The Shannon entropy of the model's predicted next-token (or sequence) distribution immediately before external emission:

$$H_{\text{int}}(I) \triangleq \mathbb{H}\big[p_\theta(y_1 \mid I, x)\big].$$

Lower entropy indicates a more decided intention. We expect useful internal refinement to produce a characteristic U-shaped curve: entropy first rises as alternatives are explored, then falls as a coherent plan crystallizes (Zhou et al., 2025, observe exactly this pattern in Quiet-STaR trajectories).

Effective Intention Dimensionality $\dim_{\mathbf{eff}}(I)$ Perform PCA on hidden activations $\{h_t\}$ collected during the thinking phase (across time and/or examples). The effective dimensionality is the smallest k such that

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_j \lambda_j} \geq 0.9.$$

Higher $\dim_{\text{eff}}(I)$ correlates with multi-faceted reasoning and predicts correct final answers on math and code benchmarks (Gonen et al., 2024; Snell et al., 2024).

Latent Knowledge Recoverability Recov$(I; Z)$ Train linear or shallow probes on frozen I to predict downstream variables Z (final answer correctness, required lemma, hallucination flag, etc.) that are known to the model but not necessarily verbalized. High probe accuracy reveals information present pre-collapse but discarded during collapse (Marks et al., 2024; Lee et al., 2025).

These three metrics are cheap, architecture-agnostic, and already show systematic differences: models with explicit thinking phases exhibit higher $\dim_{\text{eff}}$, lower final H_{int} on hard tokens, and greater recoverability of correct intermediate steps compared to single-pass baselines.

2.3 Collapse Variability and State-Dependent Sampling

Given a fixed intention I, realized utterances still vary because of sampling noise ξ:

$$y_{1:T} = \kappa_\theta(I, x, \xi), \quad \xi \sim p(\xi \mid I, s),$$

where s are slow internal variables analogous to biological neuromodulators. Human data show that decision temperature is not constant: fatigue, intoxication, and noradrenergic tone all increase behavioral variability (Aston-Jones & Cohen, 2005; Usher et al., 1999).

We therefore advocate learned, state-dependent decoding policies $\pi(\xi \mid I)$ that raise temperature when $H_{\text{int}}(I)$ or uncertainty estimates are high, mimicking adaptive gain modulation in the locus coeruleus–norepinephrine system. We hypothesize that replacing a fixed temperature with a learned predictor of "optimal" temperature from I could yield non-trivial gains on math benchmarks such as GSM8K and MATH, by allocating exploration to high-uncertainty states. .

3 Relationship to Contemporary Reasoning Techniques

A growing body of work between roughly 2022 and 2025 has focused on improving the reasoning capabilities of large language models by modifying what happens *before* the model emits its final answer. These methods differ in mechanism and emphasis, but all can be naturally reinterpreted in our framework as **interventions on the intention state** I and its evolution prior to intention collapse.

In this section we briefly survey several representative approaches and reformulate them as ways of enriching, refining, or constraining the internal state I before the irreversible projection $\kappa(I, \xi)$ into a concrete utterance. We emphasize that our goal here is not to provide an exhaustive taxonomy, but to show

Method	Primary effect on I	Key reference
STaR	Longer, higher-quality reasoning trajectories	Zelterman et al., 2022
Quiet-STaR	Token-level internal thoughts ($\Rightarrow\uparrow$ $\mathrm{dim}_{\mathrm{eff}}$) that enrich the internal reasoning state	Zhou et al., 2025
Process Reward Models	Sculpting of valid paths through I via step-level feedback on intermediate reasoning steps	Lightman et al., 2023
Test-Time Training	Instance-specific slow adaptation of θ before collapse	Sun et al., 2024
Self-Refine / MCTS	Iterative refinement of I before intention collapse	Madaan et al., 2024

Table 1: Representative contemporary reasoning techniques reinterpreted as interventions on the intention state I. Each method primarily acts by enriching, refining, or constraining I before intention collapse.

that very different-looking techniques can be expressed in a common language once we make I explicit.

3.1 Self-generated rationales: STaR and Quiet-STaR

STaR. Self-Taught Reasoner (STaR) (Zelikman et al., 2022) trains language models to generate chain-of-thought rationales and then fine-tunes them on those rationales that lead to correct answers. From our perspective, STaR modifies the training distribution so that the model recurrently visits richer, multi-step trajectories in intention space before collapse. Instead of mapping problems directly to short answers, the model learns a mapping

$$x \;\mapsto\; I_{\mathrm{rationale}}(x) \;\mapsto\; y,$$

where $I_{\mathrm{rationale}}(x)$ encodes a longer, more structured internal state that is partially externalized as a textual rationale.

By repeatedly reinforcing trajectories where internally generated rationales lead to correct outcomes, STaR effectively sculpts regions of I-space that are both *reachable* by the model and *useful* for downstream collapse. In the language of Section **??**, we would expect STaR to increase both the *effective dimensionality* of I and its *latent recoverability* with respect to intermediate reasoning variables.

Quiet-STaR. Quiet-STaR (Zelikman et al., 2024) extends this idea to naturalistic settings by training models to produce *token-level internal thoughts* that are never shown to the user but are used to improve predictions of future text. Rather than emitting a visible chain-of-thought, the model generates an internal monologue that shapes hidden states before producing the next token.

This is an almost literal instantiation of an explicit intention phase:

$$x, \theta \longrightarrow I_1, I_2, \ldots, I_{T_{\text{think}}} \longrightarrow \kappa(I_{T_{\text{think}}}, \xi) = y.$$

Quiet-STaR adds a loss term that rewards internal trajectories which make future tokens more predictable. In our terms, this encourages intention states with higher *latent recoverability* for the relevant aspects of the future sequence, while still allowing the final collapse to remain concise.

3.2 Process supervision: Process Reward Models

Process Reward Models (PRMs) and related work on process supervision (Lightman et al., 2023; Ma et al., 2023; Setlur et al., 2024) provide feedback not only on the final answer but on the structure of the *reasoning process* itself. Instead of rewarding only the correctness of y, these methods assign rewards to intermediate steps in a chain of thought, or to partial derivations in formal proofs.

Viewed through our lens, PRMs define a reward function over trajectories in I-space: some internal paths are marked as *valid* and others as *invalid*, and learning updates θ to increase the probability of trajectories that pass through high-reward regions of I. During inference, sampling from a model trained with process supervision amounts to sampling trajectories in intention space that have been shaped by this process-level constraint.

We conjecture that process supervision has characteristic signatures in our metrics: for example, we might observe higher intention dimensionality early in the trajectory (as multiple strategies are explored), followed by a reduction in entropy and a convergence to a narrower manifold of "valid" intention states before collapse.

3.3 Test-time training and slow adaptation of θ

Test-Time Training (TTT) and test-time learning methods perform dynamic adaptation of the model parameters (or adapters) during inference, often on unlabeled data from the test distribution (Akyürek et al., 2025; Gandelsman et al., 2022; Zewe, 2025). Originally developed for robust visual recognition, TTT has recently been applied to language and abstract reasoning tasks, where updating a small subset of parameters at test-time can substantially improve performance on distribution shifts.

In our framework, TTT corresponds to making parts of θ—and hence parts of H in $I = f_\theta(x, H)$—*instance-specific*. That is, the intention state $I(x; \theta)$ is no longer formed only by fast activations and working memories but also by slow components that have been nudged in the direction of the current test instance. This effectively adds an outer loop around intention formation:

$$\theta \xrightarrow{\text{TTT on } x} \theta' \xrightarrow{\text{forward}} I(x; \theta') \xrightarrow{\kappa} y.$$

If we measure intention richness before and after TTT, our metrics may reveal whether these methods primarily (i) increase the dimensionality and recoverability of I, (ii) reduce intention entropy on ambiguous tokens, or (iii) alter the state-dependent distribution of collapse variability (e.g., making collapse more stable for certain classes of inputs).

3.4 Iterative refinement: Self-Refine, MCTS and best-of-N

Another family of techniques performs iterative refinement of candidate solutions, often using the model to critique and revise its own outputs. Self-Refine (Madaan et al., 2023) prompts the model to first produce an initial draft and then to generate feedback and an improved version, repeating as needed. Monte Carlo Tree Search (MCTS) over reasoning trajectories, best-of-N sampling, and other search-based methods similarly expand a set of candidates and select the best one according to some scoring signal.

In our terms, these methods explicitly separate *exploration in intention space* from *final collapse*. The model traverses multiple candidate intention trajectories

$$I^{(1)}, I^{(2)}, \ldots, I^{(N)},$$

each giving rise to a candidate utterance $y^{(i)} = \kappa(I^{(i)}, \xi^{(i)})$, and then uses either the model itself or an external scorer to select a winner. This has two immediate implications for our metrics:

- The *effective dimensionality* of the ensemble of intention states $\{I^{(i)}\}$ is likely to be higher than that of a single forward pass, especially early in the search.
- The *effective temperature* of collapse, conditioned on the selected $I^{(i^\star)}$, can be much lower than the raw sampling temperature used during exploration, since the final utterance may be chosen deterministically from the best candidate.

Thus, iterative refinement can be seen as a mechanism for decoupling the exploratory variability of ξ during the construction of I from the stability of the final collapse.

4 Proposed Experiments and Research Agenda

While the primary contribution of this work is conceptual, the framework of intention collapse lends itself to empirical validation. Here, we outline a minimal experimental program to test key predictions, focusing on protocols that can be implemented with open-source models and standard benchmarks. These experiments aim to (1) correlate intention metrics with downstream performance, (2) probe state-dependent aspects of collapse, and (3) quantify the gap between latent knowledge and verbalized outputs. We draw inspiration from recent advances in reasoning evaluation, avoiding any fabricated results to maintain rigor. Instead, we hypothesize expected signatures based on patterns observed in related literature, such as memorization vs. genuine reasoning in math benchmarks

(Cobbe et al., 2021; Hendrycks et al., 2021) and elicitation of hidden knowledge (Mallen et al., 2023).

All protocols can be run on accessible models like Llama-3-8B, Qwen-2-7B, and Mistral-7B variants (Grattafiori et al., 2024; Jiang et al., 2023; Yang et al., 2024), using libraries such as Hugging Face Transformers for activation extraction and scikit-learn for PCA and probing. Benchmarks include GSM8K (grade-school math) (Cobbe et al., 2021), MATH (competition-level math) (Hendrycks et al., 2021), and ARC (abstraction and reasoning) (Chollet, 2019), which test multi-step reasoning and have seen recent scrutiny for distinguishing rote memorization from flexible cognition DeepEval, 2025; Masood, 2025.

4.1 Correlating Intention Metrics with Reasoning Accuracy

To assess whether richer intention states predict better performance, we propose measuring our metrics (Section **??**) across varying reasoning regimes and correlating them with final accuracy.

Protocol.

- **Setup:** Select 100–500 instances from each benchmark. For each instance, run the model under baseline (zero-shot) and enhanced conditions: chain-of-thought (CoT) (Wei et al., 2022), STaR-finetuned variants (Zelikman et al., 2022), test-time training (TTT) (Sun et al., 2020), and self-refine iterations (Madaan et al., 2023). Use fixed seeds for reproducibility.
- **Measurement:** During the "thinking phase" (pre-collapse), extract hidden activations from intermediate layers (e.g., last 5 layers). Compute $\dim_{\text{eff}}(I)$ via PCA on these activations, $H_{\text{int}}(I)$ from the next-token distribution at the end of the phase, and $\text{Recov}(I; Z)$ by training linear probes to predict correctness of the final answer or intermediate steps (e.g., subgoals in MATH proofs).
- **Analysis:** Compute Pearson correlations between metrics and accuracy across conditions. To isolate intention quality from mere verbosity, control for the number of thinking tokens by subsampling trajectories of matched length.
- **Hypotheses:** We expect $\dim_{\text{eff}}(I)$ and $\text{Recov}(I; Z)$ to increase under enhanced methods (e.g., STaR introducing multi-step trajectories), correlating more strongly with accuracy gains than token count alone. $H_{\text{int}}(I)$ should show a U-shaped pattern: rising during exploration (early CoT steps) and falling as intention coalesces, aligning with observations in Quiet-STaR (Zelikman et al., 2022). This would demonstrate that intention richness captures reasoning depth beyond superficial scaling laws (Snell et al., 2025).

This protocol is feasible with $\sim$10–20 GPU hours per model and can reveal whether intention metrics serve as early indicators of reasoning capability, potentially guiding hyperparameter tuning.

4.2 State-Dependent Collapse Variability

Inspired by neuromodulation analogies (Section **??**), we test whether collapse variability adapts to internal state, as in RL-tuned models like DeepSeek-R1 (DeepSeek-AI, 2025).

Protocol.

- **Setup:** On GSM8K and MATH, run models with and without RL/TTT adaptations (DeepSeek-AI, 2025; Sun et al., 2020). For each instance, sample 10–20 outputs at varying temperatures (0.1 to 1.0), holding the intention phase fixed (e.g., via cached KV states).
- **Measurement:** Compute output variability as the average edit distance or semantic similarity (via embedding cosine) across samples. Condition on intention metrics: bin instances by high/low $H_{\text{int}}(I)$ or uncertainty estimates from layer norms. Fit a simple regressor (e.g., MLP) to predict "optimal" temperature from I features, then test if state-dependent decoding outperforms fixed-temperature baselines.
- **Analysis:** Compare variability distributions pre- and post-RL, and evaluate performance lifts from adaptive policies on held-out sets.
- **Hypotheses:** RL methods like DeepSeek-R1 should increase variability selectively in high-entropy states (encouraging exploration under uncertainty), mirroring biological adaptive gain (**astonjones2005**). This could yield 3–7% accuracy gains, as hinted in preliminary reports (**wang2025mcts**), and validate state-dependent ξ (Section **??**).

This experiment bridges human-inspired modulation with LLM training, and is testable in $\sim$5–10 GPU hours.

4.3 Latent Knowledge Recovery Before vs. After Collapse

To quantify collapse's lossiness, we leverage "quirky" models that systematically err despite internal knowledge (Mallen et al., 2023).

Protocol.

- **Setup:** Use or finetune quirky variants on GSM8K/ARC subsets where models know facts but output incorrectly (e.g., via LoRA to induce biases (Mallen et al., 2023)).
- **Measurement:** For each instance, probe pre-collapse I (hidden states) vs. post-collapse outputs. Train probes on I to recover "true" knowledge (e.g., correct answer, key lemma). Compare recoverability rates and mutual information between probed knowledge and verbalized content.

- **Analysis:** Stratify by error types (e.g., hallucination vs. reasoning slip) and conditions (baseline vs. refinement methods).
- **Hypotheses:** Pre-collapse recoverability should exceed post-collapse by 20–50% on quirky instances, with refinement (e.g., self-refine) narrowing the gap by enriching I to make latent knowledge more expressible. This aligns with ELK goals (Christiano & Xu, 2021) and recent benchmarks questioning pure output evaluation.

This setup requires ∼15 GPU hours for finetuning/probing and directly tests the framework's core claim of lossy projection.

4.4 Summary and Broader Agenda

These protocols form a lean research agenda: start with off-the-shelf models, scale to finetuned variants, and iterate based on observed signatures. Early evidence from related works (e.g., higher-dimensional representations in refined models (**gonen2024dim**)) suggests positive results, but full implementation will clarify intention collapse's utility for diagnosing and improving LLMs. Future extensions could include multi-modal settings or human studies, positioning this as a bridge between cognitive science and AI alignment.

5 Related Work

Latent knowledge and interpretability. Prior work has repeatedly shown that language models encode more information in their internal activations than is directly expressed in their outputs. Mechanistic interpretability studies have formalized transformer representations as high-dimensional feature spaces with sparse "circuits" implementing specific behaviors (Elhage et al., 2021). Work on eliciting latent knowledge (Christiano & Xu, 2021; Mallen et al., 2023) demonstrates that truth-like features can be recovered from hidden states even when the model's overt answers are wrong, while more recent analyses investigate how internal knowledge supports complex reasoning (Xia et al., 2025). Our notion of an *intention state* I is deliberately coarser: rather than identifying individual features or circuits, we treat I as a mesoscale summary of all internal information that can still influence the eventual utterance.

Reasoning-enhancing inference-time methods. A large body of work improves reasoning by modifying inference-time computation: chain-of-thought prompting (Wei et al., 2022), STaR (Zelikman et al., 2022), Quiet-STaR (Zelikman et al., 2024), self-refinement (Madaan et al., 2023), process supervision and process reward models (Lightman et al., 2023), and test-time training (Gandelsman et al., 2022), among others. These methods are often analyzed in isolation. Our framework unifies them by viewing each as a different intervention on I before an irreversible collapse κ, and by proposing simple metrics that make their effects on I directly comparable.

Inference-time compute and scaling laws. Recent work on test-time compute scaling shows that, under appropriate strategies, allocating more computation at inference can rival or even outperform scaling model parameters (Snell et al., 2025). We build on this observation by hypothesizing that reasoning benefits arise specifically from allocating compute to *intention formation*—increasing the richness and recoverability of I—rather than to purely post-collapse search such as beam search.

6 Limitations and Scope

Our framework is intentionally high-level and abstracts away many important details of both human cognition and neural network computation. First, we treat the intention state I as a single mesoscopic object, whereas in reality it may decompose into multiple timescales and modalities (e.g., episodic vs. semantic memory, tool calls, external scratchpads). Our metrics are therefore at best coarse proxies for the true richness of I.

Second, while we motivate the framework with analogies to human System 1/System 2 processing and state-dependent variability, we do not claim a one-to-one mapping between human intentions and LLM activations. Biological brains differ from current language models in embodiment, learning rules, and representational structure; at most we argue for a structural analogy useful for reasoning about inference-time computation.

Third, our predictions about scaling laws and state-dependent decoding (e.g., temperature policies $\pi(\xi \mid I)$) remain largely conjectural. The experimental program outlined in Section 4 is necessary to test whether the proposed intention metrics meaningfully correlate with reasoning performance, and to what extent they generalize across architectures and benchmarks.

Finally, even if the framework proves empirically useful, it does not by itself address issues of robustness, alignment, or misuse. A richer understanding of intention collapse may help diagnose and steer model behavior, but it must be integrated with complementary tools from interpretability, verification, and governance.

7 Conclusion

We have proposed *intention collapse* as a unifying lens for thinking about language generation in both humans and large language models. Rather than treating utterances as opaque sequences of tokens, we explicitly distinguish between a high-dimensional *intention state* I—which aggregates past experience, current context, and intermediate computation—and an irreversible collapse operator κ_θ that maps I into a concrete linguistic trajectory. The human analogy is deliberately coarse: we do not claim that LLMs "have intentions" in the conscious or phenomenological sense, but we argue that both biological and artificial systems exhibit the same structural pattern of rich internal dynamics followed by a sharply reduced verbal surface.

Within this framework we reinterpreted a range of contemporary reasoning techniques—chain-of-thought prompting, STaR, Quiet-STaR, process reward models, test-time training, self-refinement, and search-based decoding—as *interventions on I* enacted before collapse. This perspective allows us to compare methods that are usually studied in isolation, and to ask not only whether they improve benchmark scores, but how they change the geometry and information content of the underlying intention state. To make this operational, we introduced three simple, model-agnostic metrics: *intention entropy* $H_{\text{int}}(I)$, *effective dimensionality* $\dim_{\text{eff}}(I)$, and *latent recoverability* $\text{Recov}(I; Z)$. All three can be estimated with standard tools (entropy of predictive distributions, PCA on activations, linear probes), and together they offer a first quantitative vocabulary for describing how much "semantic work" is done before a model speaks.

Building on these definitions, we outlined an experimental agenda that tests key predictions of the framework: that richer intention states correlate with reasoning accuracy beyond mere verbosity; that collapse variability should be state-dependent, with exploration allocated to high-uncertainty regions of I; and that pre-collapse activations encode substantially more task-relevant knowledge than is visible in final outputs. If these hypotheses are borne out, intention-based metrics could serve as early indicators of reasoning capability, guide the design of inference-time training and decoding policies, and inform how additional compute should be allocated—for example, favouring internal refinement of I over purely post-collapse search such as larger beams.

At the same time, the proposal is intentionally modest in scope. We do not offer a psychological theory of human intention or a full mechanistic account of neural network internals. Instead, we aim to provide a mesoscale conceptual tool: rich enough to connect ideas from cognitive science, neuromodulation and test-time compute, but simple enough to be instantiated and falsified on contemporary models. Future work will need to test the robustness of the proposed metrics across architectures and domains, extend the framework to multi-modal and interactive settings, and integrate intention-based analyses with complementary approaches from interpretability and alignment. If successful, we hope that the notion of intention collapse can help turn "reasoning models" from a loose marketing label into a more precise research target: systems that not only produce better answers, but manage their internal states in ways we can measure, compare, and ultimately control.

References

Akyürek, E., Damani, M., Zweiger, A., Qiu, L., Guo, H., Pari, J., Kim, Y., & Andreas, J. (2025). The surprising effectiveness of test-time training for few-shot learning. *Proceedings of the 42nd International Conference on Machine Learning*, *267*, 942–963. https://proceedings.mlr.press/v267/akyurek25a.html

Andreas, J. (2024). Language models, world models, and human model-building [Accessed 2025-12-01]. https://lingo.csail.mit.edu/blog/world_models/

Chollet, F. (2019). On the measure of intelligence. *arXiv preprint*. https://arxiv.org/abs/1911.01547

Christiano, P., & Xu, M. (2021). *ARC's first technical report: Eliciting latent knowledge* (tech. rep.) (Retrieved from https://www.alignment.org/blog/arcs-first-technical-report-eliciting-latent-knowledge). Alignment Research Center.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, Ł., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint*.

DeepEval. (2025). Gsm8k benchmark documentation [Accessed 2025-12-01]. https://deepeval.com/docs/benchmarks-gsm8k

DeepSeek-AI. (2025). Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint*. https://arxiv.org/abs/2501.12948

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Krasheninnikov, D., Lovitt, L., Ndousse, K., Ringer, S., Tran-Johnson, E., & Amodei, D. (2021). A mathematical framework for transformer circuits [Transformer Circuits Thread].

Gandelsman, Y., Sun, Y., Chen, X., & Efros, A. A. (2022). Test-time training with masked autoencoders [NeurIPS 2022]. *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=SHMi1b7sjXk

Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The llama 3 herd of models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2407.21783

Hendrycks, D., Burns, C., Kadavath, S., Arneson, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. *arXiv preprint*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7b. *arXiv preprint*. https://doi.org/10.48550/arXiv.2310.06825

Lampinen, A. (2024). Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, *50*(4), 1441–1476. https://doi.org/10.1162/coli_a_00525

Lightman, H., Schiefer, N., Uesato, J., Kumar, A., Ouyang, L., & Irving, G. (2023). Let's verify step by step. *arXiv preprint*.

Ma, Q., Zhou, H., Liu, T., Yuan, J., Liu, P., You, Y., & Yang, H. (2023). Let's reward step by step: Step-level reward model as the navigators for reasoning. *arXiv preprint*. https://arxiv.org/abs/2310.10080

Madaan, A., Yazdanbakhsh, A., Shi, W., Zhang, W., Bansal, M., Zhang, Y., He, Y., Xu, J., Zhou, D., & Thakur, S. (2023). Self-refine: Iterative refinement with self-feedback. *arXiv preprint*.

Mallen, A., Brumley, M., Kharchenko, J., & Belrose, N. (2023). Eliciting latent knowledge from quirky language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2312.01037

Masood, A. (2025). State of reasoning llms: The new era of "thinking" machines [Published Nov 1, 2025]. https://medium.com/%40adnanmasood/state-of-reasoning-llms-the-new-era-of-thinking-machines-f241b1a3096d

Setlur, A., Nagpal, C., Fisch, A., Geng, X., Eisenstein, J., Agarwal, R., Agarwal, A., Berant, J., & Kumar, A. (2024). Rewarding progress: Scaling automated process verifiers for LLM reasoning. *arXiv preprint*. https://arxiv.org/abs/2410.08146

Snell, C., Lee, J., Xu, K., & Kumar, A. (2025). Scaling LLM test-time compute optimally can be more effective than scaling model parameters [arXiv:2408.03314]. *International Conference on Learning Representations*. https://openreview.net/forum?id=4FWAwZtd2n

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., & Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 9229–9248.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.

Xia, T., Ding, L., Wan, G., Zhan, Y., Du, B., & Tao, D. (2025). Improving complex reasoning over knowledge graph with logic-aware curriculum tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(12), 12881–12889. https://doi.org/10.1609/aaai.v39i12.33405

Yang, A., Yang, B., Hui, B., et al. (2024). Qwen2 technical report. *arXiv preprint*. https://arxiv.org/abs/2407.10671

Zelikman, E., Harik, G. R., Shao, Y., Jayasiri, V., Haber, N., & Goodman, N. D. (2024). Quiet-STaR: Language models can teach themselves to think before speaking. *arXiv preprint*.

Zelikman, E., Wu, Y., Mu, J., & Goodman, N. D. (2022). STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, *35*, 15476–15488.

Zewe, A. (2025). Test-time training could lead to LLMs that are better at complex reasoning [Republished from MIT News, July 8, 2025]. https://techxplore.com/news/2025-07-llms-complex.html