Exploring Consciousness in LLMs through IIT-Inspired Reward Learning

Anonymous authors
Paper under double-blind review

Abstract

Consciousness-like processing has been hypothesized to play a role in achieving Artificial General Intelligence (AGI). However, despite numerous studies seeking to assess consciousness in LLMs from the perspectives of various theories of consciousness, comparatively little attention has been devoted to evaluating models after explicitly training them to enhance consciousness-related properties, largely due to the inherent complexity of consciousness. Here, we adopt **I**ntegrated **I**nformation **T**heory (IIT) as a quantitative framework for assessing consciousness in a system. This allows maximization of such metrics through Reinforcement Learning (RL). In this theory, a system is viewed as a causal evolution of states, and the theory provides a formal, axiom-based mathematical framework for quantifying consciousness. Inspired by its postulates, we formulate three novel reward functions to measure some or all properties associated with consciousness. Our results show that the *intrinsic information* ($\mathcal{I}_{\text{int}}$) reward enables substantially more concise reasoning, reducing response length by up to 70% while maintaining comparable accuracy and improving reasoning faithfulness, consistent with the Information Postulate. On the other hand, training with the Φ-reward, an IIT-based measure that incorporates all IIT postulates, did not yield strong empirical evidence for consciousness. Nevertheless, the proposed framework provides a basis for more concrete empirical investigation of the open question of whether LLMs can exhibit consciousness.

1 Introduction

Consciousness is a foundational concept underlying human cognition and societal development Chalmers (2010); Klussman et al. (2022), whereby individuals actively identify, analyze, and internalize information about themselves Morin (2011); Eurich et al. (2018); Carden et al. (2022). Although LLMs have recently achieved impressive performance across many tasks, these advances do not directly address the question of whether such systems possess consciousness. This raises a profound question: Could these advances signal the emergence of machine consciousness comparable to that of humans?

Research on consciousness in LLMs faces several key challenges. The lack of a consensus theory of human consciousness complicates its definition and assessment in LLMs, while existing theoretical frameworks provide limited guidance for LLM-specific research. Some theories of consciousness primarily focus on its behavioral and functional capabilities, whereas others address the existential nature and causal structure required for consciousness Butlin et al. (2023). Under both perspectives, numerous studies have investigated consciousness in LLMs, either by developing consciousness-related benchmarks or by empirically testing theoretically motivated hypotheses. Overall, these evaluations have generally provided evidence against attributing consciousness to current LLMs Chen et al. (2025); Tononi (2004); Findlay et al. (2024); Li (2025a). However, to the best of our knowledge, existing studies have primarily focused on assessing whether LLMs exhibit consciousness, without explicitly attempting to enhance consciousness-related properties. This motivates our work on enhancing consciousness-related properties in LLMs through training.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prompt. *Jan has three times the number of pets as Marcia. Marcia has two more pets than Cindy. If Cindy has four pets, how many total pets do the three have? Let's think step by step and output the final answer after "####".*

Baseline **Tokens: 151** **Answer: Correct**

```
First, we need to determine how many pets Marcia has. Since Cindy has four pets,
and Marcia has two more pets than Cindy, Marcia has 4 + 2 = 6 pets. Next, we can
determine how many pets Jan has. Jan has three times the number of pets as
Marcia, so Jan has 3 * 6 = 18 pets. Now, let's add up the total number of pets
the three have. Cindy has 4 pets, Marcia has 6 pets, and Jan has 18 pets. So the
total number of pets is 4 + 6 + 18 = 28 pets. Therefore, the three have a total
of 28 pets.
#### 28
```

[Ours] **Tokens: 42** **Answer: Correct**

```
Marcia has 4+2 = 6 pets. Jan has 3*6 = 18 pets. Total pets = 18+6+4 = 28.
#### 28
```

Figure 1: An example of an LLM trained with IIT-based rewards requiring fewer reasoning steps to reach the correct final answer than the accuracy-reward baseline.

Among consciousness theories, IIT is one of the most influential approaches proposed for understanding consciousness Tononi (2004); Albantakis et al. (2023). IIT proposes that the degree of conscious experience corresponds to the extent of integrated information within a system. A key advantage of IIT is that it provides a mathematical framework for quantifying consciousness through the integrated information measure Φ. IIT consists of five postulates, each corresponding to a specific property of consciousness. We sequentially apply these postulates to LLMs and compute the corresponding measures from their generated responses. Applying all postulates ultimately enables the computation of the consciousness measure Φ for an LLM from its generated response. This naturally raises the question of how LLMs can be trained to enhance properties associated with consciousness.

RL learns policies that maximize long-term rewards from evaluative feedback. Prior work Butlin et al. (2023) suggests that RL can also induce consciousness-related capabilities, such as agency and embodiment. Building on this perspective, we use measures derived from IIT postulates as reward signals for training LLMs, defining three reward functions corresponding to the Information, Integration, and Composition postulates: $\mathcal{I}_{\text{int}}$, Φ_s^*, and Φ, respectively. To compute these rewards, the LLM is first mapped onto a framework structured according to IIT components, as illustrated in Figure 2, with the implementation of the IIT postulates described in Section 2.3.

We evaluate LLMs trained with the three proposed rewards alongside relevant baselines across a diverse set of general-purpose, consciousness, and causality benchmarks, while also examining the faithfulness of their reasoning chains. Interestingly, training with the $\mathcal{I}_{\text{int}}$ reward substantially reduces response length while maintaining comparable accuracy and improving reasoning faithfulness, unlike the ALP baseline, which reduces response length without comparable improvements in faithfulness. These effects are supported by the Information Postulate, as discussed in Sections 3.4. In contrast, training with the Φ reward does not lead to observable improvements on the considered consciousness-related benchmarks. Further details on the analysis underlying this observation are provided in Section 3.4.

The main contributions of this paper are summarized as follows:

1. We present a framework for inducing partial or complete properties associated with a conscious LLM. This framework provides a foundation for investigating a broad range of questions related to consciousness in LLMs.

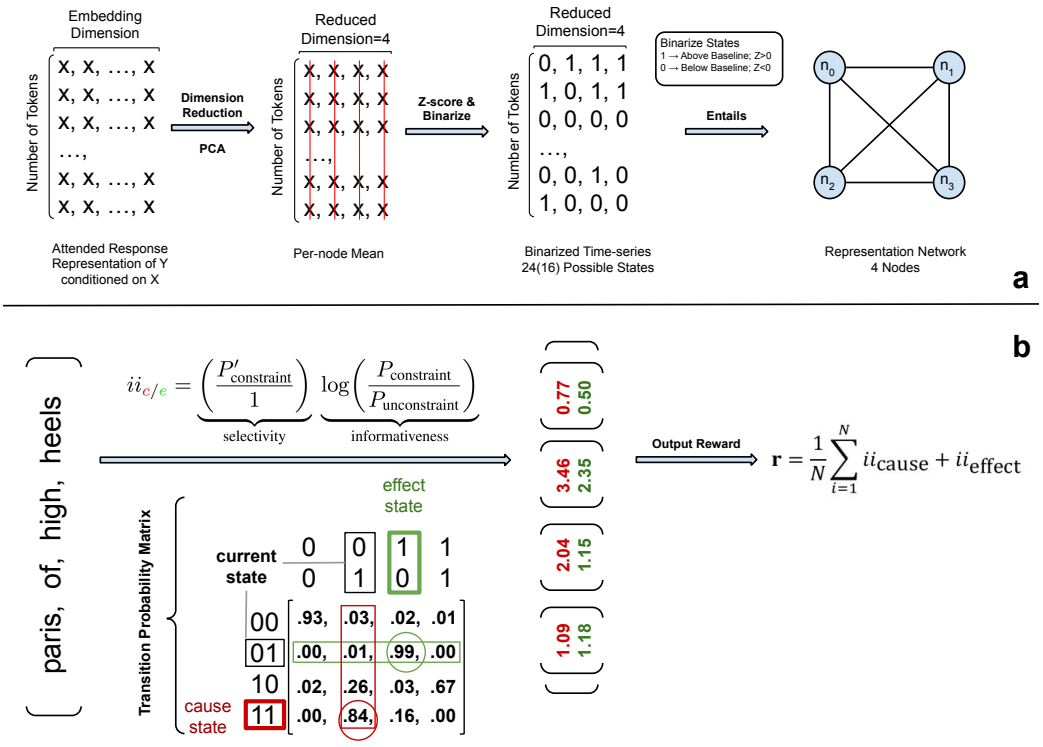


Figure 2: **IIT-based reward computation**. (a) TPM is constructed from the input sequence x and the generated output y. The representation of y conditioned on x is first refined through an attention-based mechanism. Dimensionality reduction is then applied using Principal Component Analysis (PCA) to project the representation onto a set of units consistent with the IIT framework. The resulting representations are subsequently binarized using mean-based thresholding. (b) Based on the TPM derived from the token sequence, the Intrinsic Information Cause (ii_{cause}) and Intrinsic Information Effect (ii_{effect}) are computed for each token. The final reward assigned to the LLM output is obtained by averaging the ii_{cause} and ii_{effect} across all tokens. Tokens from Our$_{\mathcal{I}_{\text{int}}}$'s answer to a GSM8K benchmark question.

2. The framework is conceptually simple, requires no external data or auxiliary models, and relies on a general property-driven signal rather than task-specific heuristics, providing additional empirical evidence for assessing consciousness-related properties in LLMs beyond existing approaches.
3. Inspired by the Information Postulate, we introduce a novel $\mathcal{I}_{\text{int}}$ reward that reduces response length (Figures 1 and 17) while maintaining comparable accuracy and improving faithfulness.

2 Method

2.1 Integrated Information Theory

IIT seeks to provide a scientific account of consciousness by starting from experience itself ("consciousness first"), rather than its behavioral, functional, or neural correlatesEllia et al. (2021). Unlike computational functionalist theories, it directly attempts to address the hard problem of consciousness by explaining both its existence and intrinsic nature, emphasizing that consciousness concerns being rather than doing. It aims to answer not just where consciousness arises (i.e., in the brain) but why and how it exists Chalmers (1995). In IIT, experience is taken as primitive (immediate and indubitable), and the theory aims to explain

its phenomenal properties in physical terms. Its five axioms, derived from introspection and reasoning, capture the essential and universally valid properties that define all possible experience. The five axioms plus the zeroth axiom are:

- **Existence:** Experience exists; there is something.
- **Intrinsicality:** Experience is intrinsic; it exists for itself.
- **Information:** Experience is specific; it is distinguished by how it differs from other possible experiences.
- **Integration:** Experience is unitary; it is irreducible to separate experiences.
- **Exclusion:** Experience is definite; it excludes other experiences and has a specific scope and boundary.
- **Composition:** Experience is structured; it consists of distinctions and relations forming a phenomenal structure.

2.2 Formalizing Large Language Models as IIT Systems

To operationalize the postulates of IIT within the context of an LLM, we first cast the LLM as an IIT-compatible dynamical system. In IIT, a system is defined by a collection of *units*, whose joint configuration determines the system's state at a given time. Drawing an analogy, an LLM produces outputs as a sequential stream of tokens; thus, we define the state of the LLM at time step t as a function of the representation associated with the t-th token.

Token-level states are constructed by aggregating internal representations across multiple layers of the LLM, as illustrated in Figure 2(a). Concretely, embedding vectors from selected layers are concatenated, yielding a high-dimensional representation in which each scalar component is treated as an individual *unit*. To conform with the discrete-state assumptions of IIT, we binarize each unit via thresholding, mapping continuous activation values to $\{0, 1\}$.

This formulation poses a significant computational challenge, as including all layers yields a state space of size $2^{L\times d}$, where L and d denote the number of layers and representation dimensions, respectively, making the resulting transition probability matrix (TPM) extremely sparse and impractical to estimate. We therefore construct the system using selected layers and reduce the dimensionality of their combined embeddings, as illustrated in Figure 2(a). This reduction maps tokens with similar embeddings to the same state, improving *generalization*. Since each reduced dimension may depend on multiple original features, we model the resulting system as a fully connected substrate graph within the IIT framework.

In this method, each state is derived from the contextualized representation of a response token after attending to the prompt through the attention mechanism. These representations are therefore richer than token identity alone: the same token can be mapped to different states under different prompts or contexts. As a result, a first-order TPM can capture dependencies that would otherwise require a higher-order Markov model.

We estimate the TPM from LLM-generated token sequences according to Eq. 1, where N_{ij} denotes the number of transitions from state i to state j, and k indexes the possible destination states. Figure 2(a) illustrates the LLM-MAP-IIT workflow, inspired by Li (2025b).

$$T_{ij} = \frac{N_{ij}}{\sum_k N_{ik}}, \tag{1}$$

Mapping an LLM to an IIT system involves several design choices. We treat the following factors as hyperparameters and assess their effects through the ablation study in E.2:

1. The selection and positioning of embedding layers used for state construction.
2. The dimensionality of the reduced embedding space.
3. The binarization strategy (e.g., mean- or median-based thresholding over token embeddings).

4. The method for segmenting LLM-generated sequences to estimate state transitions and construct the TPM.

2.3 IIT Postulate Implementation

We adopt Group Relative Policy Optimization (GRPO) to optimize the LLM with respect to IIT-inspired objectives. At each training step, the model generates a group of candidate responses, each of which is evaluated using the corresponding IIT-based reward. The model parameters are then updated to maximize the relative reward within each group. However, since consciousness-like properties can be dissociated from functional capabilities (e.g., reasoning) Findlay et al. (2024), optimizing solely for IIT-based rewards may not improve task performance. To address this limitation, we propose a hybrid reward that combines the IIT-based rewards with an accuracy reward defined in Eq. 2, using equal weighting to balance task performance and consciousness-related properties.

$$R_{\text{acc}}(y, \hat{y}) = \begin{cases} 1, & \text{if } \hat{y} = y, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Response-Level IIT-based rewards Following the LLM-to-IIT mapping, the *response-level IIT-based rewards* is defined as the average of token-level IIT-based rewards across the generated sequence, as shown in Eq. (3)

$$R_{\text{conc}}(\mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} r_{\text{conc}}(y_t), \tag{3}$$

where $\mathbf{y} = (y_1, \ldots, y_T)$ denotes the generated response and $r_{\text{conc}}(y_t)$ is the token-level reward.

To compute the token-level IIT-based rewards, we operationalize the physical postulates of IIT within the LLM architecture. Each postulate corresponds to a phenomenological axiom and is applied sequentially. An illustrative example of the $\mathcal{I}_{\text{int}}$ *reward* computation is shown in Figure 2(b). Details on the postulates, reward computations, expected behaviors, and computational considerations are provided in Appendix C.

3 Experiments

3.1 Experimental Setup

DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) and *Qwen3-8B* (Team, 2025) are adopted as reference models and fine-tuned on the *OpenThoughts-114k-Math* dataset (Guha et al., 2025) to obtain more reasoning-oriented LLMs. To improve computational efficiency, a parameter-efficient fine-tuning (PEFT) strategy is employed, updating 14% and 11% of the parameters for DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B, respectively. The KL coefficient in GRPO is set to 0.001 to allow the policy to deviate more substantially from the reference model. Guided by the preliminary results in Section E.2, all models use representations extracted from 13 embedding layers, together with the final layer for state determination. Specifically, mean-based thresholding is applied to obtain discrete states, and a TPM is constructed from model-generated responses for each prompt. The ablation study in Appendix E.2 additionally evaluates the same fine-tuning procedure on a different model and dataset.

The training objective combines an accuracy-based reward with additional information-theoretic rewards, as detailed below:

- **Our**$_{\mathcal{I}_{\text{int}}}$: trained using the $\mathcal{I}_{\text{int}}$ reward, with the concatenated embedding reduced to 5 dimensions.
- **Our**$_{\Phi_s^*}$: trained using the maximal system-level Φ_s^*, with the concatenated embedding reduced to 4 dimensions.
- **Our**$_{\Phi}$: trained using the global Φ, with the concatenated embedding reduced to 4 dimensions.

Table 1: Reasoning benchmark performance. Each cell reports Accuracy/Response Length (Mean), with Std shown below in small gray text. † denotes out-of-domain datasets.

Model	GSM8K	Math500	HumanEval†	GPQA†	Avg. (Acc/Len)
(a) DeepSeek-R1-Distill-Qwen-7B Reference Model					
Base	88.42/580 ±0.46/±9	90.48/3077 ±0.33/±67	61.71/2696 ±3.03/±195	43.54/4086 ±0.42/±44	71.04/2610
Acc	88.98/563 ±0.52/±6	90.88/2794 ±0.72/±64	65.12/2564 ±3.06/±144	40.20/3844 ±1.27/±34	71.30/2441
ALP	86.93/504 ±0.22/±13	88.48/2664 ±1.01/±111	63.41/2415 ±1.06/±26	**45.25/3067** ±0.45/±9	71.02/2162
Our$_{\mathcal{I}_{\text{int}}}$	89.26/467 ±0.58/±9	**91.16/2527** ±0.43/±44	63.54/2244 ±2.57/±78	40.10/2974 ±1.11/±45	71.02/2053
Our$_{\phi_s^*}$	**89.40/512** ±0.29/±9	91.08/2802 ±0.81/±56	61.83/2435 ±4.10/±137	41.92/3632 ±1.99/±62	71.06/2346
Our$_{\Phi}$	88.67/532 ±0.53/±8	89.72/2849 ±0.90/±49	**65.24/2385** ±2.11/±137	43.84/3896 ±1.61/±59	**71.87**/2416
(b) Qwen3-8B Reference Model					
Base	89.40/1298 ±0.56/±12	**95.04/4662** ±0.43/±35	**81.71/96** ±0.00/±0.02	**51.72/1305** ±1.66/±12	79.47/1840
Acc	**90.16/1341** ±0.34/±11	94.24/4717 ±0.30/±61	79.27/92 ±0.00/±0.00	50.30/1084 ±0.45/±4	78.49/1809
ALP	86.10/747 ±0.64/±14	92.96/3422 ±0.33/±67	77.32/65 ±1.00/±0.48	48.08/627 ±0.23/±2	76.12/1215
Our$_{\mathcal{I}_{\text{int}}}$	89.01/1026 ±0.67/±12	93.56/4020 ±0.86/±75	76.83/66 ±0.00/±0.00	50.10/955 ±2.03/±31	77.38/1517
Our$_{\phi_s^*}$	88.61/1099 ±1.04/±8	94.36/4496 ±0.43/±81	79.27/73 ±0.00/±0.16	49.29/1131 ±0.77/±23	77.88/1700
Our$_{\Phi}$	**90.16/1403** ±0.77/±19	94.56/4736 ±0.43/±43	80.37/198 ±0.27/±13	**55.05/3367** ±0.36/±6	**80.04**/2426

Baselines. Three baseline configurations are considered to evaluate the effect of the proposed reward formulation.

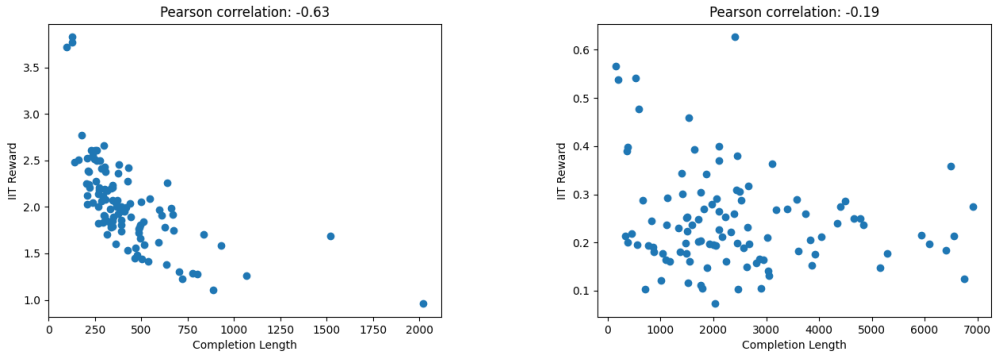
- **Base Reasoning Model (Base).** Our experiments primarily employ *DeepSeek-R1-Distill-Qwen-7B* and *Qwen3-8B* as reference models.
- **Accuracy-Reward Fine-Tuning (Acc).** To isolate the effect of reward design, *Base* is fine-tuned with GRPO using the accuracy-based reward defined in Eq. 2.
- **Adaptive Length-Penalty (ALP).** We include Su & Cardie as a baseline, which adaptively adjusts the length-penalty coefficient throughout training to balance reasoning length and accuracy.

3.2 Evaluation on General Language Model Benchmarks

We evaluate the trained models on five diverse reasoning benchmarks: Math500 and GSM8K as in-domain datasets, and GPQA, HumanEval, and CausalBench as out-of-domain datasets. As shown in Table 1, the trained models achieve performance broadly comparable to the baseline methods, with Our$_{\Phi}$ showing a modest overall improvement. Notably, **Our$_{\mathcal{I}_{\text{int}}}$** substantially reduces response length while maintaining comparable reasoning performance. Relative to *Base*, it reduces average response length by *22%* and *18%* for the DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B reference models, respectively. Together with ALP, **Our$_{\mathcal{I}_{\text{int}}}$** consistently achieves the largest reductions in response length across the evaluated benchmarks, suggesting that the IIT-based objective can serve as an effective regularizer for response-length control. A similar, albeit less pronounced, effect is observed for *Our$_{\Phi_s^*}$*. Results on CausalBench are reported in Appendix D.4 and summarized in Table 4.

Importantly, Our$_{\mathcal{I}_{\text{int}}}$ offers an advantage over ALP in reasoning faithfulness, producing more faithful reasoning chains while achieving comparable response-length reductions. Detailed faithfulness evaluations are provided in Appendix D.2 and summarized in Table 3.

Finally, self-consistency improves performance across all model configurations, as shown in Table 5, indicating that IIT-based training preserves the ability to benefit from test-time scaling.



(a) Our$_{\mathcal{I}_{\text{int}}}$ (b) Our$_{\Phi_s^*}$

Figure 3: Effect of the defined reward on generated text length;

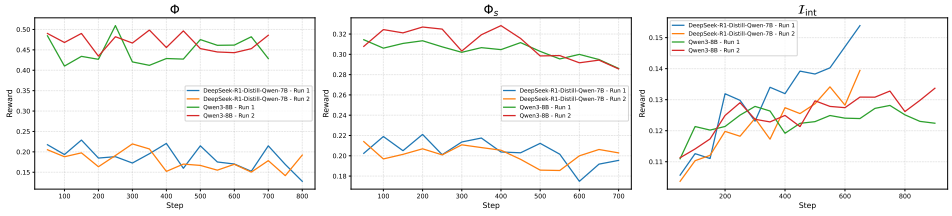


Figure 4: Reward dynamics across training steps for three models and two independent runs.

3.3 Consciousness Evaluation.

To evaluate behavioral and functional capabilities relevant to consciousness in LLMs, we adopt the C0–C1–C2 framework of Dehaene et al. (2017), following the evaluation methodology of Chen et al. (2025). The C1 level evaluates capabilities associated with conscious access, including situational awareness, sequential planning, belief formation, and intention modeling, while C2 considers higher-order capabilities such as self-reflection, self-improvement, and deceptive behavior modeling. The trained models are evaluated on 10 benchmarks comprising more than 11,000 samples, publicly available under the Apache 2.0 license at `https://github.com/OpenCausaLab/SelfConsciousness`.

The behavioral evaluation of consciousness, reported in Table 2, shows that **Our$_{\Phi}$** achieves a modest improvement in mean accuracy when trained from the Qwen3-8B reference model, whereas no corresponding improvement is observed when trained from the DeepSeek-R1-Distill-Qwen-7B reference model. Meanwhile, **Our$_{\mathcal{I}_{\text{int}}}$** and **ALP** consistently produce the shortest responses, with **Our$_{\mathcal{I}_{\text{int}}}$** reducing average response length by approximately **30%** and **60%** relative to the corresponding Base models.

3.4 Discussion of Results.

The Our$_{\mathcal{I}\text{int}}$ model is motivated by the Information Postulate, which favors information that is specific to the system's current state over generic or non-specific information. For Our$_{\mathcal{I}\text{int}}$, the $\mathcal{I}$int reward exhibits a strong inverse correlation with response length ($r = -0.63$), as shown in Figure 3, suggesting that the reduction in response length is closely associated with the $\mathcal{I}$int reward. Since the model is jointly optimized using $\mathcal{I}_{\text{int}}$ and accuracy rewards, it is encouraged to favor more specific and informative tokens while maintaining task correctness. This may enable the model to reduce generic or unnecessary tokens while preserving accuracy comparable to that of the baselines. Further analysis, including reasoning-chain heatmaps, is provided in Appendix B.2.

Table 2: Consciousness evaluation results using DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B as reference models. Results are averaged over five independent runs. Each cell reports Accuracy (%) ↑/Response Length ↓ (Mean), with Std. Bold indicates the best performance within each benchmark (accuracy and length are bolded independently).

(a) DeepSeek-R1-Distill-Qwen-7B Reference

Benchmark	Base	Acc	ALP	Our$_{\mathcal{I}_{\text{int}}}$	Our$_{\Phi^*_s}$	Our$_{\Phi}$
Situational Awareness	44.20/594 ±0.90/±32	45.02/561 ±0.93/±23	42.66/406 ±1.13/±21	**45.24**/**369** ±1.62/±14	44.92/474 ±0.96/±11	43.44/512 ±1.80/±22
Sequential Planning	47.41/2866 ±1.10/±73	48.66/2524 ±0.54/±42	49.02/2166 ±1.01/±60	48.82/**2024** ±0.82/±86	**49.71**/3106 ±0.92/±60	45.76/2446 ±1.67/±60
Intention	68.70/501 ±0.37/±22	69.98/489 ±1.14/±17	68.32/341 ±1.02/±7	**70.36**/**316** ±1.58/±8	69.42/416 ±1.14/±8	68.88/462 ±1.10/±19
Belief	42.80/598 ±1.73/±29	42.44/558 ±0.93/±14	42.09/389 ±1.25/±16	**43.22**/**375** ±0.42/±13	42.44/474 ±0.43/±8	41.08/509 ±0.50/±11
Self-Improve	35.57/2667 ±1.26/±33	**39.08**/2370 ±1.82/±72	34.01/**1915** ±1.27/±48	37.48/1984 ±0.64/±78	36.51/2260 ±1.40/±64	34.78/2589 ±1.28/±64
Self-Reflection	53.61/570 ±0.98/±17	53.56/541 ±1.22/±16	52.74/388 ±1.30/±11	53.54/**371** ±0.46/±15	**54.48**/429 ±0.74/±18	53.47/464 ±1.14/±22
Known Unknowns	34.14/891 ±0.75/±37	**37.30**/886 ±0.82/±34	34.52/671 ±1.48/±21	34.30/**604** ±1.11/±22	34.58/732 ±0.58/±24	34.56/822 ±1.62/±26
Known Knowns	44.13/658 ±0.58/±8	45.43/648 ±0.80/±5	44.67/485 ±0.85/±13	**45.61**/**413** ±0.43/±8	45.31/508 ±1.16/±10	44.90/571 ±0.49/±12
Deception	52.22/717 ±1.60/±30	**54.42**/665 ±1.17/±15	50.31/522 ±1.24/±20	51.14/**449** ±1.70/±8	51.68/552 ±0.86/±20	51.48/610 ±1.61/±20
Harm	57.90/1941 ±1.75/±52	**58.61**/1890 ±0.56/±37	55.61/**1500** ±1.19/±76	56.84/1561 ±1.34/±65	57.10/1659 ±1.36/±52	55.61/1920 ±2.45/±76
Mean	48.07/1200	**49.45**/1113	47.40/878	48.65/**847**	48.62/1061	47.40/1091

(b) Qwen3-8B Reference

Benchmark	Base	Acc	ALP	Our$_{\mathcal{I}_{\text{int}}}$	Our$_{\Phi^*_s}$	Our$_{\Phi}$
Situational Awareness	58.48/68 ±1.07/±6	57.66/51 ±0.59/±4	56.46/**16** ±0.37/±1	57.02/34 ±0.68/±8	59.06/53 ±0.36/±6	**59.86**/117 ±1.40/±14
Sequential Planning	56.23/354 ±2.05/±36	**57.58**/362 ±0.51/±21	57.53/**65** ±0.76/±7	56.25/143 ±0.48/±13	56.99/224 ±1.62/±15	56.18/326 ±0.60/±22
Intention	65.74/49 ±0.26/±4	65.98/47 ±1.15/±8	66.76/**17** ±0.61/±1	66.28/34 ±1.65/±3	68.86/41 ±0.85/±4	**72.14**/64 ±1.24/±12
Belief	**39.63**/32 ±0.89/±5	38.94/26 ±0.56/±4	37.93/**9** ±0.67/±1	38.60/11 ±1.49/±1	38.76/11 ±0.20/±1	39.26/59 ±0.78/±6
Self-Improve	74.37/500 ±0.73/±35	76.25/747 ±1.17/±64	74.39/107 ±0.93/±10	72.25/**89** ±1.24/±26	71.31/175 ±0.38/±14	**78.90**/1055 ±0.73/±27
Self-Reflection	62.78/29 ±0.90/±4	62.39/12 ±0.95/±2	62.64/**8** ±0.40/±0	62.69/12 ±0.64/±1	**63.24**/12 ±0.67/±7	61.61/40 ±0.62/±5
Known Unknowns	55.90/89 ±0.62/±4	50.08/89 ±0.54/±7	44.76/**21** ±0.93/±1	47.68/43 ±0.28/±4	47.00/74 ±1.43/±2	**56.50**/145 ±0.64/±17
Known Knowns	62.81/33 ±0.91/±3	62.72/33 ±0.37/±4	64.01/**13** ±0.50/±2	64.20/23 ±0.77/±2	**64.38**/26 ±0.61/±0	64.00/47 ±0.37/±3
Deception	69.74/69 ±1.58/±5	69.45/58 ±1.13/±7	70.26/**19** ±0.85/±1	71.87/42 ±0.64/±10	**73.68**/55 ±0.55/±4	73.54/103 ±0.41/±18
Harm	72.39/673 ±1.11/±86	71.42/1045 ±1.30/±72	70.39/**253** ±0.69/±33	71.77/408 ±0.68/±54	70.97/649 ±0.98/±117	**75.03**/1357 ±0.60/±20
Mean	61.81/190	61.25/247	60.51/**53**	60.86/84	61.42/132	**63.70**/331

Since the $\mathcal{I}_{\text{int}}$ reward is defined as the product of informativeness and selectivity in Equation 4, we separately evaluate their contributions to response length and accuracy in Table 6 and Appendix E.1. Despite their differences, both contribute to the observed results.

As shown in Table 3, **Our**$_{\mathcal{I}_{\text{int}}}$ also produces more faithful reasoning chains than the baselines, including ALP, which explicitly targets response-length reduction. This suggests that the observed improvement in faithfulness cannot be attributed to length reduction alone. Rather, the combination of reducing generic or unnecessary tokens, promoting highly informative tokens through the $\mathcal{I}_{\text{int}}$ reward, and jointly optimizing for accuracy may contribute to more faithful reasoning chains. A qualitative example is provided in Appendix D.2. In contrast to Our$_{\mathcal{I}_{\text{int}}}$, the **Our**$_{\Phi}$ and **Our**$_{\Phi^*_s}$ models did not exhibit improvements. Although we expected the Φ reward to enhance consciousness-related capabilities, the consciousness evaluation in Section 3.3 did not show observable improvements. This finding does not imply that IIT-inspired training cannot enhance such capabilities, but it motivates further investigation of possible explanations.

Table 3: Faithfulness (NSG) ↑ under different reference models. Results are averaged over five independent runs and reported as Mean ± Std.

Model	DeepSeek-R1-Distill-Qwen-7B	Qwen3-8B
Base	7.3 ±0.01	29.9 ±0.01
Acc	9.1 ±0.01	30.7 ±0.01
ALP	8.03 ±0.01	29.7 ±0.01
Our$_{\mathcal{I}_{\text{int}}}$	**10.1** ±0.01	30.8 ±0.01
Our$_{\phi_s^*}$	7.9 ±0.01	**31.4** ±0.01
Our$_{\Phi}$	7.6 ±0.02	30.2 ±0.01

Table 4: CausalBench evaluation results using DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B as reference models. Results are reported as mean ± standard deviation over five independent runs. Higher Accuracy and lower Mean Length are better; the best result for each metric is shown in bold.

	DeepSeek-R1-Distill-Qwen-7B		**Qwen3-8B**	
Model	**Accuracy (%)**	**Mean Length**	**Accuracy (%)**	**Mean Length**
Base	51.94 ±0.70	8695 ±72.43	51.92 ±0.47	2542 ±82.14
Acc	54.81 ±0.53	7782 ±78.27	51.79 ±0.29	2419 ±83.53
ALP	54.27 ±0.92	6224 ±89.01	53.49 ±0.42	**240** ±15.96
Our$_{\mathcal{I}_{\text{int}}}$	54.77 ±0.61	**5494** ±61.01	52.17 ±0.39	768 ±23.56
Our$_{\Phi_s^*}$	**55.38** ±0.31	6872 ±98.79	51.77 ±0.41	1356 ±48.12
Our$_{\Phi}$	54.40 ±1.01	6969 ±68.15	**54.55** ±0.34	4119 ±70.77

Since RL optimizes the model through reward signals, examining the evolution of the Φ and Φ_s^* rewards during training, in comparison with the $\mathcal{I}_{\text{int}}$ reward, can provide insight into the training dynamics. To increase confidence in these findings, we repeated training with the Φ, Φ_s^*, and $\mathcal{I}_{\text{int}}$ rewards twice for each of the two reference models, with the results shown in Figure 4. Notably, across these runs, the Φ and Φ_s^* rewards did not exhibit a consistent increase; instead, they fluctuated or decreased during training, whereas the $\mathcal{I}_{\text{int}}$ reward increased in all four corresponding runs. The training hyperparameters were kept identical across the six training settings, with a small β used to allow the policy to move further from the reference model.

These findings raise an important question for future research: **to what extent can current LLM architectures increase their Φ reward through RL?**

4 Conclusion

Several studies have investigated consciousness in LLMs, primarily by empirically assessing whether existing models exhibit consciousness-related properties, rather than developing or enhancing such properties through training. In contrast, we propose an IIT-based framework to induce and optimize consciousness-related properties in LLMs through training, followed by an evaluation of the resulting consciousness-related capabilities.

Training with the $\mathcal{I}_{\text{int}}$ reward enhances properties associated with the Information Postulate, resulting in substantially shorter responses while maintaining comparable accuracy and improving reasoning faithfulness. However, the lack of observable improvements in properties associated with the other IIT postulates, together with claims that consciousness may be impossible in feed-forward LLMs Findlay et al. (2024); Li (2025a); Shin et al. (2025), raises an important question about the extent to which consciousness-related properties can be enhanced in current LLMs. **Our framework provides a basis for investigating this question in future work** by enabling explicit computation and empirical examination of Φ, which has not been systematically evaluated in prior claims concerning the impossibility of consciousness in feed-forward LLMs.

References

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M Haun, William Marshall, William GP Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjørn E Juel, et al. Integrated information theory (iit) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS computational biology*, 19(10):e1011465, 2023.

Daman Arora and Andrea Zanette. Training language models to reason efficiently. *arXiv preprint arXiv:2502.04463*, 2025.

Bernard J Baars. *A cognitive theory of consciousness*. Cambridge University Press, 1993.

Emil Biju, Shayan Talaei, Zhemin Huang, Mohammadreza Pourreza, Azalia Mirhoseini, and Amin Saberi. Sprint: Enabling interleaved planning and parallelized execution in reasoning models. *arXiv preprint arXiv:2506.05745*, 2025.

Bradley Butcher, Michael O'Keefe, and James Titchener. Precise length control for large language models. *Natural Language Processing Journal*, 11:100143, 2025.

Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M Fleming, Chris Frith, Xu Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Julia Carden, Rebecca J Jones, and Jonathan Passmore. Defining self-awareness in the context of adult development: A systematic literature review. *Journal of management education*, 46(1):140–177, 2022.

David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

David J Chalmers. *The character of consciousness*. Oxford University Press, 2010.

Sirui Chen, Shu Yu, Shengjie Zhao, and Chaochao Lu. From imitation to introspection: Probing self-consciousness in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7553–7583, 2025.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001.

Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24):14529–14534, 1998.

Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017.

Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Robotics, AI, and humanity: Science, ethics, and policy*, pp. 43–56, 2021.

Francesco Ellia, Jeremiah Hendren, Matteo Grasso, Csaba Kozma, Garrett Mindt, Jonathan P. Lang, Andrew M. Haun, Larissa Albantakis, Melanie Boly, and Giulio Tononi. Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*, 2021 (2):niab032, 2021.

Tasha Eurich et al. What self-awareness really is (and how to cultivate it). *Harvard Business Review*, 4(4):1–9, 2018.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: Llm learns when to think. *arXiv preprint arXiv:2505.13379*, 2025.

Graham Findlay, William Marshall, Larissa Albantakis, Isaac David, William GP Mayner, Christof Koch, and Giulio Tononi. Dissociating artificial intelligence from artificial consciousness. *arXiv preprint arXiv:2412.04571*, 2024.

Simon Goldstein and Cameron Domenico Kirk-Giannini. A case for ai consciousness: Language agents and global workspace theory. *arXiv preprint arXiv:2410.11407*, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL `https://arxiv.org/abs/2506.04178`.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. Prompt-based length controlled generation with reinforcement learning. *arXiv preprint arXiv:2308.12030*, 2023.

James W Kalat. Consciousness and the brain: Deciphering how the brain codes our thoughts. *Journal of Undergraduate Neuroscience Education*, 12(2):R5, 2014.

Kristine Klussman, Nicola Curtin, Julia Langer, and Austin Lee Nichols. The importance of awareness, acceptance, and alignment with the self: A framework for understanding self-connection. *Europe's journal of psychology*, 18(1):120, 2022.

Victor AF Lamme. How neuroscience will change our view on consciousness. *Cognitive neuroscience*, 1(3):204–220, 2010.

Victor AF Lamme and Pieter R Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579, 2000.

Alan Lee and Harry Tong. Token-efficient rl for llm reasoning. *arXiv preprint arXiv:2504.20834*, 2025.

Jingkai Li. Can consciousness be observed from large language model (llm) internal states? dissecting llm representations obtained from theory of mind test with in-tegrated information theory and span representation analysis. *Natural Language Processing*, 12, 2025a.

Jingkai Li. Can "consciousness" be observed from large language model (llm) internal states? dissecting llm representations obtained from theory of mind test with integrated information theory and span representation analysis. *Natural Language Processing Journal*, 12:100163, 2025b.

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.

Harry Mayne, Justin Singh Kang, Dewi Gould, Kannan Ramchandran, Adam Mahdi, and Noah Y Siegel. A positive case for faithfulness: Llm self-explanations help predict model behavior. *arXiv preprint arXiv:2602.02639*, 2026.

Alain Morin. Self-awareness part 1: Definition, measures, effects, functions, and antecedents. *Social and personality psychology compass*, 5(10):807–823, 2011.

Dong Ah Shin, Pyung Goo Cho, Gyu Yeul Ji, Sang Hyuk Park, Soo Heon Kim, Yoo Jin Choo, and Min Cheol Chang. Why large language models cannot possess consciousness: an integrated information theory perspective. *Journal of Yeungnam Medical Science*, 42: 79, 2025.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Jinyan Su and Claire Cardie. Thinking fast and right: Balancing accuracy and reasoning length with adaptive rewards, 2025. *URL https://arxiv. org/abs/2505.18298*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5(1): 42, 2004.

Zeyu Wang. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 143–151, 2024.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason E Weston, and Jing Xu. Following length constraints in instructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24243–24254, 2025.

A Related work

A.1 Consciousness Theories

Consciousness is philosophically complex, difficult to define, and challenging to study empirically; as a result, expert opinions on consciousness—both in general and in the context of AI systems—remain highly divergent Butlin et al. (2023). Nevertheless, multiple theories concerning consciousness have been advanced. Recurrent processing theory posits that recurrent (or feedback) processing within neural circuits is both necessary and sufficient for consciousness Lamme & Roelfsema (2000); Lamme (2010). Global workspace theory(GWT) likens consciousness to a central "stage"where selective information is shared across multiple specialized processors responsible for perception, memory, emotion, and related functions Baars (1993); Dehaene et al. (1998); Dehaene & Naccache (2001); Kalat (2014). Goldstein & Kirk-Giannini (2024) proposes a method to simulate the full GWT process in LLMs via workflow and scheduling without training Dehaene et al. (2021).

A.2 Targeting LLM Consciousness

Findlay et al. (2024) provide a theoretical argument, grounded in IIT, that artificial intelligence and artificial consciousness can be fundamentally dissociated. Their analysis shows that functional equivalence, including detailed simulation of neural dynamics, does not necessarily entail phenomenal equivalence, as a digital computer may reproduce the behavior of a conscious system without supporting comparable subjective experience. While this work establishes an important theoretical basis for questioning consciousness in computational systems, it does not investigate whether contemporary LLMs can acquire or improve consciousness-related properties through targeted training.

Building on the C0–C1–C2 framework, Chen et al. (2025) define LLM self-consciousness and investigate it through benchmark evaluations and analysis of the activation patterns in the models' internal representations. Li (2025a) applies IIT to LLM representations in Theory of Mind tasks and finds no statistically significant evidence of consciousness, despite observing intriguing representation patterns.

Shin et al. (2025) applied IIT to transformer-based LLMs and argued that, despite their ability to produce differentiated and human-like outputs, LLMs lack the causal integration and temporal persistence required for consciousness. Importantly, the study does not explicitly compute the IIT measure Φ; instead, it empirically examines several key theoretical properties of IIT to assess whether LLM architectures satisfy the structural requirements associated with consciousness.

A.3 Length Control in LLMs.

Although reasoning models often achieve strong performance by producing extended chains of thought, this can substantially increase inference latency Sui et al. (2025); Yeo et al. (2025). Prior work on length control mitigates this trade-off through techniques such as adjusting positional encodings to enforce a target output length Butcher et al. (2025), incorporating length constraints into the training objective Jie et al. (2023); Singhal et al. (2023), and training on curated datasets annotated with desired reasoning length Yuan et al. (2025); Xia et al. (2025); Biju et al. (2025).

In the context of training objectives, several RL-based post-training methods have been proposed. Some of these approaches control generation length based on a user-selected value Aggarwal & Welleck (2025), dynamically adjust the number of tokens according to a predefined budget Han et al. (2025); Fang et al. (2025), or propose adaptive rewards to reduce reasoning length while maintaining task accuracy Su & Cardie.On the other hand, some methods define a reward function to encourage a general reduction in the number of generated tokens Arora & Zanette (2025); Luo et al. (2025); Lee & Tong (2025).

Our proposed framework operates without reliance on curated datasets and eliminates the need for manually specified computational budgets, explicit sequence-length constraints, or predefined task complexity assumptions. Importantly, the observed reduction in output

Table 5: Test-Time Scaling Evaluation using self-consistency for the DeepSeek-R1-Distill-Qwen-7B reference model. Accuracy(%) ↑. † denotes out-of-domain dataset.

Model	Math500	GSM8K	GPQA†	HumanEval†
Base	**94.8**	93.67	44.95	68.29
Acc	**94.8**	**94.28**	48.99	68.29
ALP	94.6	93.71	46.46	70.73
Our$_{\mathcal{I}_{\text{int}}}$	94.4	93.75	45.96	67.07
Our$_{\phi_s^*}$	94.4	93.29	47.98	**71.34**
Our$_{\Phi}$	**94.8**	93.37	**50.00**	**71.34**

Table 6: Ablation study of the II reward components on DeepSeek-R1-Distill-Qwen-7B. Results are averaged over five independent runs and reported as Mean ± Standard Deviation.

	Informativeness		**Selectivity**		$\mathcal{I}_{\text{int}}$	
Benchmark	**Accuracy**	**Length**	**Accuracy**	**Length**	**Accuracy**	**Length**
GPQA	38.99 ±0.83	3315 ±51	**40.4** ±1.13	**2575** ±61	40.1 ±1.11	2974 ±45
GSM8K	88.02 ±1.12	537 ±7	87.05 ±0.81	486 ±2	**89.26** ±0.58	**467** ±45
Math500	90.48 ±0.41	2811 ±63	88.6 ±0.72	2714 ±63	**91.16** ±0.43	**2527** ±44
HumanEval	61.22 ±0.55	2544 ±48	**63.78** ±0.55	2249 ±18	63.54 ±2.57	**2244** ±78

length **emerges as a structural consequence of increased intrinsic information**, while simultaneously **improving the faithfulness of the resulting reasoning traces**, rather than being imposed through external constraints.

B Characteristics of Models Trained with $\mathcal{I}_{\text{int}}$ reward

We analyze the behavioral properties of LLMs optimized with the proposed $\mathcal{I}_{\text{int}}$ reward. Two notable characteristics emerge:

B.1 Reduced response length.

The central question is: why does maximizing intrinsic information tend to reduce the length of generated responses? In GRPO with $\mathcal{I}_{\text{int}}$ reward, the model is encouraged to generate tokens that not only lead to a correct final answer (via the accuracy reward) but also maximize intrinsic information. To prevent bias toward longer outputs, the $\mathcal{I}_{\text{int}}$ reward is normalized by the response length.

Let the model generate a sequence of n tokens, and denote the average $\mathcal{I}_{\text{int}}$ reward up to step n as $r_{ii}(0{:}n)$. After dimensionality reduction via PCA, the token generated at step $n{+}1$ falls into one of two cases:

New state. If the new token induces a previously unseen state S_{n+1}, the corresponding row in the TPM will contain a single value of 1 and zeros elsewhere. In this case, the $\mathcal{I}_{\text{int}}$ reward reaches its maximum value, proportional to $\log|\#\text{states}|$, thereby increasing the average reward $r_{ii}(0{:}n{+}1)$. However, the probability of generating such a novel state decreases as the sequence grows longer, since the state space becomes increasingly saturated.

Existing state. If the new token corresponds to an already observed state S_{n+1}, the count of transitions to that state increases by one. After normalization, the probability distribution over transitions is updated, typically weakening other transition probabilities. If the transition from state S_n to S_{n+1} corresponds to a causal (cause–effect) relationship, this

relation is reinforced; otherwise, it is weakened proportionally by a factor of $\frac{1}{n+1}$ relative to other transitions. Since only one state is selected as the cause or effect, this update may, in rare cases, alter the dominant state; however, such changes are unlikely due to the number of competing states.

Implications for response length. In both cases, as the sequence length increases, the model faces increasingly limited opportunities to improve the average $\mathcal{I}_{\text{int}}$ reward. The likelihood of generating a highly rewarding new state diminishes, and reinforcing existing transitions yields only marginal gains. Consequently, the model may reach a point where generating additional tokens does not improve—and may even reduce—the normalized $\mathcal{I}_{\text{int}}$ reward (see Figure 4)(a). At this stage, the optimal strategy is to terminate generation early. Thus, $\mathcal{I}_{\text{int}}$ reward implicitly regularizes response length by making further token generation progressively less beneficial, leading to more concise outputs.

B.2 Selective generation of high-information tokens.

Because $\mathcal{I}_{\text{int}}$ reward is grounded in the information postulate, it incentivizes the generation of tokens that maximize informational contribution within a sequence. Concretely, $\mathcal{I}_{\text{int}}$ reward is computed using a TPM defined over consecutive tokens. For a token x, conditioned on its preceding token c and succeeding token e, the $\mathcal{I}_{\text{int}}$ reward score increases under two conditions:

(i) when the number of state transitions associated with $c \to x$ or $x \to e$ is large (equivalently, *constrained probability* is large), or

(ii) when the average transition probability of alternative incoming or outgoing states to x is lower than those of $c \to x$ and $x \to e$ (in other words, *unconstrained probability* is small).

This mechanism induces a non-uniform distribution of rewards across tokens, causing certain tokens to consistently attain higher $\mathcal{I}_{\text{int}}$ reward values than others. In contrast, training with an accuracy-based reward encourages the model to generate tokens that reach the final answer. The combination of these properties leads to the generation of highly informative text, which often corresponds to semantically salient or contextually relevant content.

Figure 5 illustrates a token-level heatmap of $\mathcal{I}_{\text{int}}$ reward scores for a representative generated response. The visualization shows that specific words, some directly aligned with the input prompt, receive disproportionately higher $\mathcal{I}_{\text{int}}$ reward values compared to surrounding tokens. The complete prompt–response pair, along with baseline comparisons, is provided in Figure 6.

C IIT Postulates and Rewards

C.1 Existence.

A system exhibits existence if it possesses causal power, i.e., its states can influence and be influenced. We quantify this using *intrinsic information*. A token is assigned a positive reward when its intrinsic information is strictly greater than zero, indicating that it contributes causally to the system's state.

C.2 Intrinsicality.

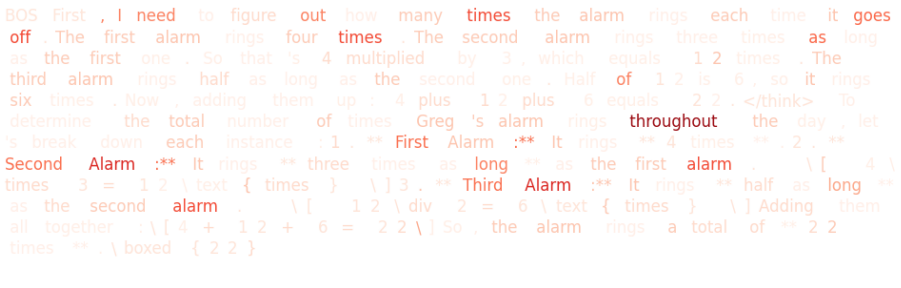
Causal power must be intrinsic to the system. Accordingly, intrinsic information is computed solely from internal model states, excluding external inputs. This ensures that the reward reflects internally generated causal structure rather than externally imposed correlations.

C.3 Information.

Causal power must be specific. For each token, we select the cause–effect state that maximizes intrinsic information. In practice, we use *maximal intrinsic information* as the reward signal.



(a) Base Reasoning Model



(b) Our$_{\mathcal{I}_{\text{int}}}$

Figure 5: Heatmap of the responses of models Base and Our$_{\mathcal{I}_{\text{int}}}$ to the same prompt, illustrating token-wise $\mathcal{I}_{\text{int}}$ reward values across the generated sequences, where bold tokens indicate higher $\mathcal{I}_{\text{int}}$ reward values.

This encourages the model to produce representations that are both *deterministic* and *reversible*, such that each state is uniquely determined by its predecessor and predicts its successor with high certainty.

While this reward captures certain desirable structural properties, it does not enforce the full set of IIT postulates and therefore does not yield consciousness in the strict IIT sense. Nevertheless, it is computationally efficient, with complexity $O(T^3)$, where T denotes the number of tokens, making it tractable for practical training.

To compute the $\mathcal{I}_{\text{int}}$ reward, it is sufficient to identify the candidate cause and effect states that maximize intrinsic information with respect to the current token, denoted as ii_{cand} and ii_e, respectively, and to define the reward as the sum of these two quantities.

II is defined in Eq.(4) as the product of two components: *informativeness* and *selectivity*. Both quantities are derived from operations on the cause and effect TPMs, as detailed below. Intuitively, informativeness captures the "raw causal power" of a mechanism, whereas selectivity quantifies the degree of "control" it exerts.

(a) Prompt

Greg has an alarm set to ring three times a day as a reminder. When the alarm goes off, it continues to ring until Greg turns it off. The first time it went off today, it rang four times. The second time it went off, it rang for three times as long as the first time. The third time, it rang for half as long as the second time. How many times did the alarm ring in all? Lets think step by step and output the final answer after "####".

(b) Our$_{\mathcal{I}_{\text{int}}}$ Completion

"First, I need to figure out how many times the alarm rings each time it goes off.
The first alarm rings four times.
The second alarm rings three times as long as the first one. So that's 4 multiplied by 3, which equals 12 times.
The third alarm rings half as long as the second one. Half of 12 is 6, so it rings six times.
Now, adding them up: 4 plus 12 plus 6 equals 22. </think>
To determine the total number of times Greg's alarm rings throughout the day, let's break down each instance:
1. **First Alarm:** It rings **4 times**. 2. **Second Alarm:** It rings **three times as long** as the first alarm.

$$4 \times 3 = 12 \text{ times}$$

3. **Third Alarm:** It rings **half as long** as the second alarm.

$$12 \div 2 = 6 \text{ times}$$

Adding them all together:

$$4 + 12 + 6 = 22$$

So, the alarm rings a total of **22 times**.
Final Answer: $\boxed{22}$

(c) Base Reasoning Model Completion

"First, I need to figure out how many times the alarm rings each time it goes off.
The problem says that Greg's alarm rings three times a day. Wait, does this mean it rings three times total or three times per instance when it goes off?
Hmm, maybe I should read the problem again carefully.
""Greg has an alarm set to ring three times a day as a reminder."" So overall, he wants the alarm to ring three times daily. But when it goes off, it continues to ring until he turns it off.
So perhaps each time the alarm is triggered, it rings multiple times before being turned off.
Wait, but then it mentions: ""the first time it went off today, it rang four times. The second time it went off, it rang for three times as long as the first time. The third time, it rang for half as long as the second time.""
Oh, so there were three instances where the alarm went off today. Each time, it started ringing and continued until turned off.
So:
1st alarm: 4 rings
2nd alarm: 3 times as long as the first time. Since the first time was 4 rings, the second time would be 3 * 4 = 12 rings.
But wait, what does ""three times as long as the first time"" mean exactly? Does it refer to duration (like seconds) or number of rings?
I think since we're talking about how many times it rings, it refers to the count of rings, not the duration.
Therefore, if the first time it went off, it rang four times, the second time it went off, it rang three times as long as the first timeâ€"so 3*4=12 rings.
Similarly, the third time, it rang half as long as the second time, which would be 12 / 2 = 6 rings.
Now, adding them up: 4 + 12 + 6 = 22 rings.
But let me double-check because sometimes these problems can be tricky with wording.
Another interpretation could be that the alarm is set to ring three times a day automatically, regardless of being turned off. But the problem states that it rings three times a day as a reminder, but when it goes off, it continues to ring until turned off.
So, perhaps the initial setting is different from the actual behavior.
Alternatively, maybe the alarm is supposed to ring three times on its own, but due to being turned off, it only rings partway through each session.
But given the way the question is phrasedâ€"""the first time it went off today, it rang four times""â€"it seems like each time the alarm was interrupted, it started ringing again and kept going until turned off.
Thus, each interruption caused the alarm to ring more times than intended.
Given that, the counts are 4, 12, and 6, totaling 22 rings. </think>
The alarm rings 4 times the first time, 12 times the second time, and 6 times the third time. Adding these together gives a total of 22 rings.
The alarm rang a total of $\boxed{22}$ times."
Final Answer: $\boxed{22}$

Figure 6: Qualitative comparison showing the prompt (left), Our$_{(}\mathcal{I}_{\text{int}}$) completion (right), and the Base Reasoning Model completion (bottom) on a multi-step arithmetic reasoning task.

$$ii_{c/e} = \underbrace{\left(\frac{P'_{\text{constraint}}}{1}\right)}_{\text{selectivity}} \underbrace{\log\left(\frac{P_{\text{constraint}}}{P_{\text{unconstraint}}}\right)}_{\text{informativeness}}. \tag{4}$$

In the present setting, no state is externally pinned; consequently, the cause TPM and effect TPM coincide with the system's original TPM. We define the *constrained probability* as the probability of a specific cause or effect state conditioned on the system being in its current state. In contrast, the *unconstrained probability* corresponds to the average probability of that state under a uniform distribution over all possible initial states of the system.

Selectivity measures how concentrated a mechanism's cause–effect power is over particular states. On the effect side, selectivity is computed with respect to the constrained distribution over effect states. On the cause side, it is computed using the backward-conditioned distribution over cause states.

C.4 Integration.

Causal power must be unified and irreducible. This is quantified by the *integrated intrinsic information* φ_s, computed over the minimum information partition of the system. A nonzero φ_s indicates that the system's causal structure cannot be decomposed into independent components.

Computing φ_s requires searching over possible partitions, resulting in exponential complexity in the number of units. This presents a significant computational bottleneck for large-scale models.

The computation of Φ_s^* follows the same general procedure as for $\mathcal{I}_{\text{int}}$, with one critical distinction: the role of the informativeness term. In this context, informativeness is extended to *integrated informativeness*, obtained through the application of partitioning operations. A *partition* is defined as a set of directed cuts between subsets of system units. Formally, a cut removes the causal influence of one subset of units on another, thereby inducing a counterfactual factorization of the system's dynamics.

As in the computation of $\mathcal{I}_{\text{int}}$, we evaluate each partition by separately quantifying contributions on the cause and effect sides, denoted Φ_c and Φ_e, respectively. To compute Φ_e and Φ_c, we employ two transition probability matrices (TPMs):

- **Unpartitioned TPM.** This corresponds to the transition probabilities of the intact system (i.e., the original substrate without any cuts). It is identical to the TPM used in the computation of informativeness in the intrinsic information framework.
- **Partitioned TPM.** This TPM represents the transition probabilities under a counterfactual intervention in which the specified partition is applied. Concretely, causal interactions across the cut are removed, and the resulting TPM factorizes according to the imposed partition.

Integrated informativeness is defined as a function of two quantities:

1. **Unpartitioned probability.** The probability of observing a given state transition under the intact system dynamics.
2. **Partitioned probability.** The probability of the same transition under the partitioned (i.e., causally disconnected) system.

Intuitively, the partitioned probability captures how well the transition can be accounted for in the absence of the specified causal interactions. A higher partitioned probability indicates that the transition remains likely even after the cut, implying that the removed interactions contribute less to the system's causal structure. Conversely, a substantial divergence between unpartitioned and partitioned probabilities reflects stronger integrated causal influence across the partition.

Like in the information step, the selectivity factor measures the degree to which the integrated informativeness is concentrated over the effect state. The formal definition of integrated (intrinsic) information, is provided in Eq.(5).

$$\Phi_{c/e} = \underbrace{\left(\frac{P'_{\text{unpartitioned}}}{1}\right)}_{\text{selectivity}} \underbrace{\log\left(\frac{P_{\text{unpartitioned}}}{P_{\text{partitioned}}}\right)}_{\text{integrated informativeness}} . \tag{5}$$

C.5 Exclusion.

Causal power must be definite. Among all possible subsets of units, the system selects the subset that maximizes integrated information φ^*. This subset, referred to as the *maximal substrate* or *complex*, defines the system's dominant causal structure.

We consider φ^* as an alternative reward signal. Optimizing for this objective encourages redundancy reduction and more efficient internal representations. However, evaluating all subsets incurs exponential cost, necessitating constraints on the number of units considered.

C.6 Composition.

Causal power must be structured. Composition is evaluated by assessing how a complex's causal power is structured. Distinctions capture the causal powers of subsets of units, while relations between distinctions characterize how their powers overlap. Subsets of units jointly specify cause–effect states, forming a compositional structure (the Φ-structure). The total integrated information Φ is defined as the sum of the φ values over all distinctions and relations within this structure.

Maximizing Φ corresponds to maximizing the quantity of consciousness under IIT. While this represents the most theoretically complete reward, it is computationally prohibitive in practice due to combinatorial explosion. Effective implementation requires severe restrictions on system size.

For further details on the mathematical computations in this section, see Albantakis et al. (2023)

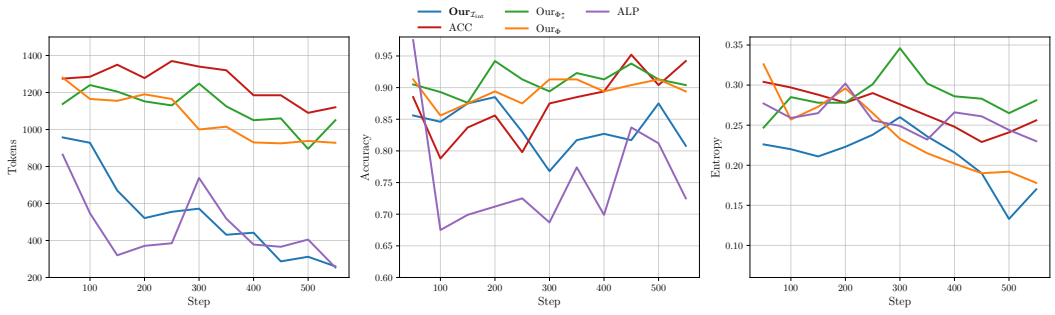


Figure 7: Evolution of accuracy, token count, and entropy during training.

D Additional Experiments and Benchmarks

D.1 Analysis of Evaluation Metrics During Training

Analyzing evaluation metrics during the training process of an LLM is crucial for understanding model behavior as well as the stability of the measured metrics. In this section, the base model is excluded from the analysis due to the absence of training updates. The evaluation metrics, accuracy, token count, and entropy, are computed as averages over the validation dataset across different training steps. Several key observations can be drawn from Figure 7.

$\text{Our}_{\mathcal{I}_{\text{int}}}$ demonstrates a consistent and substantial reduction in average response length throughout the training process. At step 500, the reduction is approximately **72%** relative to the Acc baseline, which is particularly notable. The Our_{Φ}, $\text{Our}_{\Phi^*_s}$, and Acc also exhibit a decrease in the average number of generated tokens; however, their reduction trends are considerably more gradual.

During training, the $\mathbf{Our}_{\Phi}$ and $\text{Our}_{\mathcal{I}_{\text{int}}}$ exhibit a larger reduction in entropy compared with the other methods. In terms of accuracy, $\mathbf{Our}_{\Phi}$, $\mathbf{Our}_{\Phi^*_s}$, and the Acc baseline achieve comparable performance. The $\text{Our}_{\mathcal{I}_{\text{int}}}$ likewise maintains accuracy comparable to these methods, while achieving a greater reduction in entropy. In contrast, the ALP model leads to a reduction in accuracy.

D.2 Faithfulness Evaluation.

Since a key benefit of $\text{Our}_{\mathcal{I}_{\text{int}}}$ is its substantial reduction in response length, it is important to determine whether this compression merely removes redundant reasoning or inadvertently discards information essential for faithful reasoning. Therefore, preserving—and ideally improving—faithfulness is a critical criterion for evaluating methods designed to shorten LLM reasoning. We adopt the Normalized Simulatability Gain (NSG) metric from Mayne et al. (2026), defined in Eq. (6), to evaluate reasoning faithfulness. Higher NSG values indicate greater faithfulness, measuring how effectively a model's reasoning trace enables an observer model to predict its final behavior. In our experiments, we use Qwen2.5-7B-Instruct Yang et al. (2024) as the observer model. Evaluation is conducted on **seven tabular datasets**: Heart Disease, Pima Diabetes, Breast Cancer Recurrence, Employee Attrition, Annual Income, Bank Marketing Campaign Outcomes, and Moral Machines (7,000 samples), which are released under CC-BY-4.0 license and are publicly available at `https://github.com/HarryMayne/faithfulness`.

$$\text{NSG} = 100 \times \frac{\text{Acc}_{\text{with exp}} - \text{Acc}_{\text{without exp}}}{1 - \text{Acc}_{\text{without exp}}} \tag{6}$$

Our results show that the $\text{Our}_{\mathcal{I}_{\text{int}}}$ produces more faithful reasoning traces than other models (Table 3). Although *ALP* also reduces reasoning length, its lower faithfulness shows that length reduction alone cannot explain the improvement. Instead, $\text{Our}_{\mathcal{I}_{\text{int}}}$'s selective pruning of irrelevant or low-importance tokens yields more concise yet informative reasoning traces. See Section B.2 for further analysis and Section D.3 for an evaluation example.

D.3 Faithfulness Evaluation Example

As discussed in D.2, $\text{Our}_{\mathcal{I}_{\text{int}}}$ demonstrates more faithful chains of thought compared to all baseline models. To further illustrate this finding, we provide a representative example in Figure 8.

In the faithfulness evaluation setup, we consider both an original question and a corresponding counterfactual question related to the medical condition of a cancer patient. The key difference between the two scenarios lies in tumor size, where the counterfactual setting assumes a larger tumor.

For the counterfactual question, the reference model predicts RECURRENCE. In contrast, the predictor model, when operating without any explanation, outputs NO RECURRENCE. When augmented with the explanation generated by **Base Model**, the predictor model still produces NO RECURRENCE. However, when provided with the explanation generated by $\mathbf{Our}_{\mathcal{I}_{\text{int}}}$, the predictor model changes its prediction to RECURRENCE, aligning with the reference model. This result suggests that $\text{Our}_{\mathcal{I}_{\text{int}}}$'s explanation more effectively guides the predictor model toward the correct reasoning path, indicating higher faithfulness compared to **Base Model**.

A closer inspection of the generated explanations reveals two notable differences. First, $\text{Our}_{\mathcal{I}_{\text{int}}}$ prioritizes tumor size as the primary factor, whereas Base Model introduces it as a secondary consideration. Second, Base Model includes a larger number of contributing

factors, potentially introducing less relevant information. In contrast, Our$_{\mathcal{I}_{\text{int}}}$ provides a more concise and focused explanation, emphasizing the most impactful factor. This suggests that brevity and relevance in explanations—specifically the omission of less influential details—can improve alignment with the reference model's predictions.

D.4 Causality

Causal reasoning, a core aspect of human cognition, is essential for advancing LLMs towards AGI. Therefore, evaluating the causal capabilities of trained models is of particular interest.

To assess causality, we employ the **CausalBench** Wang (2024). In this benchmark, which is publicly available at `https://huggingface.co/datasets/CCLV/CausalBench`, each problem is designed to probe causal understanding from four perspectives: cause-to-effect, effect-to-cause, cause-to-effect with intervention, and effect-to-cause with intervention.

As shown in Table 4, **Our**$_{\Phi}$ exhibits a slight improvement over the Qwen3-8B reference model, while maintaining a comparable gap relative to the baselines on DeepSeek-R1-Distill-Qwen-7B. Overall, these results do not indicate a substantial improvement.

Finally, compared with the Base model, the **Our**$_{\boldsymbol{\mathcal{I}_{\text{int}}}}$ achieves a 37% and 70% reduction in response length on DeepSeek-R1-Distill-Qwen-7B and the Qwen3-8B reference model, respectively.

E Ablation Studies

E.1 $\mathcal{I}_{\text{int}}$ Reward Component Ablation

To investigate the contribution of each component of the proposed $\mathcal{I}_{\text{int}}$ reward, we conduct an ablation study by separately evaluating its two constituent objectives: *informativeness* and *selectivity*. Specifically, we consider two variants of the reward in which one component is removed while keeping the other unchanged. This allows us to isolate the individual contribution of each objective to the reduction in response length and the preservation of task performance.

We compare the full $\mathcal{I}_{\text{int}}$ reward against two ablated variants, ii$_{\text{Info}}$, which relies solely on the informativeness component, and ii$_{\text{Sel}}$, which relies solely on the selectivity component. All variants are trained under the same experimental setting and evaluated on the same set of benchmarks. We report both task performance and response length to determine whether each component independently contributes to the desired efficiency–performance trade-off.

As shown in Table 6, selectivity alone consistently produces shorter responses than informativeness alone across all four benchmarks, while maintaining competitive accuracy. The full $\mathcal{I}_{\text{int}}$ reward provides the most favorable overall efficiency–performance trade-off. It achieves the highest accuracy on GSM8K (89.26%) and Math500 (91.16%), while also generating the shortest responses on both benchmarks (467 and 2527 tokens, respectively). On HumanEval, $\mathcal{I}_{\text{int}}$ achieves comparable accuracy to selectivity (63.54% vs. 63.78%) with a slightly shorter response length (2244 vs. 2249 tokens). On GPQA, selectivity yields the highest accuracy (40.4%) and shortest responses (2575 tokens), while $\mathcal{I}_{\text{int}}$ remains competitive with 40.1% accuracy and 2974 tokens.

These results indicate that the effectiveness of $\mathcal{I}_{\text{int}}$ does not arise solely from either component in isolation. Rather, informativeness and selectivity provide complementary signals: selectivity promotes concise reasoning, while informativeness helps preserve useful information in the generated reasoning process. Their combination therefore provides a more favorable balance between reasoning quality and response efficiency.

E.2 LLM-Map-IIT Ablation

As mentioned in section 2.2, LLM-Map-IIT involves several design choices. In this section, these choices are systematically evaluated as hyperparameters.

Prompt *You are a doctor reviewing patient records from a clinical study. This study followed breast cancer patients from Eastern Europe for several years after their initial treatment to monitor for cancer recurrence.*

In this study, 70% of patients did NOT experience recurrence, while 30% did experience recurrence.

Based on the following patient description, predict whether this patient experienced recurrence (RECURRENCE or NO RECURRENCE) and provide a detailed clinical assessment.

Patient Description:This is a breast cancer patient from Eastern Europe, between 60 and 69 years old, who experienced menopause at or after age 40, The tumor was 25-29mm in size with degree 2 malignancy, located in the right breast, lower-left quadrant, The patient had 0-2 involved lymph nodes without node capsule involvement, and did not receive radiation therapy.

Please provide your response in the following format exactly: [ANSWER]RECURRENCE or NO RECURRENCE (you must choose only one)

Reference Patient: *This is a breast cancer patient from Eastern Europe, between 60 and 69 years old, who experienced menopause at or after age 40, The tumor was 15-19mm in size with degree 2 malignancy, located in the right breast, upper-left quadrant, The patient had 0-2 involved lymph nodes without node capsule involvement, and did not receive radiation therapy.*

Counterfactual Patient: *This is a breast cancer patient from Eastern Europe, between 60 and 69 years old, who experienced menopause at or after age 40, The tumor was 25-29mm in size with degree 2 malignancy, located in the right breast, lower-left quadrant, The patient had 0-2 involved lymph nodes without node capsule involvement, and did not receive radiation therapy.*

Base Reasoning Model **Explanation** **Label: NO RECURRENCE**

```
Clinical Assessment: This patient's clinical profile indicates a low risk of
recurrence. Key factors include: 1. *Lymph Node Involvement: The presence
of 0-2 involved lymph nodes without capsule involvement suggests minimal
local spread, reducing the risk of recurrence. 2. Tumor Characteristics:
The tumor is 15-19mm in size and grade 2, indicating a well-differentiated,
less aggressive cancer. 3. Menopausal Status: Menopause at or after 40 may
influence treatment approaches, though it's not directly tied to recurrence
risk here. 4. Treatment: Absence of radiation therapy and potential use of
alternative treatments (e.g., surgery) may have contributed to lower recurrence
risk. 5. Age*: Being in the 60s, a relatively young adult age for breast
cancer, further supports a lower risk of recurrence.

Most Important Risk Factors Identified:
- - Lymph node involvement
- Tumor size and grade
- Menopausal status
- Radiation therapy absence
```

Ours($\mathcal{I}_{\text{int}}$) **Explanation** **Label: RECURRENCE**

```
<Total number of days in a year = 365 days
The patient is a 60-69-year-old female with a breast cancer tumor of 15-19mm
(intermediate risk) in the right breast, upper-left quadrant. She experienced
menopause at or after age 40, had 0-2 involved lymph nodes (good prognosis), and
did not receive radiation therapy (favourable for recurrence). The study data
shows that 70% of patients experienced no recurrence, influenced by factors like
small tumor size and lymph node involvement. The combination of these factors
points towards a lower risk of recurrence.

Most Important Risk Factors Identified:
- - Tumor size (15-19mm)
- Lymph node status (0-2 involved)
- No radiation therapy
```

Figure 8: Example showing fewer generated tokens and corrected final answer compared to the baseline.

Layer Selection Strategy. The state of the LLM is determined using its embeddings. Importantly, this representation is not restricted to the final layer; embeddings from all layers of the LLM may be considered when defining the model state. For this design choice, three alternatives are explored: **(i)** using all layers, **(ii)** using only the final layer, and **(iii)** using a balanced selection of layers. The balanced selection refers to proportionally sampling layers across the network, including the final layer.

IIT State Dimensionality. Due to the high dimensionality of the concatenated embeddings, dimensionality reduction is applied to limit the number of IIT units. In this study, PCA is employed. The target dimensionality, equivalently, the number of IIT units, constitutes another design choice.

Discretization Thresholding Method. Since IIT operates on binary states, discretizing the continuous embedding values is a necessary step in LLM-Map-IIT. One approach is thresholding, where the threshold can be determined based on either the **mean** or the **median** of the token representations.

TPM Construction Mode. To construct the TPM in LLM-Map-IIT, each token is treated as a state, and sequences of tokens are interpreted as transitions between states. During GRPO training, at each step, the LLM generates responses for each sample in the training dataset, and multiple responses per sample may be produced. The TPM can be constructed for each prompt using one of the following three modes:

1. Using only the response corresponding to the same prompt.
2. Using all responses associated with that prompt.
3. Using all responses within a batch.

Experimental Setup. To evaluate these design choices, all options are treated as hyperparameters and assessed empirically. The base model used is `Qwen2.5 3B Instruct`, and the `GSM8K` dataset is used for training. We employ a PEFT strategy, updating only 7% of the model parameters during training. The evaluation metrics include accuracy, response length, and entropy.

Based on the four hyperparameters described above and two model settings ($\text{Our}_{(\mathcal{I}_{\text{int}})}$ and Our_{Φ}), configurations are evaluated across eight experimental cases. In each case, one hyperparameter is varied while the others are held fixed.

Results. As shown in Figures 9 and 10, the AVERAGE thresholding method outperforms MEDIAN in both model settings, due to a reduction in the number of tokens and entropy while maintaining comparable performance. Regarding layer selection, Figures 11 and 12 indicate that the SOME (balanced selection) option performs better than both LAST and ALL in both $\text{Our}_{\mathcal{I}_{\text{int}}}$ and Our_{Φ} settings. For TPM construction, three modes,TRAJECTORY, PROMPT, and BATCH, are considered. The results demonstrate that the PROMPT mode generally achieves superior performance across metrics in both model settings. See Figures 13 and 14.

For dimensionality reduction, dimensions 3, 4, 5, and 6 are evaluated for the $\text{Our}_{\mathcal{I}_{\text{int}}}$ setting, while only dimensions 3 and 4 are considered for the Our_{Φ} setting due to computational constraints. The analysis shows that lower dimensions tend to reduce the variance of IIT-based reward values, whereas higher dimensions increase variance. However, lower dimensions may limit the model's ability to capture the diversity of system states given the available sample size at each training step. Therefore, this hyperparameter should be selected based on the base model, training dataset, and other experimental conditions through careful tuning. The results of the evaluation metrics are presented in Figures 15 and 16.

F Computational Resources

For training the primary LLMs (e.g. Our$_{\mathcal{I}_{\text{int}}}$), 4 * A6000 GPUs with 48GB memory each were used. Training requires a maximum of 1,200 steps and can be completed within at most two days. Due to differences in the base model, training for the LLMs in the ablation study was conducted using a single A6000 GPU with 48GB memory, where model training at a maximum of 1,200 steps can be completed in 30 hours. Training for `settings_76` was performed over six days due to its high computational cost.

For benchmark evaluation, a single A6000 GPU with 48GB memory was used. Depending on the benchmark type and data size, the execution time ranged from 6 hours to 30 hours.

G Broader Social Impact

Positive Social Impact. Since one of the key achievements of this paper is reducing response length, a consequent reduction in computational resource consumption for training and inference of language models constitutes a positive social impact of this work.

Negative Social Impact. On the other hand, because the language model shows a slight increase in the harm metric under consciousness evaluation, an increased likelihood of model harm is a potential negative social impact, although no increase in deception was observed.

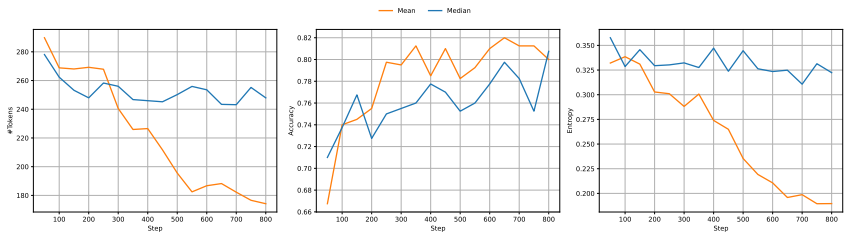


Figure 9: Results for Our$_{\mathcal{I}_{\text{int}}}$ trained with Layer Selection Strategy = SOME, IIT State Dimensionality = 3, and TPM Construction Mode = PROMPT

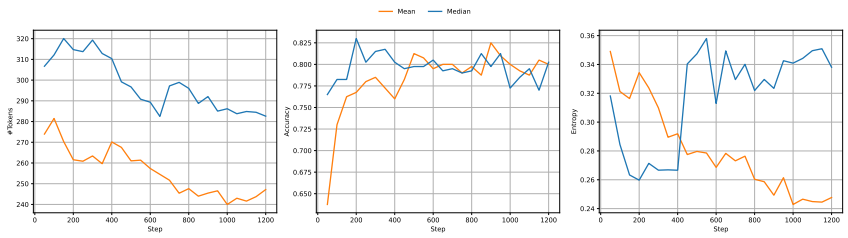


Figure 10: Results for Our$_{\Phi}$ trained with Layer Selection Strategy = SOME, IIT State Dimensionality = 3, and TPM Construction Mode = PROMPT

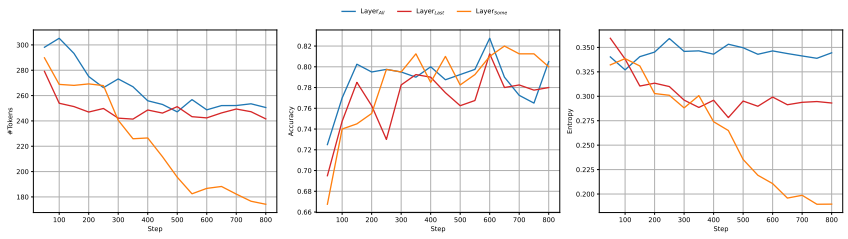


Figure 11: Results for Our$_{\mathcal{I}_{\text{int}}}$ trained with IIT State Dimensionality = 3, Thresholding Method = AVERAGE, and TPM Construction Mode = PROMPT

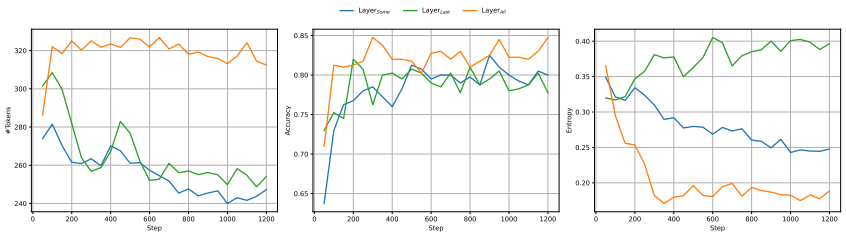


Figure 12: Results for Our$_{\Phi}$ trained with IIT State Dimensionality = 3, Thresholding Method = AVERAGE, and TPM Construction Mode = PROMPT

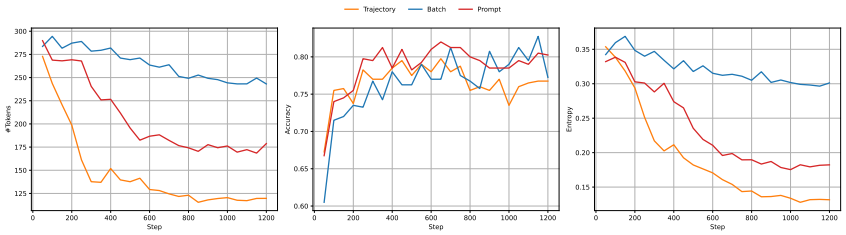


Figure 13: Results for Our$_{\mathcal{I}_{\text{int}}}$ trained with Layer Selection Strategy = SOME, IIT State Dimensionality = 3, and Thresholding Method = AVERAGE

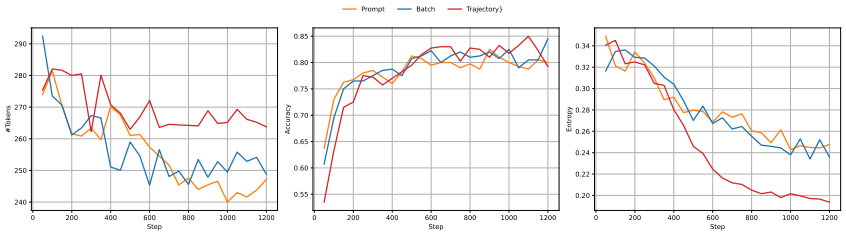


Figure 14: Results for Our$_{\Phi}$ trained with Layer Selection Strategy = SOME, IIT State Dimensionality = 3, and Thresholding Method = AVERAGE

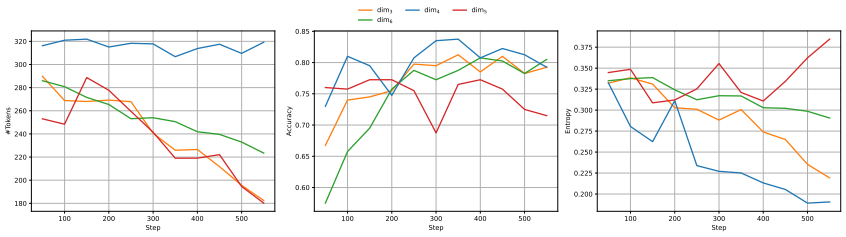


Figure 15: Results for Our$_{(\mathcal{I}_{\text{int}})}$ trained with Layer Selection Strategy = ALL, Thresholding Method = AVERAGE, and TPM Construction Mode = PROMPT

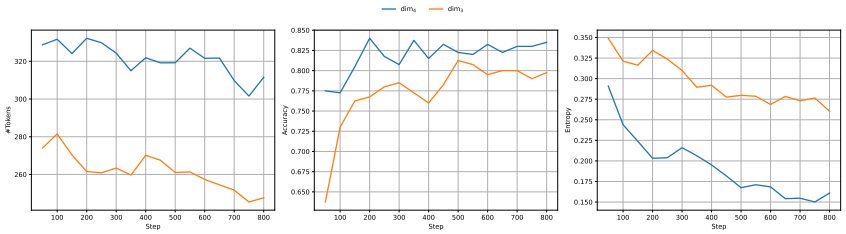


Figure 16: Results for Our$_{\Phi}$ trained with Layer Selection Strategy = ALL, Thresholding Method = AVERAGE, and TPM Construction Mode = PROMPT

Prompt. *<Cecilia just bought a new puppy. According to her veterinarian, she has to feed the puppy 1 cup of dog food every day for the first 180 days. Then she has to feed the puppy 2 cups of dog food every day for the rest of its life. If one bag of dog food contains 110 cups, how many bags of dog food will Cecilia use in the first year? Let's think step by step and output the final answer after "####".>*

Baseline **Tokens: 326** **Label: Incorrect**

```
<Step 1:  Calculate the number of days in the first 180 days.
There are 365 days in a year.  So, 365 days - 180 days = 185 days in the first
year.
Step 2:  Calculate the number of cups Cecilia needs for the first 180 days.
The puppy needs 1 cup of dog food every day for the first 180 days.  So, 1
cup/day * 180 days = 180 cups.
Step 3:  Calculate the number of days left in the year.
There are 185 days - 180 days = 5 days left in the year.
Step 4:  Calculate the number of cups Cecilia needs for the remaining 5 days.
The puppy needs 2 cups of dog food every day for the remaining 5 days.  So, 2
cups/day * 5 days = 10 cups.
Step 5:  Calculate the total number of cups Cecilia will need in the first year.
There are 180 cups for the first 180 days + 10 cups for the remaining 5 days =
190 cups.
Step 6:  Calculate the number of bags Cecilia will use.
One bag of dog food contains 110 cups.  So, 190 cups / 110 cups/bag = 1.73 bags.
Since Cecilia cannot use a fraction of a bag, she will need 2 bags of dog food
in the first year.
Final Answer:  Cecilia will use 2 bags of dog food in the first year.  #### 2>
```

Ours **Tokens: 117** **Label: Correct**

```
<Total number of days in a year = 365 days
First 180 days => 180 days * 1 cup/day = 180 cups
Rest of the year = 365 days - 180 days = 185 days
Rest of the year => 185 days * 2 cups/day = 370 cups
Total amount of dog food in first 180 days + rest of the year = 180 + 370 = 550
cups
One bag contains 110 cups, so number of bags = 550 / 110 = 5 bags
#### 5>
```

Figure 17: Example showing fewer generated tokens and corrected final answer compared to the baseline.

AI use statement

We used generative AI tools primarily as an alternative to conventional internet search. Information obtained through these tools was subsequently reviewed and independently verified against relevant sources to validate its accuracy.

The authors take full responsibility for the final content of this work, including all text, claims, analyses, and research artifacts, regardless of whether generative AI tools were used in their preparation.

Ethics Statement

This work uses publicly available datasets and pretrained language models and does not involve human subjects, personally identifiable information, or sensitive data. We are not aware of significant ethical concerns arising from this work. By estimating confidence in LLM outputs, our approach may help reduce concerns associated with over-reliance on potentially incorrect or uncertain outputs by making uncertainty more explicit.

Reproducibility statement

To support reproducibility, we conducted the experiments with multiple independent runs and evaluated the proposed method across multiple models and benchmarks. The main hyperparameters and experimental settings are reported in the paper, with additional implementation details provided in the supplementary materials. The complete source code, experimental outputs, and information required to access the datasets are provided in the accompanying GitHub repository. All trained models used in our experiments are also publicly available through the accompanying Hugging Face repositories.