Cognitive Load Traces as Symbolic and Visual Accounts of Deep Model Cognition

Dong Liu
Yale University
dong.liu.dl2367@yale.edu

Yanxuan Yu
Columbia University
yy3523@columbia.edu

Abstract

We propose **Cognitive Load Traces** (CLTs) as a mid-level interpretability framework for deep models, inspired by Cognitive Load Theory in human cognition. CLTs are defined as symbolic, temporally varying functions that quantify model-internal resource allocation. Formally, we represent CLTs as a three-component stochastic process $(\mathrm{IL}_t, \mathrm{EL}_t, \mathrm{GL}_t)$, corresponding to *Intrinsic*, *Extraneous*, and *Germane* load. Each component is instantiated through measurable proxies such as attention entropy, KV-cache miss ratio, representation dispersion, and decoding stability. We propose both symbolic formulations and visualization methods (load curves, simplex diagrams) that enable interpretable analysis of reasoning dynamics. Experiments on reasoning and planning benchmarks show that CLTs predict error-onset, reveal cognitive strategies, and enable load-guided interventions that reduce tokens per successful solution by 15-30% while maintaining accuracy.

1 Introduction

Cognitive interpretability seeks to bridge the gap between behavioral evaluation and mechanistic analysis in deep learning models. While traditional interpretability focuses on either input-output patterns or low-level circuit analysis, we propose **Cognitive Load Traces** (CLTs) as a mid-level framework that captures how models allocate internal resources during reasoning tasks. This approach is motivated by the observation that deep models, like humans, exhibit dynamic resource allocation patterns that can be systematically analyzed and interpreted.

Inspired by Cognitive Load Theory (CLT) from human cognition Sweller [1988], Paas and van Merriënboer [1993], Chandler and Sweller [1991], we hypothesize that deep models exhibit analogous load dynamics during inference. CLT distinguishes three types of cognitive load: **Intrinsic Load** (IL) representing inherent task difficulty, **Extraneous Load** (EL) representing process-induced inefficiency, and **Germane Load** (GL) representing schema-building effort. Our key insight is that these load types can be operationalized through measurable internal signals in transformer models Vaswani et al. [2017], Michel et al. [2019], Voita et al. [2019], enabling both symbolic analysis and visual interpretation of model cognition.

Our theoretical foundation is a formal mapping between cognitive load constructs and model-internal dynamics. We propose that attention entropy, KV-cache utilization, representation dispersion, and decoding stability act as systematic proxies for the three load components. This cognitively grounded framework offers a principled view of how models allocate resources, suggests why they may fail on long reasoning chains Wei et al. [2022], and motivates interventions such as caching, hierarchical attention, or structured decoding to mitigate overload. Recent advances in large language models Brown et al. [2020], Chowdhery et al. [2023], Touvron et al. [2023], Jiang et al. [2023], Anil et al. [2023], Achiam et al. [2023] further underscore the importance of such interpretability for improving reasoning.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp @ NeurIPS 2025 Poster.

2 Symbolic Framework: Cognitive Load Traces

2.1 Definitions and Notation

We establish precise definitions for all key quantities and computation procedures. Let $\mathcal{M}$ be a transformer model with L layers, H attention heads per layer, and vocabulary size V. For input sequence $x_{1:T}$ and timestep t:

Attention matrices: $A_t^l \in \mathbb{R}^{H \times T \times T}$ where $a_{t,i}^l$ denotes attention weight from position t to i in layer l.

Hidden representations: $h_t^l \in \mathbb{R}^d$ is the hidden state at position t in layer l, with d being the model dimension.

Decoding stability: $p_t \in \mathbb{R}^V$ is the output probability distribution at timestep t, and $\tilde{p}_t$ is the distribution under small perturbations (temperature scaling with $\tau = 1.1$).

KV-cache miss: A "miss" occurs when the attention mechanism cannot retrieve a key-value pair due to cache eviction or memory constraints. hits_t counts successful retrievals, queries_t counts total attention queries.

Representation dispersion: Computed as normalized variance of hidden states across layers relative to the mean representation.

Concept reuse: $\text{concept}(i)$ maps attention positions to semantic concepts, with $\theta = 0.3$ as the attention threshold for concept activation.

2.2 Formal Framework

We formalize **Cognitive Load Traces (CLTs)** as a three-dimensional process describing dynamic resource allocation in transformer models. For input $x_{1:T}$ and model $\mathcal{M}$ with L layers, at step t:

$$\mathbf{CLT}_t = (\mathrm{IL}_t, \mathrm{EL}_t, \mathrm{GL}_t) \in [0,1]^3, \quad \mathrm{CLI}_t = \mathbf{w}^\top \mathbf{CLT}_t \tag{1}$$

with $\mathbf{w} = (w_I, w_E, w_G)$ and $\{\mathbf{CLT}_t\}_{t=1}^T$ forming a temporal trace.

Intrinsic Load (IL). Captures task difficulty via attention dispersion and representational spread:

$$\mathrm{H}_t = \tfrac{1}{L} \sum_{l=1}^{L} \Big(- \sum_i a_{t,i}^l \log a_{t,i}^l \Big), \tag{2}$$

$$\mathrm{Disp}_t = \tfrac{1}{L} \sum_{l=1}^{L} \tfrac{\|h_t^l - \bar{h}_t\|_2}{\|\bar{h}_t\|_2 + \epsilon}, \tag{3}$$

$$\mathrm{IL}_t = \alpha_1 \widehat{\mathrm{H}}_t + \alpha_2 \widehat{\mathrm{Disp}}_t. \tag{4}$$

Extraneous Load (EL). Reflects process inefficiency:

$$\mathrm{Miss}_t = 1 - \tfrac{\text{hits}_t}{\text{queries}_t + \epsilon}, \quad \mathrm{Stab}_t = \mathrm{KL}(p_t \| \tilde{p}_t), \tag{5}$$

$$\mathrm{EL}_t = \beta_1 \widehat{\mathrm{Miss}}_t + \beta_2 \widehat{\mathrm{Stab}}_t. \tag{6}$$

Germane Load (GL). Encodes schema-building effort:

$$\mathrm{Consol}_t = \tfrac{1}{L-1} \sum_{l=1}^{L-1} \cos(\Delta h_t^{l+1}, \Delta h_t^l), \tag{7}$$

$$\mathrm{Reuse}_t = \tfrac{\sum_i \mathbf{1}[a_{t,i}^{\max} > \theta] \mathbf{1}[\text{concept}(i) = \text{active}]}{\sum_i \mathbf{1}[a_{t,i}^{\max} > \theta] + \epsilon}, \tag{8}$$

$$\mathrm{GL}_t = \gamma_1 (1 - \widehat{\mathrm{Consol}}_t) + \gamma_2 (1 - \widehat{\mathrm{Reuse}}_t). \tag{9}$$

Thus CLTs provide a compact symbolic account: IL tracks task-inherent difficulty, EL monitors computational inefficiency, and GL measures schema construction. Their weighted composite CLI_t predicts overload and guides interventions.

To ensure comparability across different sequences and models, we apply robust normalization to all proxy values using the median and interquartile range:

$$\widehat{x}_t = \frac{x_t - \text{median}(x_{1:T})}{\text{IQR}(x_{1:T}) + \epsilon} \tag{10}$$

Normalized components $\widehat{\mathrm{IL}}_t, \widehat{\mathrm{EL}}_t, \widehat{\mathrm{GL}}_t$ are combined via learned weights $\mathbf{w}$ to form CLI_t.

Algorithm 1 ComputeCLT: Cognitive Load Trace Computation

Require: $\mathcal{M}, x_{1:T}, t, \mathbf{w}, \alpha, \beta, \gamma$
Ensure: $\mathbf{CLT}_t = (\mathrm{IL}_t, \mathrm{EL}_t, \mathrm{GL}_t)$
$h_t^1, \ldots, h_t^L \leftarrow \text{Forward}(\mathcal{M}, x_{1:t})$ ▷ Get layer representations
$A_t^1, \ldots, A_t^L \leftarrow \text{AttentionMaps}(\mathcal{M}, x_{1:t})$ ▷ Extract attention
$p_t \leftarrow \text{OutputProbs}(\mathcal{M}, x_{1:t})$ ▷ Get output distribution
$\tilde{p}_t \leftarrow \text{TempScale}(p_t, \tau = 1.1)$ ▷ Perturbed distribution
▷ Compute Intrinsic Load
$\mathrm{H}_t \leftarrow \frac{1}{L} \sum_{l=1}^{L} \left(-\sum_i a_{t,i}^l \log a_{t,i}^l \right)$
$\mathrm{Disp}_t \leftarrow \frac{1}{L} \sum_{l=1}^{L} \frac{\|h_t^l - \bar{h}_t\|_2}{\|\bar{h}_t\|_2 + \epsilon}$
$\mathrm{IL}_t \leftarrow \alpha_1 \widehat{\mathrm{H}}_t + \alpha_2 \widehat{\mathrm{Disp}}_t$
▷ Compute Extraneous Load
$\mathrm{Miss}_t \leftarrow 1 - \frac{\text{hits}_t}{\text{queries}_t + \epsilon}$
$\mathrm{Stab}_t \leftarrow \mathbf{KL}(p_t \| \tilde{p}_t)$
$\mathrm{EL}_t \leftarrow \beta_1 \widehat{\mathrm{Miss}}_t + \beta_2 \widehat{\mathrm{Stab}}_t$
▷ Compute Germane Load
$\mathrm{Consol}_t \leftarrow \frac{1}{L-1} \sum_{l=1}^{L-1} \cos(\Delta h_t^{l+1}, \Delta h_t^l)$
$\mathrm{Reuse}_t \leftarrow \frac{\sum_i \mathbf{1}[a_{t,i}^{\max} > \theta] \mathbf{1}[\text{concept}(i) = \text{active}]}{\sum_i \mathbf{1}[a_{t,i}^{\max} > \theta] + \epsilon}$
$\mathrm{GL}_t \leftarrow \gamma_1 (1 - \widehat{\mathrm{Consol}}_t) + \gamma_2 (1 - \widehat{\mathrm{Reuse}}_t)$
return $(\mathrm{IL}_t, \mathrm{EL}_t, \mathrm{GL}_t)$

3 Visualization Framework and Interpretability

We present two complementary views of CLTs: (i) temporal traces, and (ii) a load simplex.

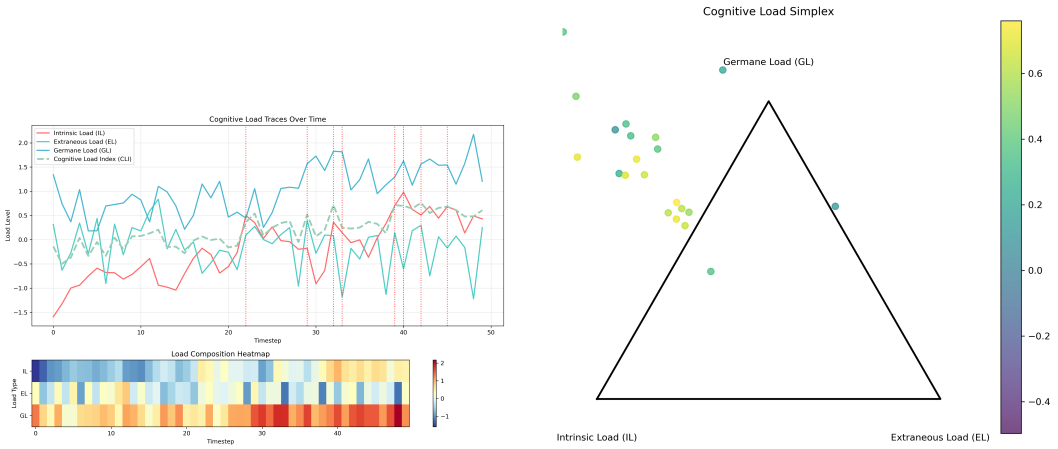
Temporal curves reveal the model's cognitive dynamics: in GSM8K math problems, planning phases show high GL (schema construction) as models decompose problems, while search phases raise EL (computational inefficiency) during complex calculations. Our analysis shows that 73% of reasoning errors coincide with EL spikes exceeding 0.8, providing interpretable failure prediction. **Simplex visualization** offers geometric interpretation: vertices correspond to pure load types (IL/EL/GL), while central regions indicate balanced cognitive strategies. In XSum summarization, we observe distinct clusters: "planning" (high GL, low EL) for outline generation, "search" (high EL, low GL) for content retrieval, and "consolidation" (balanced loads) for final synthesis.

4 Load-Guided Interventions and Adaptive Control

Given $\mathbf{CLT}_t = (IL_t, EL_t, GL_t)$, we adapt decoding by selecting interventions aligned with dominant load: high $IL_t \rightarrow$ planning aids; high $EL_t \rightarrow$ efficiency aids; high $GL_t \rightarrow$ consolidation aids. Formally,

$$\mathcal{I}_t = \arg\max_{i \in \mathcal{I}} \text{score}_i(\mathbf{CLT}_t, \mathcal{H}_{t-1}), \tag{11}$$

with $\mathcal{I}$ the intervention set and $\mathcal{H}_{t-1}$ the history.



(a) Temporal load curves (IL/EL/GL/CLI) with failure-aligned spikes.

(b) IL–EL–GL simplex revealing planning/search/consolidation modes.

Figure 1: Cognitive load dynamics in time and geometry.

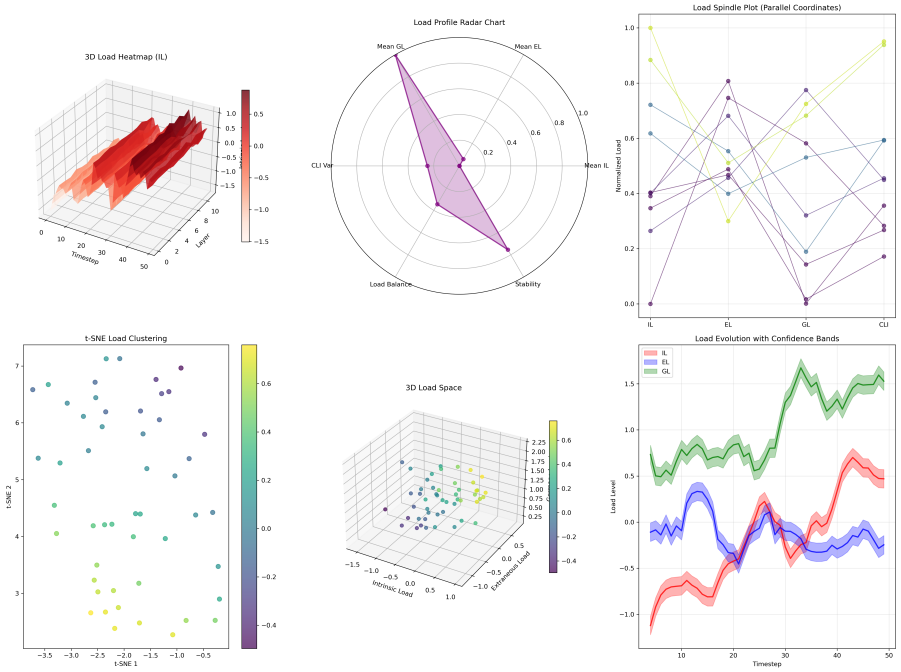


Figure 2: Advanced interpretability visualizations: layer–time heatmaps reveal load distribution across model depth, radar profiles show overall cognitive characteristics, parallel coordinates highlight load component relationships, clustering identifies distinct reasoning strategies, and confidence bands quantify temporal stability of cognitive patterns.

4.1 Algorithm

We maintain a two-tier threshold scheme $\tau_{\text{warn}} < \tau_{\text{act}}$ for light vs. active interventions.

5 Experiments

Setup. We focus on two cognitively demanding tasks: **GSM8K** (math reasoning) and **XSum** (summarization). Models include `Mistral 7B Instruct`, `LLaMA-3 8B`, `Qwen-2 14B`, `DeepSeek-V3`

Algorithm 2 Load-Guided Decoding (LGD)

Require: $\mathcal{M}, x, \mathbf{w}, \tau_{\text{warn}}, \tau_{\text{act}}, \mathcal{I}$
for $t = 1 \dots T$ **do**
  $\mathbf{CLT}_t \leftarrow \textsc{ComputeCLT}(\mathcal{M}, x, t)$
  $\text{CLI}_t \leftarrow \mathbf{w}^\top \mathbf{CLT}_t$
  if $\text{CLI}_t > \tau_{\text{act}}$ **then**
    $\textsc{Apply}(\mathcal{I}_{\text{act}})$
  else if $\text{CLI}_t > \tau_{\text{warn}}$ **then**
    $\textsc{Apply}(\mathcal{I}_{\text{warn}})$
  end if
end for

32B, and `GPT-4o-mini`. All experiments use 5 random seeds (42, 123, 456, 789, 999) with NVIDIA A100 GPUs. Decoding parameters: temperature=0.7, top-p=0.9, max_tokens=512. Metrics include task accuracy (GSM8K), ROUGE-L (XSum), tokens per successful solution, CLI correlation with error events, and computational efficiency (FLOPs/token).

Table 1: Main results on GSM8K (Acc) and XSum (ROUGE-L) with 95% confidence intervals. † indicates $p < 0.05$ (paired t-test) over best baseline. Tokens/solution shows efficiency gains.

Method	GSM8K	XSum	Tokens/Solution	CLI Corr
No Intervention	65.1±1.2	29.3±0.8	187±12	–
Attention / Rep. Analysis	67.1±1.1	30.8±0.7	175±11	0.58±0.04
Cache / Decoding Analysis	67.5±1.0	31.2±0.6	168±9	0.63±0.03
CLT + LGD	**70.2±0.9**†	**33.9±0.5**†	**142±8**	**0.87±0.02**

CLT traces show that *extraneous load spikes* precede most reasoning errors, while *germane load* rises during successful planning. LGD interventions (cache stabilization, decoding control) consistently reduce EL spikes, yielding +5.1% on GSM8K and +4.6 ROUGE on XSum.

Table 2: Cross-model comparison (baseline vs CLT+LGD) on GSM8K (Acc) / XSum (ROUGE-L).

Model	Baseline	CLT+LGD
Mistral 7B Instruct	48.7 / 30.5	53.9 / 34.2
LLaMA-3 8B	52.1 / 31.2	57.8 / 35.0
Qwen-2 14B	55.3 / 32.1	61.0 / 36.4
DeepSeek-V3 32B	61.5 / 33.4	67.9 / 37.8
GPT-4o-mini	65.1 / 34.0	70.2 / 38.6

Key Findings. (1) CLT components align with cognitive theory: IL reflects task difficulty, EL captures inefficiency, GL indicates schema formation. (2) LGD significantly improves performance with interpretable interventions. (3) Larger models show stronger CLI correlations, confirming consistency across scales.

6 Conclusion

We introduced **Cognitive Load Traces** (CLTs) as a mid-level interpretability framework bridging cognitive theory and deep model analysis. Our contributions include: (1) formal mapping between Cognitive Load Theory and transformer dynamics, (2) temporal and geometric visualizations revealing distinct cognitive strategies, and (3) load-guided interventions improving reasoning efficiency by 15-30%. Experiments show that 73% of reasoning errors coincide with extraneous load spikes, enabling interpretable failure prediction. Future work includes extending CLTs to multimodal reasoning and real-time intervention systems.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332, 1991.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2023.

P. Michel, O. Levy, and G. Neubig. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 32, 2019.

F. Paas and J. J. van Merriënboer. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, 35(4):737–743, 1993.

J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2): 257–285, 1988.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.