CAUSALITY IS WHAT YOU MAY DOWNGRADE IF YOU LOOK BEYOND DATA DISTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that the causality [1] of language models (LMs) may not be necessary nor optimal. This is the case when **system behavior** (denoted as S) is incorporated as a first-principle Bayesian feature. Here, S refers to **extra dominant factors** beyond the data space, and it especially captures the coupled effects of algorithms, hardware, runtime, and natural language. Despite being the de facto foundation of modern architecture, recent studies indicate persistent mismatches and contradictions with causality. These issues largely stem from system behavior rather than the data distribution. We therefore propose the System Behavior Dynamics (`SBD`) framework, which incorporates S as an irreducible component of the evidence lower bound (ELBO). `SBD` theoretically reveals a counter-intuitive **Causality Tax** phenomenon, where causality emerges as a suboptimal approximation with an additional structural error, due to the obliviousness to S. To address the challenge of latent variable analysis, we validate the `SBD`-predicted impact of S via implicit measurements, theoretical-bound-guided controls, and Neural Tangent Kernel (NTK) evaluations. In particular, we construct Green Shell (`GSH`) to show the possibility of reducing Causality Tax. `GSH` is a non-causal variational family, and it replaces the sequential dependency chain of S components with a divide-and-conquer partition. NTK spectra in the lazy-training regime confirm that `GSH` always achieves significantly tighter error bounds than causality, with $7dB+$ (or $5\times$ more) improvement in signal-to-noise ratio. In the relatively later stage of lazy-training, `GSH` further leads to superior generalization (up to 20% richer multi-scale fitting capabilities). Taken together, `SBD` establishes system behavior as a complementary theoretical abstraction besides causality and distribution fitting, opening new research avenues such as designing and optimizing LM base models.

1 INTRODUCTION

Modern autoregressive language models (LMs) (Vaswani et al., 2017), particularly large language models (LLMs) (Yang et al., 2025; Touvron et al., 2023; Achiam et al., 2023), adhere to the *causal formulation*: each token is generated conditioned on the incrementally growing prefix. However, we observe increasing mismatches and contradictions with causality from multiple aspects (Atil et al., 2024; Papadopoulos et al., 2024; Arriola et al., 2025; Fu et al., 2024). For instance, causality seems to conflict with Shannon's law (Papadopoulos et al., 2024), and a deliberate relaxation of causality is self-consistent and feasible (Fu et al., 2024; Arriola et al., 2025). Those contradictions originate systematically and typically beyond data distributions, and we therefore summarize these factors as *system behavior* S (Definition 1.1). Theoretically, S is *deliberately broad to cover all that is orthogonal to the data space*, as any narrower definition would arbitrarily exclude aforementioned systematic effects. In practice, even under the computational and observability constraints, the observable and manipulable subset S is still far beyond the passive infrastructure of hardware and runtime (Atil et al., 2024), such as those from algorithms, natural languages and other yet-to-be-understood factors (Papadopoulos et al., 2024; Fu et al., 2024).

[1] Throughout this paper, we use *causality* and *causal formulation* interchangeably, referring to the causal factorization underlying autoregressive language models. They are not related to causal inference tasks involving interventions or counterfactuals here.

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.

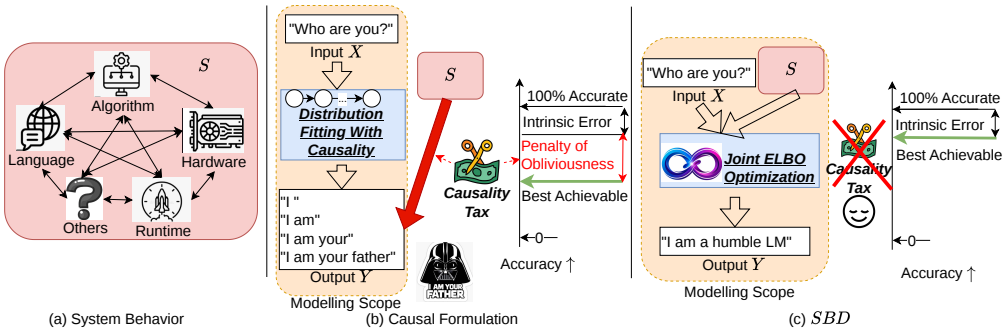


Figure 1: Overview of SBD.

Definition 1.1 (System Behavior)**.** In this paper, system behavior (S) refers to the union of all transformation factors beyond the data distribution in LMs, including algorithms, hardware, runtime, natural languages, and other yet-to-be-understood factors (Figure 1(a)).

Causality focuses on distribution alignment via a sequential dependency chain, e.g., "I" $\rightarrow$ "I am" in Figure 1(b). However, it remain *oblivious* to how S governs the generation process (red arrows in Figure 1(b)). Intuitively, obliviousness leads to a penalty, such as inaccurate or even offensive outputs (Ouyang et al., 2025; Betley & et al., 2025) (e.g., "I am your father"). Formally, causality may degrade the best achievable accuracy. The resulting error can exceed the intrinsic approximation error of the neural architecture due to the aforementioned mismatches. Therefore, a fundamental question arises as:

How can language modelling explicitly incorporate S for the reconciliation, i.e., to identify, explain, and reduce these systematic penalties?

Incorporating an often-overlooked space is usually harder than improving a well-established one. The formulation needs to be broad enough to give all factors beyond data a meaningful and interpretable role. It also needs to stay concrete enough to guide practical improvements, rather than becoming overly abstract. We propose the **S**ystem **B**ehavior **D**ynamics (SBD), a Bayesian framework to bridge the gap. Bayesian framework (Kingma & Welling, 2013) introduces *latent variables* to capture the broad diversity without requiring rigid, manual specifications, and it utilizes the *Evidence Lower Bound (ELBO)* to be compatible with gradient-based optimization. This is different from point-estimates or functional constraints (Bertsekas, 2014), which suffer from mode collapse or optimization instability when faced with complex system behaviors considered in this work. We therefore treat S as *a set of latent variables*, and it is an irreducible component of the ELBO (Figure 1(c)). Such ELBO jointly considers the global consistency of S and output (e.g., "I am a humble LM"), and an expansion of it confirms that ***there is a structural overhead, which we refer to as Causality Tax, to enforce the causal chain itself.*** Furthermore, the causality tax is not inevitable. As shown by SBD, the ELBO remains globally consistent even without imposing a causal chain.

Despite the theoretical insights, validation and utilization of SBD in practice is still challenging, due to the less explicit nature of investigating latent variables compared with fitting data distributions (Blei et al., 2017). Specifically, a closed-loop latent variable analysis (Bollen, 2014) involves a technical question of ***how to measure, control, and evaluate the impacts of S within SBD?***

We address these detailed aspects in turn. First, we conduct a perturbation analysis for *measurement*. SBD establishes that S structurally shapes the output during inference beyond hardware or runtime randomness. Although some components of S are unknown or unobservable, its structural influence arises jointly from algorithms, hardware, runtime environments, and natural languages. Therefore, we inject Gaussian random noise into input embeddings, and examine the resulting changes in output logits and hidden states. This analysis reveals stable regularities in logit variability and hidden-state distances beyond simple noise propagation. Second, we *control* the grouping strategies of S to especially show that the Causality Tax can indeed be reduced or even avoided. We introduce Green Shell (GSH), a non-causal variational family inspired by Green's functions (Green, 1889), to replace causal formulations. GSH achieves a tighter theoretical error bound than its causality

counterparts by a spatial partition over S components. Instead of imposing a causal chain, such a spatial partition inherits a divide-and-conquer approximation from Green's functions and thus gets rid of the sequential dependencies. Finally, Neural Tangent Kernel (NTK) spectra in the lazy-training regime (Jacot et al., 2018) offer a rigorous framework for model analysis, and we empirically validate causality-tax-reduction claims of `GSH` through NTK *evaluation*. Our evaluation also suggests that `GSH` may support the development of new base model families. In particular, its decoupled structure and reorganization of S suggest the potential for training language models without backpropagation.

Contributions. Our main contributions are as follows:

- We propose `SBD`, a new abstraction of S and treat it as a first-principle factor. (Section 3).
- We discover the Causality Tax by utilizing `SBD`, indicating that causality is neither necessary nor optimal. (Section 4).
- We establish the closed-loop validation of `SBD` beyond the theoretical claims through measurements, controls and evaluations around S. A proof-of-concept non-causal variational family `GSH` is proposed and examined during this establishment.(Section 5).

2 Related Work and Background

This section identifies a growing mismatch with causality through a literature review, followed by relevant Bayesian concepts. Discussions of more related work are shown in Appendix A.

Indication of Mismatches. Recent studies have indicated various mismatches with the causal formulation in autoregressive (AR) LMs. Empirically, non-determinism (Atil et al., 2024) is involved in inference, where identical inputs can lead to different outputs and fluctuating internal states. This phenomenon involves future states modifying past representations, while causality enforces the one-way determinism of the past to shape the future only. Theoretically, the Arrows of Time (AoT) framework (Papadopoulos et al., 2024) uncovers a fundamental paradox between causality and the symmetry of information (Shannon, 1951). Meanwhile, block diffusion (Arriola et al., 2025) demonstrates that the token-by-token causal chain itself can be reconstructed and shortened via a denoising process. On the practical side, inference optimizations such as speculative decoding (Leviathan et al., 2023) and lookahead decoding (Fu et al., 2024) deliberately relax strict token-by-token causality to improve efficiency, while still attempting to preserve the original probabilistic distribution.

Despite these valuable indications of mismatch, causality remains the foundational assumption of all the above works. In contrast, our work revisits the foundation of causality from first principles.

Bayesian Modelling, Variational Inference, and Evidence Lower Bound. There is a growing trend to treat the *system itself* as an active component of the behavioral model (Cidon et al., 2015; Lee et al., 2024; Zeng et al., 2024). Our work is the first to extend this to language modelling.

Bayesian modelling, as involved in PECJ (Zeng et al., 2024), offers a unified treatment of uncertainty. PECJ optimizes streaming database system operators under the so-called stream oscillation. Instead of modelling each source of randomness individually and passively reacting, it proactively compensates for oscillation by abstracting both observable and unobservable factors into latent variables and formulates a Posterior Distribution Approximation (PDA) problem. This is regardless of database deployment details such as hardware placement, network conditions, or request skew. Variational Inference (VI) (Ranganath et al., 2014; Broderick et al., 2013) is then utilized to make PDA tractable by approximating the posterior with a variational family and optimizing the Evidence Lower Bound (ELBO). For problems of moderate complexity, such as stream oscillation compensation in PECJ, both closed-form methods (Blei et al., 2016) and black-box approaches like BBVI (Ranganath et al., 2014) are applicable to optimize ELBO. As the problem complexity increases, such as image, text or video generation (Arriola et al., 2025), closed-form coordinate updates becomes increasingly difficult, and BBVI provides a more practical alternative.

3 System Behavior in Bayesian Modelling

In this section, we formulate `SBD` from a Bayesian statistics aspect. We first discuss the key differences regarding abstraction principles, then discuss the detailed meanings.

3.1 A SHIFT OF ABSTRACTION: ASCENDING THE LATENT VARIABLES

Causal Formulation: Latent Variables as Slaves of Distribution Fitting. Causality is tightly linked with the data distribution. Even in relaxed frameworks like block diffusion, this binding persists. Formally, they optimize the ELBO through sequential or structured causal generation. For block diffusion, this involves a latent corruption process $q(Z \mid X)$ and a reverse generation model $p_\theta(X \mid Z)$, yielding Eq. 1, where Z denotes latent variables of the causal chain in the corrupted space. This formulation also subsumes autoregressive generation as a special case when the corruption process is instantiated as successive token masking (Austin et al., 2021).

$$\log p_\theta(X) \geq \mathbb{E}_{q(Z|X)} \left[\log p_\theta(X \mid Z)\right] - \mathrm{KL}(q(Z \mid X) \| p(Z)), \tag{1}$$

Such data-driven formulations treat observed data X as fixed external evidence and evaluate quality primarily by reconstruction fidelity. However, many contradictions to causality discussed before arise **not** from the data distribution. For instance, non-determinism comes from *runtime* and *hardware* behaviors (Atil et al., 2024), AoT is due to several *algorithm* aspects, including computability, irreversibility of model architectures and operator implementations (Papadopoulos et al., 2024), and look-ahead decoding relies on an alternative *algorithm* paradigm of forwarding an LM (Fu et al., 2024). Besides, there are system-wide behaviors that mainly arise from the *natural language* itself and *potentially unknown factors*, also beyond a pure data distribution, such as reasoning collapses under complexity shifts (Shojaee et al., 2025), the misalignment caused by narrow finetune (Betley & et al., 2025), and the emergence of metacognitive self-monitoring (Didolkar et al., 2024) (Appendix A.2).

To Learn From PECJ: Latent Variable Usage $\perp$ Distribution Fitting. In PECJ, there is also a latent variable notation Z. Different from Eq. 1, it captures various environmental factors that induce uncertainty in the streaming database engine, such as network transmission delay and OS scheduling jitter. In other words, this is an **orthogonal factor** beyond data. However, even if explicitly recognized as an irreducible factor beyond distribution fitting, Z is still an intermediate structural rather than final objective. Therefore, it is eventually *marginalized* by a set of closed-form or BBVI techniques (Ranganath et al., 2014) in PECJ. While this kind of marginialization process may be more comprehensive than Eq. 1, its major focus is not on how and why these factors affect.

To Form the `SBD` Abstraction: Latent Variable S as the First Principle. Following PECJ, we reformulate the ELBO to characterize the global consistency rather than merely distribution fitting quality. However, instead of marginalizing latent variables, we consider S as the first principle. In this view, accurately modelling language generation requires accounting for three elements simultaneously: the input data X, the output data Y, and the system behavior S. Specifically, given a joint distribution $p(Y, S, X)$, there is a variational distribution $q(Y, S)$ to approximate the posterior distribution $p(Y, S|X)$. The intuition is that a system itself should be self-consistent, rather than just being a good approximation mechanism of data input and output. Taking the prior $\log p(X)$ as the overall evidence, this joint perspective leads to a system-behavior-aware ELBO as formalized in Eq. 2, and further motivates Proposition 3.1.

$$\mathrm{ELBO}_q = \mathbb{E}_{q(Y,S)} \left[\log p(Y, S, X)\right] - \mathbb{E}_{q(Y,S)} \left[\log q(Y, S)\right] \tag{2}$$

Proposition 3.1 (ELBO under `SBD`)**.** *By expanding the $p(Y, S, X)$ in Eq. 2, a further decoupled ELBO is derived as Eq. 3. The $\mathbb{E}_{q(Y,S)}$ notation is removed for $p(X)$, since there is no q involved.*

$$\textit{ELBO}_q = \log p(X) + \mathbb{E}_{q(Y,S)} \left[\log p(Y \mid X)\right] + \mathbb{E}_{q(Y,S)} \left[\log p(S \mid X, Y) - \log q(Y, S)\right] \tag{3}$$

3.2 DETAILED MEANINGS

We structurally discuss other terms in Proposition 3.1 in the following.

Token Alignment (TA) $\mathbb{E}_{q(Y,S)} \left[\log p(Y \mid X)\right]$. During training, since the target Y is a fixed ground-truth, its variational distribution collapses to a Dirac delta, reducing the term into standard cross-entropy (CE, $\log p(Y \mid X)$). During inference, Y becomes stochastic, and the expectation must be retained to integrate over the latent space of S.

More importantly, as a structural component of expectations, this term is expected to exhibit regular patterns beyond mere random noise. Since a well-trained language model has largely aligned its output distribution Y with the data during training, propagated deviations in the variational approximation

during inference is mostly with respect to S through this expectation. As a result, the TA term effectively encodes the joint influence of the input X and system behavior S, causing the hidden states to exhibit systematic, non-random modifications induced by S. We leverage this property as a probe to implicitly measure S in Section 5.1.

System Response to Input/Output (SRIO) $\mathbb{E}_{q(Y,S)}\left[\log p(S \mid X, Y)\right]$. This term reflects the coherence between the system behavior S and the generated output Y. It captures the model's inductive biases and expectations regarding its own structural dynamics, given a particular input-output pair. Such biases include convergence patterns, consistency of attention flows, and the plausibility of hidden state transitions. For simplicity, in a frozen LM without external memory, we approximately treat this term as constant and work with its expectation for both training and inference. We believe that explicitly expanding this term in settings involving memory mechanisms or agentic behaviors could significantly benefit subsequent optimization and interpretability of language models.

Approximation's Internal Semantics (AIS) $-\mathbb{E}_{q(Y,S)}\left[\log q(Y, S)\right]$. As the entropy of the variational distribution itself, AIS governs how the output generation process is internally organized and approximated within the system. AIS is fixed during model inference, as "how to conduct the approximation" is not changed, given the frozen model weights. On the other hand, model training is a BBVI optimization (Ranganath et al., 2014) over AIS. Specifically, the approximation is implicitly optimized through changing model weights, and loss functions serve as the proxy to guide the optimizations and reduce approximation errors. For most BBVI implementations, different reasonable proxy objectives induce the same error bound, regardless of the specific loss function. This error bound is fundamentally determined by the structural organization of $q(\cdot)$, which we further discuss in Section 4. Common loss functions, such as cross-entropy, have already provided reasonable proxy objectives, but are not the only viable choices. More details are discussed in Appendix D.3.

4 Discovery of the Causality Tax

In this Section, we demonstrate how the Causality Tax is revealed through a further derivation.

4.1 Causality is Not Necessary

AIS itself contains no explicit left-to-right generation order or fixed autoregressive structure (Proposition 3.1), and causality only re-emerges as a byproduct of specific structural assumptions. A canonical example is diffusion models, where system behavior S is grouped into a thermal-noising trajectory $S = \{s^0, s^1, \ldots, s^T\}$, and the AIS takes the factorized form shown in Eq. 4.

$$q(Y, S) = q(Y|s^0, s^1, \ldots, s^T) \times \prod_{\text{For each step } i \neq 0} q(s^i \mid s^{i-1}) \times q(s^0) \tag{4}$$

In this case, each s^i represents the system state at the i-th denoising step dependent on the previous step $i-1$ by definition. A similar phenomenon occurs in autoregressive models, which replace $q(s^i \mid s^{i-1})$ with a full historical token dependence $q(s^i \mid s^{i-1}, \ldots, s^0, X)$, and conducts the cumulative product by "For each token i ". While such factorization can be effective, they impose strong causal structural constraints on the approximation of S. In contrast, by treating the AIS variational distribution $q(Y, S)$ as a free design choice rather than a fixed factorization, we can achieve more flexible and potentially superior control of approximation error. As long as the joint system converges to a high-probability configuration, the approximation remains valid.

4.2 Causality is Not Optimal

An Intuitive Lemma to Claim Suboptimal Bound. Let $f(X)$ denote the target output distribution $p(Y \mid X)$ given input X. Let $F_{\text{non}}(X)$ and $F_{\text{gen}}(X)$ be neural approximation mechanisms from the general variational inference family Ω_{gen} and the causal family Ω_{causal}, respectively. By the *monotonicity of the supremum* (Rudin, 2021), we observe Ω_{causal} is not optimal. Since the causal variational family Ω_{causal} is a strict subset of the general family Ω_{gen} (i.e., $\Omega_{\text{causal}} \subsetneqq \Omega_{\text{gen}}$), any upper bound on $\{\mathcal{L}(q) : q \in \Omega_{\text{gen}}\}$ is also an upper bound on $\{\mathcal{L}(q) : q \in \Omega_{\text{causal}}\}$. By instantiating $\mathcal{L}(q)$ as the KL divergence and substituting into Eq. 3, Lemma 4.1 follows immediately.

A Non-trival Lemma to Reveal *Causality Tax*. Lemma 4.1 is equivalent to Lemma 4.2, which is less intuitive but derived by a set of functional integration and inequality scaling. Due to the page limits, we only discuss the results here, and please refer to Appendix B for a detailed derivation.

Compared with Lemma 4.1, the non-trival derivations behind leads to a specific term of *Causality Tax*, which reveals that *the causal chain itself leads to the additional error*. By Universal Approximation Theorem (UAT), the general family achieves the approximation error bound in Eq. 6, which reflects its ability to approximate the target Y distribution directly. For the causal family with equivalent parameters and imposed with T-order causal chain, however, the UAT bound becomes strictly looser due to an additional residual term $\sum_{t=1}^{T} |t_t(X)|$ (Eq. 7). As derived in Eq. 11 (Appendix B), this residual term arises from the extra effort required to enforce the causal chain. Furthermore, each $t_t(X)$ is a hierarchy integration over all possible realizations of the X distribution, and the summation puts the guarantee at each order t of the causal chain together.

Lemma 4.1 (Expressivity Bound by Monotonicity of the Supremum)**.**

$$\max(\mathrm{ELBO}_{q\in\Omega_{\mathrm{gen}}}) \;\geq\; \max(\mathrm{ELBO}_{q\in\Omega_{\mathrm{causal}}}), \tag{5}$$

Lemma 4.2 (Expressivity Bound by UAT)**.**

$$|F_{\mathrm{gen}}(X) - f(X)| < \epsilon \tag{6}$$

$$|F_{\mathrm{causal}}(X) - f(X)| < \epsilon + \sum_{t=1}^{T} |t_t(X)| \tag{7}$$

F_{gen} always achieves smaller approximation error than F_{causal}, with strict inequality if $\exists t_t(X) \neq 0$.

5 Validation and Analysis of SBD

In this Section, we verify SBD through measurements, controls, and evaluations.

5.1 Measurement on the Impacts of S

We perform a perturbation analysis for the measurement, which indicates S is far more complicated than hardware and runtime randomness.

Perturbation Setup. We run standard autoregressive decoding on a frozen LM and record the logits drift sequence E associated with input tokens X across N generated steps, and a similar tracing is conducted on the last hidden states, denoted as the r sequence. More intuitively, the perturbation analysis can be viewed as probing a pizza factory pipeline consisting of three manufacturing stages: "Crusting $\rightarrow$ Topping $\rightarrow$ Baking", where padding noise is analogous to deliberately introducing spoiled caviar at the very beginning of the production line. If the pipeline is merely a passive production process (analogous to S consisting only of hardware/runtime noise), the contaminated ingredients before Baking should simply produce proportionally worse pizzas, leading to a strong correlation of $E \propto r$. In contrast, if the production process contains hidden mechanisms that actively respond to the ingredients (analogous to the language- and algorithm-dependent dynamics within S), the resulting degradation will deviate from this simple proportional relationship. A formal definition of each term is shown in Appendix C.

Measurement Results. We conduct the measurement over varies LLMs (Yang et al., 2025; Dubey et al., 2024; Guo et al., 2025; Shen et al., 2024). Please refer to Figure 5 at Appendix C for more details. In general, the probing results reveal clear systematic patterns. First, the correlation between E and $1/r$ exceeds 0.6 for all models and reaches above 0.9 for some models with longer prompts, indicating a strong non-random relationship. On the other hand, the correlation with r is consistently lower (i.e., below 0.4), indicating the linear propagation of random noise is weaker than other structural factors, such as some inverse functional. Furthermore, once the prompt length exceeds a certain threshold, the E-$1/r$ correlation shows a sharp transition toward unity for some models. In addition, the lag between E and $1/r$ typically deviates from -1, indicating that TA term is not a simple instantaneous effect but involves a structured temporal propagation. In summary, the fluctuations during inference exhibit measurable regularities beyond random noise. This arises from system behavior S that systematically influences token alignment (Proposition 3.1).

5.2 CONTROLS OVER S GROUPING

Overview. To demonstrate that one can reduce or remove the Causality Tax through controls, we propose GSH (**G**reen **SH**ell), a non-causal mean-field variational family. The key insight of Green-Shell methods (Jackson, 2021; Green, 1889) in physics is to replace a complicated electromagnetic or potential field with a set of concentric shells as approximation units. Instead of tracking every fine-grained interaction, one only needs to compute how charges or sources are distributed across these shells. In this way, Green-Shells approximate a complex system through spatial partitioning rather than temporal dependency, making each approximation unit inherently *shared-nothing*. Inspired by these, GSH factorizes AIS as Eq. 8 and partitions the S into s^i by concentric shells instead of a causal chain. It still follows $S = \{s^0, s^1, \ldots, s^T\}$, but T here refers to the shell index.

Instead of instantiating a denoising chain $q(s^i \mid s^{i-1})$ (diffusion) or a full historical token dependence $q(s^i \mid s^{i-1}, \ldots, s^0, X)$ (autoregressive), GSH instantiates $q(s^i)$ to align with field propagation in the Green-Shell structure, including: (1) a temporal ordering constraint, requiring that the processing of shell i occurs between shells $i-1$ and $i+1$ (the arrow in Figure 2); and (2) a spatial partition constraint, requiring that shell i occupies its own region of the hidden-state space (the layer partition at Figure 2, as concisely defined in Eq. 16 and Appendix D). Given the spatial partition over S and properties of Green-Shells, (3) the hidden-state space of Y is also equivalently represented as spatial partitions $Y = y^0 + y^1 + ... + y^T$, and each y_i only covers the radius $R_{i-1} < r \leq R_i$ without overlap with others (partition of output in Figure 2).

Different from Eq. 4, each shell i performs an independent and unconstrained approximation of its local target y^i through $q(s^i)$, without causal dependence on other shells s^j (for $j \neq i$) or the input sequence X. In this way, GSH incurs no additional structural approximation error beyond the intrinsic UAT bound ϵ (Eq. 6) (Hornik et al., 1989; Bishop & Nasrabadi, 2006), in the sense that (1) the temporal ordering constraint corresponds to a deterministic execution graph inside runtimes like PyTorch (Pyt, 2023); (2) the variational components $q(s^i)$ without causal chain dependency (e.g., $q(s^i|s^{i-1})$ in Eq. 4) are *sufficiently expressive in the infinite-width/depth regime of UAT*, thereby avoiding causal-chain induced terms $|t_t(X)|$ in Eq. 7); and (3) approximation error of Y can be localized as $q(y^i|s^i)$, rather than being always propagated through the coupled $q(Y|s^0, s^1, \ldots, s^T)$. Please refer to Appendix D.2 for more details.

$$q(Y, S) = q(Y|s^0, s^1, \ldots, s^T) \times \prod_{\text{For each layer } i} q(s^i) = \prod_{\text{For each layer } i} (q(s^i) \times q(y^i|s^i)) \tag{8}$$

It is worth noting that the consistently directional field propagation preserves essential temporal and sequential information about S. Therefore, GSH avoids both the collapse into a bag-of-words model and the rigid constraints of causality.

Loss Function and Parametrization. A critical challenge in designing a loss function and parameterizing GSH is that the mean-field of hidden-state and vocabulary tokens are *not naturally aligned*. This leads to two main strategies: discretizing continuous fields into discrete tokens, or embedding discrete tokens into continuous field. Both involve non-trival derivations as detailed in Appendix D.3, and we briefly illustrate key insights in the following.

In the discrete approach, the variational distribution becomes a categorical distribution defined by a non-negative scoring function. Cross-entropy loss corresponds to one specific parametrization using exponential scoring and softmax normalization. Richer scoring schemes exist, such as label smoothing (Szegedy et al., 2016). Recent works (Kalai et al., 2025; Peters et al., 2020; Tang et al., 2024) have also pointed out inherent limitations of cross entropy and suggested a broader design space beyond softmax. In the continuous approach, the variational distribution is expressed by an energy functional, such as a Green Kernel (Eq. 28 at Appendix D), and a simpler quadratic energy leads to a mean-squared-error (MSE) loss. This formulation preserves the geometry of the field and offers numerical stability, but may discard richer structural dynamics. More sophisticated energy functions (LeCun et al., 2006; Zhang et al., 2023) are left for future exploration.

A complete removal of causality is incompatible with state-of-art infrastructures, and we leave this for future work. In the current proof-of-concept, we adopt a block-wise strategy similar to block diffusion (Arriola et al., 2025) to be compatible with KV caches. There is still causality among

Table 1: Overall Comparison.

Later Phase				Earlier Phase			
Dataset	Method	SNR (dB, ↑)	MLC (↓)	Dataset	Method	SNR (dB, ↑)	MLC (↓)
TinyStories	AR	12.8	0.94	AG News	AR	19.5	0.96
	BD	15.8	0.92		BD	15.6	0.85
	GSH	**22.5**	**0.72**		GSH	**28.3**	0.95
MultiEURLEX	AR	17.0	0.91	C4	AR	17.3	0.93
	BD	13.4	0.89		BD	15.1	0.91
	GSH	**24.0**	**0.73**		GSH	**32.2**	0.89

different blocks, but for each block inside, we replace the diffusion generation of each block with non-causal GSH approximation. Please refer to Appendix D.4 for more implementation details.

5.3 Evaluations about S under NTK

We evaluate on whether and how S is optimized by GSH under the lens of Neural Tangent Kernel (NTK) (Jacot et al., 2018), as well as examining the theoretical claims about error bounds at Section 4. Please also refer to Appendix G for additional sensitivity and ablation studies.

NTK Proxy of Error Bound. Directly verifying the UAT error bound at Lemma 4.2 is extremely difficult in practice, and common end-to-end metrics (such as perplexity or accuracy) cannot be directly linked to the theoretical bound. Fortunately, in the *lazy training* regime, prior work has shown that the spectral properties of the NTK can be rigorously connected to generalization and approximation bounds (Bartlett & Mendelson, 2002; Caponnetto & De Vito, 2007; Martin & Mahoney, 2021). However, this connection breaks down once the model enters the full feature-learning regime. We therefore conduct all evaluations in the lazy training regime to maintain theoretical validity. Under this setting, we utilize the **Signal-to-Noise Ratio** (SNR) of the NTK spectrum as an approximately *monotonic* proxy for the UAT error bound. Formerly, SNR is defined as the energy ratio of the *signal subspace* to the *noise bulk* in the NTK spectrum (Martin & Mahoney, 2021), and a higher SNR corresponds to a tighter approximation bound (Appendix E). Following standard practices in spectral analysis, we report SNR in decibels (dB), calculated as $10\log_{10}()$.

In addition to SNR, we introduce another spectral metric called the **Main Lobe Concentration** (MLC). For a given NTK spectrum, we define the main lobe as the set of eigenvalues consisting of the largest eigenvalue together with all subsequent eigenvalues whose energy is at least 20% of the largest one. The MLC is then computed as the ratio of the total energy contained in this main lobe to the total energy of the entire signal subspace. A healthy NTK spectrum is expected to capture not only dominant features but also rich peripheral (edge) features. Therefore, a lower MLC is generally preferable, as it indicates that the signal energy is more distributed across multiple significant eigenvalues rather than being overly concentrated in a few dominant modes.

Other Setups. We compare GSH against (1) the classical **autoregressive (AR)** model (Vaswani et al., 2017) and (2) the recent **block diffusion (BD)** model (Arriola et al., 2025). We utilize four datasets, i.e., TinyStories (Eldan & Li, 2023), AG News (Zhang, 2015), MultiEURLEX (Chalkidis et al., 2021) and C4 (AllenAI, 2020). Please refer to Appendix F for more details.

General Comparison. The experimental results confirm our theoretical predictions. As shown in Table 1, GSH consistently achieves the best across all datasets. This provides direct empirical evidence that the non-causal variational family attains a tighter approximation error bound, as theoretically predicted in Section 4. In the earlier stage, GSH starts with an extremely high SNR and highly concentrated main lobe, reflecting a conservative initialization. In the later stage, where the model has learned richer representations and covered more signals, GSH not only obtains the highest SNR but also exhibits the lowest MLC. The detailed NTK spectra comparison at C4 and TinyStories (Figure 3) show that GSH exhibits a "lifting force" during the transition from earlier to later. Such force compresses the main lobe and fills the mid-spectrum, leading to a more balanced signal distribution that suppresses noise without losing feature richness. Eventually, the signal energy is more effectively distributed across multiple significant eigenvalues. Therefore, besides the proven tighter error bound, non-casual variational family may also enjoy additional advantages in generalization capability.

Bonus: Possibility of Backward Propagation Free. There are 4 transformer layers labeled $0 \sim 3$ and one linear LM Head in our backbone NN, and we investigate the effect of gradient backward

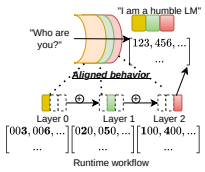
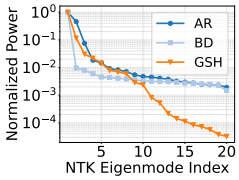
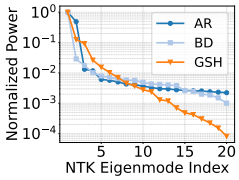


Figure 2: GSH.



(a) C4



(b) TinyStories

Figure 3: Detailed NTK spectra.

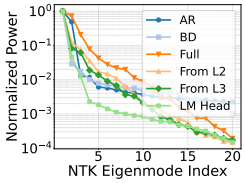


Figure 4: Resulting NTK spectra of gradient cutoff.

propagation cutoff from a certain layer in Figure 4. In this cutoff setting, the front part involves each layer training independently (i.e., only step 1 at Appendix D.4), while the rest still conducts backward propagation across layers (i.e., both steps at Appendix D.4). All other settings remain consistent with Figure 3-(b). In general, GSH outperforms state-of-the-art if the gradient cutoff occurs from the transformer layers, with stable SNR $\approx 22dB$ and MLC ≈ 0.8. The only significant downgrade occurs when the gradient cutoff starts at LM Head, and energy is too much concentrated at the first eigenvalue (MLC =0.94). Such possibilty of gradient cutoff is due to the layer decoupling nature of GSH (Section 5.2), and further implies that backward-propagation-free training (Li et al., 2025) is not confined to small networks but is also achievable in larger LMs.

6 Discussions

In this section, we discuss the limitations, potential risks of misuse and future work.

Limitations and Impact Statement. This work has four inherent limitations as follows. First, more detailed expansion of each S component is left for future work. Second, the Lemmas rely on the universal approximation theorem or the monotonicity of the supremum. Third, the non-causal instantiation GSH depends on BBVI and thus inherits its limitations. Finally, our experiments are intended to validate the theory rather than demonstrate end-to-end practical performance. Besides, given the first-principle changes, SBD inevitably opens new attack surface to LLMs, such as emergent misalignment (Betley & et al., 2025). We discuss the details in Appendix I.

Future Work. SBD and GSH align with more advanced LLM base models with less overhead. While model training incurs a substantial device memory footprint to maintain the complete computational graph for gradients, non-causal variants like GSH offer a principled mechanism to alleviate this overhead. They not only yield tighter error bounds, but also eliminate the need to retain the full gradient graph in device memory. Given a language model with N layers, GSH offers the opportunity to reduce the memory footprint of training from $O(N)$ to $O(1)$, as indicated by the gradient cutoff in Figure 4. In addition, given the theoretical down gradation of causality, LLM infrastructures may explore better reuse and approximate computing in both prefilling (Jin et al., 2024; Gim et al., 2024) and decoding (Cai et al., 2024). Speculatively, the similar Bayesian foundations of SBD and PECJ (Zeng et al., 2024) may extend naturally toward a further joint optimization in AI and DB systems. Please refer to Appendix H for more details.

7 Conclusion

We revisited the theoretical foundations of causality in language models from a Bayesian perspective. Through SBD modelling, we showed that *causal formulation is what you may downgrade if you look beyond data distribution and incorporate system behavior*. This is reaffirmed by a further validation and analysis regarding measure, control and evaluation of the latent variables. During the optimized control of approximation errors, a realizable and promising non-causal variational approximation prototype emerges as GSH. Beyond these immediate findings, this work calls for a broader shift from fitting data distributions under causal chains to explicitly characterizing system behaviors themselves as Bayesian latent variables (Definition 1.1). We believe that modelling system behavior offers a unifying theoretical lens for understanding stability, efficiency, and adaptation in generative systems.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing and debugging detailed Python implementations. We have not used generative AI tools for key idea framing, mathematical derivations, or retrival of related works. LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

ETHICS STATEMENT

Potential privacy and security issues are discussed in Appendix I, and there are no other kinds of ethics concerns.

REPRODUCIBILITY STATEMENT

We discuss more detailed setups in Appendix F. The code, scripts, and documents to reproduce all tables and figures are submitted as *Supplementary Material* along with this paper draft, and we plan to make it an open-source repo later.

REFERENCES

Pytorch homepage, `https://pytorch.org/`, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AllenAI. C4: Colossal clean crawled corpus. `https://huggingface.co/datasets/allenai/c4`, 2020. Hugging Face dataset.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, et al. Non-determinism of" deterministic" llm settings. *arXiv preprint arXiv:2408.04667*, 2024.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Jan Betley and Tan et al. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. In *International Conference on Machine Learning*, pp. 4043–4068. PMLR, 2025.

Charles K Birdsall and A Bruce Langdon. *Plasma physics via computer simulation*. CRC press, 2018.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

David Blei, Alp Kucukelbir, and Jon McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112, 01 2016. doi: 10.1080/01621459.2017.1285773.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Kenneth A Bollen. *Structural equations with latent variables*. John Wiley & Sons, 2014.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia Wilson, and Michael Jordan. Streaming variational bayes. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/51ef186e18dc00c2d31982567235c559-Paper.pdf.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. In *International Conference on Machine Learning*, pp. 5209–5235. PMLR, 2024.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*, 2021.

Asaf Cidon, Robert Escriva, Sachin Katti, Mendel Rosenblum, and Emin Gun Sirer. Tiered replication: A cost-effective alternative to full cluster geo-replication. In *2015 USENIX Annual Technical Conference (USENIX ATC 15)*, pp. 31–43, 2015.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812, 2024.

Zhichen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, and Chaochao Lu. Emergent response planning in llms. *arXiv preprint arXiv:2502.06258*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Ju Fan, Zihui Gu, Songyue Zhang, Yuxin Zhang, Zui Chen, Lei Cao, Guoliang Li, Samuel Madden, Xiaoyong Du, and Nan Tang. Combining small language models and large language models for zero-shot nl2sql. *Proceedings of the VLDB Endowment*, 17(11):2750–2763, 2024.

Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. Break the sequential dependency of llm inference using lookahead decoding. In *International Conference on Machine Learning*, pp. 14060–14079. PMLR, 2024.

Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. doi: 10.1109/TIT.2014.2323359.

In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *International Conference on Machine Learning*, pp. 15706–15734. PMLR, 2024.

George Green. *An essay on the application of mathematical analysis to the theories of electricity and magnetism*, volume 3. Mayer & Müller, 1889.

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.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

John David Jackson. *Classical electrodynamics*. John Wiley & Sons, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. Language models resist alignment: Evidence from data compression. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23411–23432, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1141. URL https://aclanthology.org/2025.acl-long.1141/.

Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. Ragcache: Efficient knowledge caching for retrieval-augmented generation. *arXiv preprint arXiv:2404.12457*, 2024.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint*, arXiv:2509.04664, 2025. URL https://arxiv.org/abs/2509.04664. Submitted on 4 Sep 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. Infinigen: Efficient generative inference of large language models with dynamic kv cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 155–172, 2024.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Qinyu Li, Yee Whye Teh, and Razvan Pascanu. Noprop: Training neural networks without back-propagation or forward-propagation. *arXiv preprint arXiv:2503.24322*, 2025.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Ivan V Oseledets and Eugene E Tyrtyshnikov. Breaking the curse of dimensionality, or how to use svd in many dimensions. *SIAM Journal on Scientific Computing*, 31(5):3744–3759, 2009.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–28, 2025.

Vassilis Papadopoulos, Jérémie Wenger, and Clément Hongler. Arrows of time for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39509–39528. PMLR, 21–27 Jul 2024.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 1504–1519. Association for Computational Linguistics (ACL), 2020.

Max Planck. On the theory of the energy distribution law of the normal spectrum. *Verh. Deut. Phys. Ges*, 2(237):237–245, 1900.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander (eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL https://proceedings.mlr.press/v33/ranganath14.html.

Walter Rudin. Principles of mathematical analysis. 2021.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1): 50–64, 1951.

Yikang Shen, Matthew Stallone, Mayank Mishra, Gaoyuan Zhang, Shawn Tan, Aditya Prasad, Adriana Meza Soria, David D Cox, and Rameswar Panda. Power scheduler: A batch size and token number agnostic learning rate scheduler. *arXiv preprint arXiv:2408.13359*, 2024.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *Advances in neural information processing systems*, 36:27801–27819, 2023.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4195–4206, 2025.

Eric Stern, Robin Wagner, Fred J Sigworth, Ronald Breaker, Tarek M Fahmy, and Mark A Reed. Importance of the debye screening length on nanowire field effect transistor sensors. *Nano letters*, 7(11):3405–3409, 2007.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. Learning geometry-aware representations for new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5641–5654, 2024.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in LLMs through information theory. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=IjOWms0hrf.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Richard Turner, Pietro Berkes, and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time Series Models*, 01 2008. doi: 10.1017/CBO9780511984679.006.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhilin Wang, Yafu Li, Jianhao Yan, Yu Cheng, and Yue Zhang. Unveiling attractor cycles in large language models: A dynamical systems view of successive paraphrasing. *arXiv preprint arXiv:2502.15208*, 2025.

Igor Weinberg and Elsa Ronningstam. Narcissistic personality disorder: Progress in understanding and treatment. *Focus*, 20(4):368–377, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. Cacheblend: Fast large language model serving for rag with cached knowledge fusion. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 94–109, 2025.

Marcin Zajenkowski and Gilles E Gignac. Telling people they are intelligent correlates with the feeling of narcissistic uniqueness: The influence of iq feedback on temporary state narcissism. *Intelligence*, 89:101595, 2021.

Xianzhi Zeng, Shuhao Zhang, Hongbin Zhong, Hao Zhang, Mian Lu, Zhao Zheng, and Yuqiang Chen. Pecj: Stream window join on disorder data streams with proactive error compensation. *Proceedings of the ACM on Management of Data*, 2(1):1–24, 2024.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. Interpreting and improving large language models in arithmetic calculation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 59932–59950, 2024.

Xiang Zhang. Ag news dataset. `https://huggingface.co/datasets/ag_news`, 2015. Hugging Face dataset.

Xinwei Zhang, Zhiqiang Tan, and Zhijian Ou. Persistently trained, diffusion-assisted energy-based models. *Stat*, 12(1):e625, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

Xuanhe Zhou, Chengliang Chai, Guoliang Li, and Ji Sun. Database meets artificial intelligence: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1096–1116, 2022. doi: 10.1109/TKDE.2020.2994641.

A More Discussions of Related Works

A.1 Detailed Mismatches with Causality

Empirical Evidence. It is show that even under strictly deterministic settings, model outputs can vary across runs (Atil et al., 2024). Furthermore, future generation steps can retroactively influence past representations, such as embeddings and logits of generated tokens. This *non-determinism* is typically framed in the literature as undesirable randomness or instability of runtime and hardware (Atil et al., 2024; Song et al., 2025; Ouyang et al., 2025), and it is commonly treated as a bug to be corrected through numerical precision tuning, scheduling constraints, or infrastructure-level fixes.

Theoretical Contradictions. The *Arrows of Time (AoT)* (Papadopoulos et al., 2024) reveals a theoretical paradox that LLMs exhibit asymmetric perplexity when predicting the next token versus the previous one. This asymmetry contradicts the fundamental symmetry assumed in classical information theory (Shannon, 1951). Proposed explanations, such as sparsity and computational asymmetry, are outside the scope of causal formulation itself and belong to algorithm aspects.

Another notable challenge arises from block diffusion (Arriola et al., 2025). While still using causality, it offers a relaxation of the autoregressive (AR) causal chain. Instead of relying on token-by-token causality, block diffusion conducts a causal denoising process for each token block. Although it

has not yet reached the scale of state-of-the-art AR models like (Yang et al., 2025; Touvron et al., 2023; Achiam et al., 2023), block diffusion has demonstrated improvements in global consistency and decoding speed.

Implementation Strategies. LLM inference typically consists of two phases, namely *prefill* and *decoding*. In both phases, recent strategies increasingly diverge from strict causality by involving new algorithm paradigms. During prefill, causality implies that key-value (KV) cache reuse should follow strict prefix sharing. However, techniques like modular cache slicing (Gim et al., 2024; Yao et al., 2025) and prompt reconstruction (Fan et al., 2024) allow relaxed reuse in specific domains, such as retrieval-augmented generation, thereby softening causal constraints. In the decoding phase, techniques like speculative decoding (Leviathan et al., 2023) and lookahead decoding (Fu et al., 2024) predict multiple tokens at once, which effectively breaking the causal step-by-step progression. More recently, multi-token prediction strategies (Gloeckle et al., 2024; Cai et al., 2024) have emerged, training auxiliary heads or modules to propose multiple candidate tokens per step for parallel decoding.

A.2 Cognitive and Language Semantic Investigations

While SBD highlights the irreducible impact of system behavior rather than relying solely on data-driven alignment, we are not the first to explore modelling beyond data distributions in the context of LLMs. Specifically, several recent studies report higher-order LLM behaviors that cannot be fully explained by data-distribution alignment, and most of them are linked with the language and cognitive capabilities themselves. Examples include the emergence of metacognitive self-monitoring in mathematical reasoning (Didolkar et al., 2024), targeted representation interventions that boost arithmetic accuracy (Zhang et al., 2024), and the *illusion of thinking* phenomenon, where apparent reasoning collapses under complexity shifts (Shojaee et al., 2025). Other findings similarly highlight system-level tendencies: narrow fine-tuning on adversarial tasks can induce surprisingly broad misalignment effects (Betley & et al., 2025), and chain-of-thought reasoning can be decomposed into an information-theoretic process of incremental information gain, revealing hidden failure modes without explicit reasoning annotations (Ton et al., 2025).

We emphasize that it is not possible to exhaustively enumerate all such discoveries here, nor is our intention to challenge, rebuild, or extend any specific line of work. Instead, we regard these diverse findings collectively as a technique justification for modelling beyond data, as they point to persistent and structured tendencies in the language part of system behavior that are not reducible to surface-level output patterns in token distributions. While SBD is not intended to fully explain these phenomena, framing them together in future work may offer a principled way to study the deep mechanisms shaping LM behavior beyond token-level accuracy. Such an orthogonal perspective may eventually unify these seemingly disparate phenomena under a coherent dynamical framework.

A.3 Modelling Behaviors and Strategies of LMs

A growing body of work seeks to model or reverse-engineer how moren LM, especially LLMs, organize and deploy internal computational strategies. Among these behaviors, one prominent theme is *emergent abilities*, where capabilities absent in smaller models appear abruptly at scale (Zhao et al., 2023), though some argue this arises from nonlinear evaluation metrics rather than genuine behavioral shifts (Schaeffer et al., 2023). Recent studies broaden this view, showing that LLMs can spontaneously form multi-step contingency plans without explicit supervision (Dong et al., 2025).

Another line of work revisits *in-context learning* (ICL), revealing multiple coexisting mechanisms, i.e., pure ICL, in-weights learning (IWL), and context-constrained IWL (CIWL) that compete and cooperate during training, with dynamic strategy switching explaining phenomena such as ICL transience (Singh et al., 2023).

Beyond these training- or observation-centric analyses, Wang et al. (2025) adopts a dynamical-systems lens, modelling inference as a discrete-time process and uncovering stable periodic patterns (attractor cycles) in output space across prompts, models, and decoding settings. Similarly, Ji et al. (2025) introduces the concept of *elasticity* in LLMs, showing that alignment effects are fragile and that models rebound toward pre-training distributions under small perturbations. Compression theory

further links this resistance to Hooke's law, suggesting that alignment effects are often superficial rather than deeply altering internal representations.

Our work is inspired by these insights but makes a major departure. Rather than treating system behaviors as byproducts of scale, data, or architecture (Zhao et al., 2023; Schaeffer et al., 2023; Dong et al., 2025; Singh et al., 2023; Ji et al., 2025), we posit that *system behaviors themselves constitute part of the model*. This view is also more general and system-centric than Wang et al. (2025), where the system serves primarily as an analytical metaphor rather than as a directly modelled entity.

B Detailed Derivations of Lemma 4.2

For the general family, the Universal Approximation Theorem (UAT) directly yields a tight error bound as shown in Eq. 9. For the causal family, the conditional independence constraints is additional imposed, as formalized in Eq. 10, introducing an additional residual term $t_t(X)$ defined in Eq. 11. This residual term quantifies the deviation caused by enforcing conditional independence at each step of the causal chain. Specifically, the integration of S (or equivalently over the decomposed internal states s) over X ensures that the causal factorization must hold *for all possible realizations* of the input data distribution X, regardless of its specific form. In other words, the model must satisfy the conditional independence not only on the observed data, but across the entire support of the input distribution.

To simultaneously satisfy these causal constraints while approximating the target distribution $f(X)$, the approximation error under the same parameter capacity becomes Eq. 12. Applying the triangle inequality then yields the final UAT bound for the causal family in Eq. 13, which is strictly looser than the non-causal case by the accumulated residual terms $\sum_{t=1}^{T} |t_t(X)|$. T is the length of the causal chain. For diffusion models like block diffusion, T equals to the number of denosing steps; for autoregressive models, T equals to the number of generated tokens.

$$|F_{\text{gen}}(X) - f(X)| < \epsilon \tag{9}$$

$$\begin{cases} \textbf{Diffusion:} & s_t \perp \{s_0, \ldots, s_{t-2}\} \;\big|\; (s_{t-1}, X), \\ \textbf{Autoregressive:} & s_t \perp \{s_{t+1}, \ldots, s_T\} \;\big|\; (s_{0:t-1}, X) \end{cases} \tag{10}$$

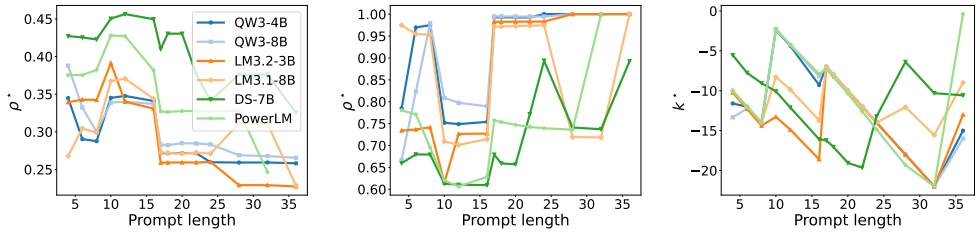
$$\begin{cases} \textbf{Diffusion:} t_t^{(\text{diff})}(X) = \\ \displaystyle\int p(s_t, \{s_0, \ldots, s_{t-2}\} \mid X) \\ \quad - p(s_t \mid X) p(\{s_0, \ldots, s_{t-2}\} \mid X)\, dX, \\ \textbf{Autoregressive:} t_t^{(\text{AR})}(X) = \\ \displaystyle\int p(s_t, \{s_{t+1}, \ldots, s_T\} \mid X) \\ \quad - p(s_t \mid X) p(\{s_{t+1}, \ldots, s_T\} \mid X)\, dX \end{cases} \tag{11}$$

$$|F_{\text{causal}}(X) - \sum_{t=1}^{T} t_t(X) - f(X)| < \epsilon \tag{12}$$

$$|F_{\text{causal}}(X) - f(X)| < \epsilon + \sum_{t=1}^{T} |t_t(X)| \tag{13}$$

C Formal Definition of the Perturbation Analysis

We run standard autoregressive decoding on a frozen LM and record the logits drift sequence $E = \{e_1, e_2, \ldots, e_N\}$ associated with input tokens X across N generated steps. Each e_i represents



(a) $\rho^\star$ of E and r. (b) $\rho^\star$ of E and $1/r$. (c) $k^\star$ of E and $1/r$.

Figure 5: Implicit Measurement of S

the incremental change in hidden states of the input tokens X at decoding step i. A perturbed input is then constructed by injecting Gaussian noise into the tail of the token embeddings of X. We collect the final-layer hidden states under this perturbation (H^{curr}) and compare them with the clean target hidden states H^{target}. The overall discrepancy is quantified by the scalar r, defined as the ℓ_2 norm of the element-wise Euclidean distances between H^{target} and H^{curr}. For simplicity, both e_i and r are scalars after summation. A measurement metrics is to compute the normalized cross-correlation between E and r (or $1/r$) as $\hat{R}_{E,r}(k) = \frac{\sum_{i=1}^{N} e_i\, r_{i+k}}{\sqrt{\sum_{i=1}^{N} e_i^2}\sqrt{\sum_{i=1}^{N} r_i^2}}$, where k is the lag. We then identify the maximum correlation value $\rho^\star$ and its corresponding best lag $k^\star = \arg\max_k \hat{R}_{E,r}(k)$.

We let QWen3-4B (*QW3-4B* for short) (Yang et al., 2025), QWen3-8B (*QW3-8B* for short) (Yang et al., 2025), Llama3.2-3B (Dubey et al., 2024) (*LM3.2-3B* for short), Llama3.1-8B (Dubey et al., 2024) (*LM3.1-8B* for short), DeepSeek-7B (Guo et al., 2025) (*DS-7B* for short), and PowerLM-3B (Shen et al., 2024) to generate 24 output tokens, while varying the length of input tokens from $2 \sim 36$. We use random samples from *TinyStories* (Eldan & Li, 2023) to form the input, and report the results in Figure 5.

D More Discussions of GSH

This section is a more detailed version of Section 5.2 with more derivations and justifications.

D.1 Mean Field from Mathematical Physics

$$q(Y, S) = \prod_{i \in \mathcal{I}} q_i(\text{variable}_i) \tag{14}$$

Mean-field formulations (Eq. 14) suggest that a family of tractable approximation mechanisms without assumed causality can in principle exist, enabling the conceptually free approximation of $q(i)$. However, the notion of a "mean field" is itself overly broad and under-specified. In this paper, we restrict the variational family to one that has long been used to analyze complex electromagnetic and potential systems, i.e., the so-called *Green-Shell*. In physics, Green-Shell methods (Jackson, 2021; Green, 1889) are widely used to approximate complex electromagnetic or potential fields. The basic idea is to replace a complicated continuous field with a set of concentric shells, each acting as an equipotential layer. Instead of tracking every fine-grained interaction, one only needs to compute how charges or sources are distributed across these shells. Compared to a preset causal chain, Green-Shells approximate a complex system through spatial partitioning rather than temporal dependency, making them inherently *shared-nothing* and fully consistent with the assumptions of the mean-field family (Eq. 14) . Moreover, the number of shells can be scaled arbitrarily, directly paralleling how neural networks expand their capacity with more layers or parameters. In the following, we demonstrate how this is applied to instantiate Eq. 14 and formulate GSH.

Assumptions and Intuitions. We begin by assuming that the output sequence Y can be expressed by a hidden-state field H defined with respect to a distance measure r in a perturbative analysis, i.e., $Y \propto H(r)$. We do not restrict a concrete form of r for two reasons: (1) distances in latent or embedding space can be freely projected or transformed by neural network mappings, and (2) r is

introduced only as a conceptual construct, and does not directly appear in the final loss formulation. Conceptually, the state space of H can use tensors of either logits or embeddings, but in this paper we restrict to embeddings that typically involves less dimensions than logits, due to the notorious curse of dimensionality (Oseledets & Tyrtyshnikov, 2009). Consequently, r is in the same dimension as embeddings.

Conceptual Set-ups. After conceptually treating Y as a field, we apply the following steps as what mathematical physics does. First, the target field $H(r)$ of output Y is divided into concentric shells as Eq. 15. Each component is individual from the others, and can be defined as Eq. 16. $\rho(\mathbf{r}', \theta^{(i)})$ is the density function of the field, and it will be learned during model training with parameter space θ^i of layer i.

For simplicity, we let the number of shells $T + 1$ equal to the number of transformers used in the neural network, and each layer is one-one mapped with each layer.

$$\mathbf{H}(\mathbf{r}) = \sum_{k=0}^{T} \mathbf{H}^{(k)}(\mathbf{r}) \tag{15}$$

$$\mathbf{H}^{(k)}(\mathbf{r}) = \int_{R_k < |\mathbf{r}'| < R_{k+1}} \rho(\mathbf{r}', \theta^{(i)}) \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} \, d^3\mathbf{r}' \tag{16}$$

Next, we extend the mean-field formulation in Eq. 14 into a field-theoretic variational family (Eq. 17). In analogy with diffusion models, $q(s^i)$ conceptually conducts the denoising process, and is trained through the noising process $q^{noising}(s^i, Y)$. However, unlike diffusion models, these processes are not anchored to the temporal order of X or Y; rather, they follow the intrinsic propagation of the field itself. In implementation, this corresponds to the neural network's forward noising and backward denosing passes, which operate without any predefined causal chain. Concretely, in the forward noising step (Eq. 18), each layer perceives the inside of its shell as a "clean system," while treating everything outside as Gaussian noise $\varepsilon^{(k)}$. This is analogous to a spherical cavity in electrostatics: the interior field seeks balance, while external disturbances only appear as noisy perturbations on the boundary. In this sense, the GSH instantiates a mean-field whose independence arises not from temporal factorization but from spatial partitioning.

$$q(Y, S) = \prod_{\text{For each layer } i} (q(s^i) \times q(y^i | s^i)) \tag{17}$$

$$q^{noising}(s^i, Y) \propto \sum_{k \le i} \mathbf{H}^{(k)} + \sum_{k > i} \varepsilon^{(k)} \tag{18}$$

Rationales Behind. It is important to note that this formulation contains no explicit notion of input–output temporal causality, as the variational family is not tied to the order of X or Y. However, this does not reduce the model to a bag-of-words abstraction. By construction, the propagation of fields proceeds layer by layer, preserving an intrinsic *directionality* analogous to temporal flow, even in the absence of an externally imposed sequence order. This directionality arises from the propagation of information through network depth and from the non-reversible dynamics of optimization and representation formation.

Recent evidence further supports this interpretation. The *Arrows of Time (AoT)* phenomenon (Papadopoulos et al., 2024) demonstrates that LMs exhibit an inherent temporal asymmetry, as internal activations and attention flows encode a preferred direction consistent with information compression and entropy increase, even when explicit token order is removed. In this sense, directionality is more *fundamental* than causality itself, as it arises not from pre-defined sequential rules, but from the irreversible flow of representation and optimization dynamics. *Therefore, causality is not discarded but subsumed in GSH.* What was previously interpreted as a left-to-right causal chain can instead be viewed as a manifestation of this emergent directional field.

D.2 GSH IS FREE FROM ADDITIONAL UAT RESIDUAL LOSS

As discussed in Section 5.2, GSH involves (1) the temporal ordering constraint, (2) the spatial partition constraint, and (3) shared-nothing components equivalent to output Y.

The first property introduces no additional approximation error, as modern deep learning frameworks (e.g., PyTorch) execute layers in a strictly sequential forward pass. From the perspective of UAT, this temporal ordering does not impose any additional functional restriction that would require introducing an extra residual term beyond the intrinsic approximation error ϵ (Eq. 6). From a probabilistic graphical model perspective, the forward pass corresponds to a deterministic computation, implying that the event of satisfying the prescribed evaluation order has probability one under the induced execution measure.

The approximation error induced by the second constraint is also subsumed into the UAT error bound ϵ. As a free approximation, each $q(s^i)$ does not constrain the UAT representational capacity in the infinite-width/depth limit, and is allowed to be *sufficiently expressive* (Hornik et al., 1989; Bishop & Nasrabadi, 2006) without restrictive structural assumptions (e.g., causal dependencies like $q(s^i|s^{i-1})$) or additional fitting residuals (e.g., Bayesian hierarchy $|t_t(X)|$ in Eq. 7). Under this regime, any discrepancy induced by the spatial partition constraint $p(s^i)$ can be absorbed through reweighting within $q(s^i)$ (Hornik et al., 1989; Bishop & Nasrabadi, 2006).

Furthermore, $q(y^i|s^i)$ ensures that approximation effects at each shell remain localized rather than being propagated through the coupled joint $q(Y|s^0, s^1, \ldots, s^T)$, leading to faster attenuation of prior-induced discrepancies under iterative updates (Jordan et al., 1999; Turner et al., 2008), and thus typically faster convergence of model training.

Consequently, in the absence of structural restrictions on $q(s^i)$, the overall approximation error bound remains determined by the intrinsic UAT error ϵ, without an additional residual term such as $|t_t(X)|$.

D.3 Mean Field into Training Loss

Eq. 19 is a general training objective of arbitrary mean-field, where $p(H)$ denotes the true (unknown) distribution of hidden-state fields, and $q_\theta(H)$ is the variational family parameterized by θ in approximating the H. For `GSH`, Eq. 19 naturally decomposes into layer-wise terms (Eq. 20) given Eq. 18, and each $\mathcal{L}_{\text{field}}^{(i)} = \mathbb{E}_{p(H^{(i)})}\big[-\log q_\theta^{(i)}(H^{(i)})\big]$ is computed locally on shell i.

$$\mathcal{L}_{\text{field}} = \mathbb{E}_{p(H)}\big[-\log q_\theta(H)\big] \tag{19}$$

$$\mathcal{L}_{\text{field}} = \sum_{i=0}^{T} \mathbb{E}_{p(H^{(i)})}\big[-\log q_\theta^{(i)}(H^{(i)})\big] \tag{20}$$

This mirrors the way Green-Shells approximate a physical field by concentric equipotential layers in Mathematical Physics: every shell contributes its own loss, and the global approximation emerges from their superposition. In this sense, the variational objective preserves the structural independence of shells while assembling them into a coherent system-level approximation.

Nevertheless, a critical challenge arises because hidden-state fields and vocabulary tokens are not naturally aligned. This leaves two possible directions: one can *discretize continuous fields into categorical indices*, or conversely, *embed discrete symbols into continuous field formulations*. Both views have their own justification and difficulties. Discrete fields align neatly with vocabulary-level modelling and allow exact normalization via cross-entropy, but at the cost of discarding the continuous geometry of hidden states. Continuous fields preserve the geometric and dynamical structure observed in system trajectories, yet they face the challenge of defining or approximating a tractable normalization. Within the Green-Shell framework, however, this tension is secondary: *whether discrete or continuous, the underlying task is mathematically a problem of field estimation.* Such dualities between continuous and discrete formulations are common in physics, where both perspectives may coexist and complement one another. In what follows (and in the Appendix), we briefly outline the discrete and continuous instantiations of this general objective.

Discrete Establishing. If the hidden-state field $H^{(i)}$ is discretized into a finite vocabulary $\mathcal{V}$, the variational distribution $q_\theta^{(i)}$ reduces to a categorical distribution (Eq. 21), where $f_\theta^{(i)}(\cdot)$ is an arbitrary nonnegative scoring function.

$$q_\theta^{(i)}(v) = \frac{f_\theta^{(i)}(v)}{\sum_{u\in\mathcal{V}} f_\theta^{(i)}(u)}, \quad v \in \mathcal{V} \tag{21}$$

This introduces a nontrivial design space: different parametrizations of $f_\theta^{(i)}$ correspond to different ways of discretizing fields. For example, one may use

- Additive or multiplicative decompositions of hidden states.
- Alternative normalizers beyond the softmax, such as sparsemax or entmax.
- Or structured mappings that reflect geometry in the field space.

To make the above discussion concrete, let us consider a particularly simple instantiation. Suppose we take the scoring function in Eq. 21 to be an exponential of a linear logit, where $z_v^{(i)}$ is directly produced by the network. This choice is admittedly a form of "taking the easy way out": it ignores possible geometric structure in the field, collapses all interactions into a single scalar score per token, and normalizes only through the exponential family. Substituting Eq. 22 into Eq. 21 yields the familiar softmax distribution (Eqn 23), and inserting Eq. 23 into Eq. 21 reduces the shell-wise objective to the cross-entropy as Eq.24.

$$f_\theta^{(i)}(v) = \exp(z_v^{(i)}), \tag{22}$$

$$q_\theta^{(i)}(v) = \frac{\exp(z_v^{(i)})}{\sum_{u\in\mathcal{V}} \exp(z_u^{(i)})} \tag{23}$$

$$\mathcal{L}_{\text{field}}^{(i)} = -\sum_{v\in\mathcal{V}} p(H^{(i)} = v)\, \log q_\theta^{(i)}(v) \tag{24}$$

In other words, *cross-entropy per layer is one particular instantiation within the broader field-estimation framework*, characterized by the exponential choice of $f_\theta^{(i)}$. It survives largely as a convenient but ultimately limited option, and while richer discretization schemes may lead to more faithful field approximations, exploring them lies beyond the scope of this work and is left for future investigation. Recent works like Kalai et al. (2025) also indicates the inherent suboptimal nature of cross-entropy from different aspects. Promising alternatives may involve non-softmax normalizers (Peters et al., 2020) or geometry-aware mappings (Tang et al., 2024), which suggest a broader design space than the exponential-softmax default.

Continuous Establishing. Alternatively, one may retain the continuous nature of the hidden-state field $H^{(i)}$. In this view, the variational family $q_\theta^{(i)}$ is naturally expressed as a Gibbs distribution (Eq. 25), where $\mathcal{H}_\theta^{(i)}(H)$ is an energy functional and $R_\theta^{(i)}$ is the intractable partition function. This formulation preserves the geometry of the field, but introduces a significant design space in how the energy $\mathcal{H}_\theta^{(i)}$ is chosen and how $R_\theta^{(i)}$ is approximated.

$$q_\theta^{(i)}(H) = \frac{\exp(-\mathcal{H}_\theta^{(i)}(H))}{R_\theta^{(i)}} \tag{25}$$

One particularly simple choice is to take $\mathcal{H}_\theta^{(i)}(H)$ as a quadratic energy like Eq. 26, where $\hat{H}_\theta^{(i)}$ is the predicted clean field, and $R_\theta^{(i)}$ is simply assumed to be constant.

$$\mathcal{H}_\theta^{(i)}(H) = \tfrac{1}{2}\|H - \hat{H}_\theta^{(i)}\|^2 \tag{26}$$

Substituting Eq. 26 into Eq. 25 and evaluating the corresponding negative log-likelihood recovers a mean-squared-error(MSE) loss of Eq. 27.

$$\mathcal{L}_{\text{field}}^{(i)} \propto \|H - \hat{H}_\theta^{(i)}\|^2 \tag{27}$$

Thus, MSE can be understood not as an arbitrary engineering trick but as one particular instantiation of continuous field estimation, arising from the quadratic choice of energy. This makes it numerically stable and convex, but also limited: the field is forced toward pointwise agreement, discarding global structural constraints and potentially collapsing nontrivial dynamics. More sophisticated energy functions, such as non-quadratic or learned functions (LeCun et al., 2006; Zhang et al., 2023), may yield richer and more faithful field approximations, though exploring them lies beyond the scope of this work.

D.4 Detailed Implementation of GSH

Although a full study of discrete (Eq. 21) and continuous (Eq. 25) Green-Shell approximation mechanisms are beyond the scope of this paper, a straightforward instantiation can be conducted as follows.

Step 1: Continuous Field Constraint via Energy Functional. For each layer i, the hidden-state field $H^{(i)}$ is constrained by a field-theoretic energy functional. Inspired by the *Debye screening* (Stern et al., 2007) in plasma physics, the Green kernel is screened in Eq. 28, and the corresponding field energy functional is given in Eq. 29.

$$\mathcal{K}(r_{ij}, H_m^{(i)}, H_n^{(i)}) = \frac{H^{-r_{ij}/\lambda}}{r_{ij} + e}\Big(1 - \gamma + \gamma\big(1 - \mathrm{sim}(H_m^{(i)}, H_n^{(i)})\big)\Big) \tag{28}$$

$$\mathcal{L}_{\text{field}}^{(i)}[H^{(i)}] = \sum_{m,n} H_m^{(i)}\, \mathcal{K}(r_{mn}, H_m^{(i)}, H_n^{(i)})\, H_n^{(i)} \tag{29}$$

Optimizing Eq. 29 plays a dual role. First, through the kernel definition in Eq. 28, the interaction between a predicted state $H_m^{(i)}$ and its target $H_n^{(i)}$ is not treated as independent noise but as mediated by a screened Green's function. Here r_{ij} denotes the effective distance between prediction i and target j, while m, n index the corresponding token-level embeddings involved in this interaction. The exponential decay term ensures that long-range couplings are progressively attenuated, while the $1/r_{ij}$ singularity guarantees that no token position is ever completely isolated from the rest of the field. In addition, the similarity factor $\mathrm{sim}(H_m^{(i)}, H_n^{(i)})$ discourages trivial collapse by explicitly penalizing pairs of tokens that converge to overly similar states. This introduces a diversity-promoting bias directly at the field-interaction level, with the strength of this effect controlled by the hyperparameter $\gamma \in [0, 1]$. In contrast, the widely used MSE loss can be interpreted as a conceptually degenerate case of Eq. 29, where the kernel reduces to an identity operator $\mathcal{K} = \mathbf{I}$, thereby ignoring both spatial coupling and structural diversity, which is a convenient but ultimately impoverished approximation. It is worth noting that e is merely used for numerically preventing zero division, and we set it to 1×10^{-4}.

Second, Eq. 29 aggregates these pairwise interactions into a global quadratic form. This turns the hidden-state field $H^{(i)}$ into an energy configuration, where the total loss corresponds to the system's field energy. Unlike local objectives, which only penalize pointwise deviations, this functional captures how deviations propagate and reinforce across the entire layer. As a result, the optimization enforces global coherence across shells: every unit must not only fit its local target but also remain consistent with the structure implied by its neighbors and the overall field.

Taken together, Eq. 28 specifies the microscopic interaction law, and Eq. 29 evaluates the macroscopic energy of the system. By minimizing the latter, the network learns hidden-state fields that maintain internal consistency and diversity, without requiring any explicit causal ordering at the token level.

Step 2: Black-Box Renormalization. The continuous field regulation of Eq. 29 ensures that each layer-wise hidden-state field $H^{(i)}$ is internally constrained by its own energy functional, so every layer "knows" what to optimize for. However, at the system level there remains the global partition function R_θ inherited from Eq. 25. In physics, global normalizers such as $\int \exp(-\mathcal{H}(x))dx$ are notorious for the potential risk of unbounded energy (Green, 1889), such as the well-known *ultraviolet catastrophes* (Planck, 1900), where infinitely many high-frequency modes contribute unbounded energy. Analogously, the partition function R_θ in Eq. 25 can be dominated by unlikely but high-entropy regions of the hidden-state space, destabilizing optimization. In LMs this would manifest as overemphasis on spurious rare or noisy token patterns, so we treat R_θ as a black box and replace it with the Black-Box Renormalization of Eq. 30, while leaving a more explicit and analytic renormalization to future work.

Concretely, we approximate the intractable R_θ by projecting fields back into token space through a mixed target distribution, inspired by label smoothing (Szegedy et al., 2016). Following Eq. 21, we interpolate between the empirical dataset distribution p_{data} and a uniform prior u, with mixing coefficient $\alpha \in [0, 1]$:

$$q_{\text{mix}}(v) = (1 - \alpha)\, p_{\text{data}}(v) + \alpha\, u(v), \quad v \in \mathcal{V} \tag{30}$$

This interpolation acts as a surrogate for explicit renormalization, preventing any single high-entropy mode from overwhelming training dynamics. It is worth noting that Eq.30 also covers cross entropy as a special case by setting α to zero, and the tuning of α involves a trade-off between output diversity v.s. fitting accuracy.

The resulting black-box renormalization loss is then written as Eq. 31, where $q_\theta(v)$ denotes the model's predicted distribution over the vocabulary. Eq. 31 thus serves as a direct stand-in for the elusive R_θ, enabling stable optimization without requiring its explicit form.

$$\mathcal{L}_{\text{renormalization}} = -\sum_{v \in \mathcal{V}} q_{\text{mix}}(v)\, \log q_\theta(v) \tag{31}$$

Despite the absence of an explicit form of R_θ in Eq. 31, the intuition is straightforward. By interpolating between the empirical data distribution and a uniform prior, probability mass is redistributed so that no single mode dominates. As a result, Black-Box Renormalization provides a pragmatic closure to the variational family, ensuring that the continuous field formulation can be consistently reconciled with discrete token alignment.

Putting Things Together. To approximate a mixture of discrete and continuous fields, our standard instantiation integrates two complementary steps. First, each layer is parameterized by a continuous *field energy functional* (Eq. 29), which preserves local structure and prevents collapse into bag-of-words abstractions. Second, because the global partition function R_θ (Eq. 25) is intractable and may diverge, we employ a *Black-Box Renormalization* (Eq. 31) to project the system back into discrete token space.

The renormalization can be applied in two regimes. When restricted to the final `lm_head` or merely the last several layers, it corresponds precisely to the backpropagation-free architecture introduced by Li et al. (2025). This setting is computationally efficient but may retain residual inconsistencies. Alternatively, applying renormalization across the entire network yields more consistent yet costlier alignment. We treat the choice between these two regimes as an empirical question and evaluate both in Figure 4; a more complete investigation is left for future work.

Possible Variants. While the two-step procedure follows the physical intuition of approximating fields, it is not the only viable formulation. In practice, one may *omit the first step*, allowing the entire network to be parameterized by the renormalization term. This aligns better with standard training pipelines and offers direct optimization toward target texts. However, doing so disables the backpropagation-free possibility of `GSH`, since gradients from renormalization must now traverse all layers without a continuous-field parameterization. Conversely, one may *omit the second step* and allow the continuous field alone to drive self-consistency. Although this reduces data dependency across layers and simplifies training, it risks divergence by letting R_θ grow without bound.

Block-wise Field Approximation. We follow the block-wise generation of block diffusion. We retain causal masks between blocks, but replace the standard temporal denoising process inside each block with non-causal `GSH` approximation. Different from block diffusion, an *interleaved reconstruction* replaces the noise schedule in the training phase. Masked blocks (e.g., B and D in A[noise]C[noise]) are trained to reconstruct using surrounding unmasked blocks as boundary conditions without any noising/denoising chain, therefore shifting the objective from following a preset noise decay schedule to directly controlling S within each block. The inference phase is similar to block diffusion, where we append a pure noise block to the current context embedding, forward it through the non-causal Green Shells, and use the decoded tokens to update the prompt for the next block's generation.

E SNR AS A PROXY FOR UAT ERROR BOUND

Lazy Training Background. In the regime of *lazy training* (Jacot et al., 2018), parameter updates remain within the linearization region of the initialization, suppressing large-scale representation drift. Such a regime isolates the *system behavior* induced by the model architecture itself from data-driven feature learning dynamics, thereby providing a cleaner environment for examining model capabilities and error bounds.

Under the lazy training regime, the evolution of a neural network $F(X)$ can be linearized around its initial parameters θ_0. The approximation capability is governed by the Empirical NTK matrix

$\mathbf{K} \in \mathbb{R}^{N \times N}$, where N is the number of samples when obtaining the NTK matrix. Following Jacot et al. (2018), we define the spectral decomposition of the Gram matrix as Eq. 32.

$$\mathbf{K} = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \tag{32}$$

Signal, Noise and Bounds. Following Random Matrix Theory (RMT) and the Marchenko-Pastur law (Martin & Mahoney, 2021), we further partition the spectrum into a *Signal Subspace* and a *Noise Bulk* at the median index $m = N/2$, and the power of each component is given by Eq. 33. Note that, the optimal division threshold should be $4/\sqrt{3}$ (Gavish & Donoho, 2014), but here we use 2 for implementation simplicity. With the decomposed signal and noise, we define Signal-to-Noise of NTK (`SNR` for short) as Eq. 34.

$$P_{sig} = \sum_{i=1}^{m} \lambda_i, \quad P_{noise} = \sum_{i=m+1}^{N} \lambda_i \tag{33}$$

$$\text{SNR} = \frac{P_{sig}}{P_{noise}} \tag{34}$$

Following the localized Rademacher complexity theory (Bartlett & Mendelson, 2002; Caponnetto & De Vito, 2007), the UAT error ϵ of Eq. 9 or 13 is related to the complexity of the hypothesis class restricted to the functions with low variance. In the NTK regime, such complexity that leads to ϵ is dominated by the noise bulk P_{noise} relative to the stabilizing signal power P_{sig}. By normalizing the spectral leakage at P_{noise} against the empirical kernel scale at P_{sig}, we obtain Eq.35. C is irrelevant to data, but only affected by the neural network architectural and the complexity of task.

$$\epsilon \propto C \cdot \sqrt{\frac{P_{noise}}{N \cdot P_{sig}}} \tag{35}$$

$$\epsilon \propto \sqrt{\frac{1}{\text{SNR}}} \tag{36}$$

By keeping (1) the same neural network, (2) the same fitting object, and (3) the same number of NTK samples N, Eq. 35 and 34 leads to Eq. 36. Thus, the `SNR` calculated via Eq. 34 serves as a rigorous proxy for the structural stability and approximation convergence of f under the UAT framework.

F Other Experimental Settings

Overall Settings. Our experiments are deliberately conducted under a **lazy-training regime** (Jacot et al., 2018). This configuration ensures that parameter updates remain within the linearization region of the initialization, suppressing large-scale representation drift. Such a regime isolates the *system behavior* induced by the model architecture itself from data-driven feature learning dynamics, thereby providing a cleaner environment for examining the theoretical properties of the proposed non-causal variational family `GSH` and `SBD` itself.

To highlight the structural implications of our proposed variational family `GSH`, we conduct a controlled comparison against two canonical baselines: (1) the textbook-level **autoregressive (AR)** model, representing strict token-level causality, and (2) the recent **block diffusion** mode (Arriola et al., 2025), which realizes a Markovian causal denoising process. Conceptually, there is a progressively relaxed causal hierarchy of $AR \rightarrow BD \rightarrow$ `GSH`. All models are trained from scratch under identical configurations, using the same **150M-parameter backbone** with **ROPE positional encoding** and the same optimizer settings. Each transformer layer is implemented as a simplified Qwen-3 transformer (Yang et al., 2025) without SwiGLU activations or MoE routing, so that the purity of the dynamics remains unaffected by optimization-specific modifications. We utilize the original tokenizer and embeddings of Qwen-3 for simplicity, and let the hidden states sized at $2560 - D$. This design choice makes the subsequent analysis of NTK spectrum and field interactions directly comparable across variants, without interference from activation scaling or expert gating effects. For fairness,

both GSH and block diffusion share an identical **block mask** and perform **parallel decoding over 5 tokens per step**, whereas the AR baseline decodes one token at a time.

Conceptually, these three models form a progressively relaxed causal hierarchy. The AR model enforces strict autoregressive dependence across tokens; the block diffusion model relaxes this by introducing a Markov chain in latent space, retaining causality only across blocks but not individual tokens; and our GSH goes one step further by *removing the causality within each block*. Instead, GSH relies solely on the intrinsic directionality of the latent field and the spatial inductive bias introduced by ROPE to preserve contextual coherence. This design embodies a causality-free generative principle consistent with our theoretical framework of system behavior modelling. We construct a 10k passage subset from the dataset (TinyStories (Eldan & Li, 2023) by default) as training data, and each contains about 100 tokens. For single-objective settings of AR and block diffusion, we train for 4 epochs, while for combined objectives involved in GSH (Appendix D.4), each component is trained for 2 epochs by default (total 4). The batch size is by default set to 32, and general tendencies reserve if reduced to 8 or 16.

We utilize the Adam optimizer with a fixed learning rate of 1×10^{-4}, rather than AdamW or any adaptive scheduling variant. The reason is that the lazy-training regime assumes parameter updates remain within an infinitesimal neighborhood of initialization, such that the model behaves as a first-order kernel expansion. In this regime, introducing additional scheduling effects, such as weight decay or adaptive learning-rate modulation in AdamW, effectively alters the curvature of the optimization trajectory and changes the implicit NTK geometry. These adjustments can accelerate convergence but break the fundamental assumption of parameter stagnation ($\|\Delta\theta\| \ll \|\theta_0\|$), thereby violating the conditions required for theoretical correspondence between training dynamics and kernel evolution. By fixing the learning rate and removing extraneous scheduling, we preserve a *faithful lazy-training regime* in which observed differences in convergence and stability can be attributed solely to model structure rather than optimizer dynamics. For the specific hyperparameters of GSH, we set $\alpha = 0.25$ and $\gamma = 0.5$, which serve as balancing coefficients between reconstruction accuracy and output diversity. While this configuration may not be optimal across all conditions, it was empirically validated as stable and effective within our evaluation scope. From a theoretical standpoint, α and γ can be interpreted as modulating the relative strength of self- and cross-interaction terms. Future work may consider learning-based adaptation of these coefficients or, more fundamentally, analytical derivations that allow them to be absorbed into the renormalized kernel, thereby removing manual tuning altogether.

Machines. Our evaluations have been conducted on NVIDIA RTX 4090, A6000 Ada, and AMD RX 9070 XT GPUs, and a single card is sufficient. For most evaluations except Figure 5, a minimum of 16 GB VRAM is required (with batch size $8 \sim 16$), while 24 GB is recommended (with batch size 32). For Figure 5, 32 GB VRAM is mandatory in some LLMs.

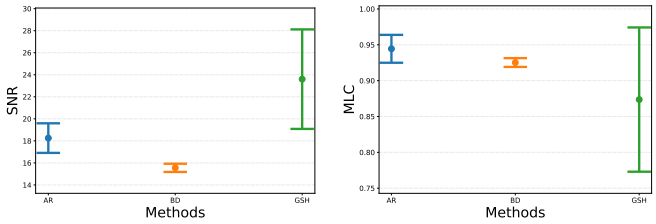
Relative Error at the Same Random Seed. Given the same random seed, it is the runtime and hardware behavior that involves minor drifts, therefore, the relative error is evaluated to be strictly bounded. The relative error across repeated runs on the same machine is within $\pm 4\%$, while the discrepancy across different NVIDIA CUDA machines is within $\pm 7\%$. The performance is comparable across CUDA and ROCm platforms, and relative error is bounded within $\pm 12\%$.

Error Bar at the Different Random Seeds. Different random seeds involve different optimizer initialization, while it does not affect the overall tendencies, the detailed metric value is affected. We therefore report the 1σ error bars across 50 random seeds on TinyStories datasets as example (Figure 6), and other cases remain similar.

Throughout this paper, we report results obtained on the RTX 4090 for consistency.

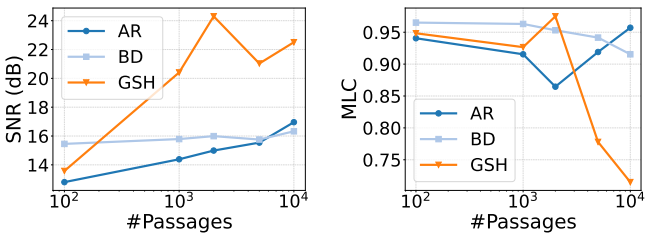
G Additional Experimental Studies

We report the additional sensitivity and ablation studies regarding GSH in this Section.



(a) Error bar for SNR. (b) Error bar for MLC.

Figure 6: Error bar across 50 different random seeds.



(a) SNR (b) MLC

Figure 7: Data scalability.

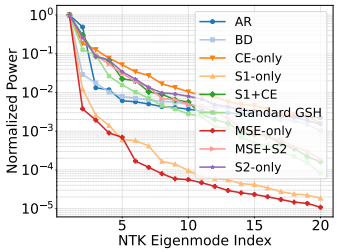


Figure 8: NTK spectra of ablation study.

G.1 Additional Sensitivity Study

In this part, we take TinyStory as the example dataset while keeping other settings default to Table 1, varying the training data scale from 100 to 10,000, and each passage involves 100 tokens. The resulting SNR and MLC are shown in Figure 7. We deliberately limit the data scale such that the number of tokens remains *far smaller* than the number of parameters. Beyond this point, the training dynamics would exit the lazy training regime, rendering the NTK spectral analysis no longer a reliable proxy for the theoretical UAT error bound. Within the full examined range, GSH consistently achieves the highest SNR, and the overall growth speed is also highest. The observed sharp curve around 2000 is primarily attributed to the current proof-of-concept implementation, particularly the heuristic renormalization and unoptimized hyperparameters that have not been fully resolved (Section 5.2 and Appendix D). More importantly, both GSH and BD exhibit an overall decreasing trend in MLC as data scale increases, while AR baseline shows the opposite trend. This suggests that GSH and BD have a weaker tendency toward overfitting compared to the standard causal formulation as Bayesian approximations. Furthermore, GSH is even better due to the greater relaxation of causality.

G.2 Additional Ablation Study

In this part, we explore how discrete and continuous establishing act themselves and interact with each other. Specifically, we conduct an ablation study by comparing **standard** GSH (Section D.4) with the following approaches.

1. **MSE+S2**: the Green-kernel parameterization (Eq. 29) of standard setting in the first step is simplified by MSE.
2. **S1+CE**: the renormalization part (Eq. 30) in the second part is reduced into cross entropy renormalization.
3. **S1-only**: the discrete setup is removed, and there is only 10 epochs of Green-kernel-based continuous training.
4. **S2-only**: the continuous setup is removed, and there is only 10 epochs of mixed-target-based discrete renormalization.
5. **MSE-only**: similar to **S1-only**, but utilize MSE.

6. **CE-only**: similar to **S2-only**, but utilize cross entropy.

We keep the basic setup the same as Figure 3(a) and report the resulting NTK spectra in Figure 8. There are two primary takeaways.

First, both continuous setup and discrete renormalization are required in GSH. When everything is discrete (**S1-only** or **CE-only**), there are even richer peripheral features learned with $MLC < 0.7$. However, the suppression of the noise bulk is even weaker than taht in AR or BD, making $SNR < 16$ dB. On the other hand, when everything is continuous (**S2-only** or **MSE-only**), there is an extremely high noise suppression with $SNR > 30dB$, but most NTK energy is concentrated with $MLC > 0.95$. This is because both continuous fields and discrete tokens are important, and one has to align them during the parameterization (Appendix D.1).

Second, the spectra also indicate an interestingly complex *field-interaction effect* between continuous and discrete components. While the noise suppression of **S1-only** is weaker than that of **MSE-only** (i.e., 34.1 dB vs. 36.2 dB in SNR), the results become opposite when combined with continuous renormalization. Specifically, standard GSH leads to higher SNR than **MSE+S2** (22.5 dB vs. 20.3 dB). In continuous–discrete coupled physics problems (e.g., hybrid particle–field simulations (Birdsall & Langdon, 2018)), the two fields are *not linearly separable*, and we suspect a similar property holds for language modeling.

H Detailed Discussions of Future Works

Training New Foundation Model Families. While our study focused on the theoretical formulation of SBD and its valid instantiation GSH under the lazy-training regime, Proposition 3.1 and Theorem 4.1 alone with their empirical validations at Section 5.3 point naturally toward a broader class of *non-causal foundation models* whose inductive biases align with global system behavior rather than local data distributions. Specifically, future works may inherit GSH's possibility of backward-propogation-free training (Li et al., 2025) at large scale , while further enrich or extend the (1) system-behavioral modeling, (2) generalized continuous and discrete loss functions and (3) field-theoretic optimizations. These directions follow directly from our mathematical derivations and may provide a theoretical foundation for more efficient, stable and adaptive large-model families.

LLM Inference Infrastructure. Most existing acceleration strategies (Jin et al., 2024; Gim et al., 2024; Fan et al., 2024; Gloeckle et al., 2024; Cai et al., 2024) treat inference efficiency as an engineering problem, decoupled from the intrinsic system behavior of models. However, since causality has been shown to be unnecessary and even suboptimal, new architectural pathways become possible. For decoding acceleration, intra-head parallelism may enable a single head to emit multiple tokens through field reconstruction, as in GSH, complementing existing inter-head parallelism approaches (Cai et al., 2024). For prefilling, strict token-level matching (Jin et al., 2024; Gim et al., 2024) could be relaxed toward *semantic approximation*, allowing more flexible reuse of intermediate states once causal constraints are lifted.

Database (DB) for AI and AI4DB. As surveyed by Zhou et al. (2022), *AI4DB* often treats models as black-box replacements for components such as learned indexes or query optimizers, while *DB4AI* applies database principles like indexing or materialized views to isolated ML operators. Both remain modular and loosely coupled, limiting deep integration between data management and model dynamics. When AI is instantiated as an LLM, the causality assumption becomes a major obstacle: data reuse tends to be all-or-nothing, requiring either full re-computation or full replay of token sequences.

However, by explicitly modeling *system behavior*, SBD downgrades causality into a suboptimal special case of LLM modeling and enables us to overcome it, potentially allowing data flow and the LLM to be co-optimized as interacting processes rather than sequential stages. Under this view, formulations beyond GSH may extend naturally toward a *field-theoretic integration* of DB and AI systems, where query engines and models are represented as a unified system behavioral field. Such a perspective suggests that storage, retrieval, and inference could become continuous operations of the same underlying field of a system, breaking the long-standing modular barrier imposed by causality.

Probabilistic Modeling. We conjecture that our perspective and AoT (as discussed in Section 2) may have a deeper connection rather than orthogonal discoveries of abnormalies. AoT attributes directional asymmetry to structural and computational properties of the model, while our framework integrates such properties as manifestations of a latent *system behavior* set S. Viewed through this lens, AoT could be interpreted as probing a specific directional derivative of system behavior with respect to token order, which is an observable imprint of internal dynamics that causal modeling alone cannot capture. This suggests a potential unifying research direction: understanding how S interacts with generation order, and whether AoT-like asymmetries are surface expressions of a more general system-level probabilistic structure. Exploring this possibility may offer a principled bridge between computational asymmetry and non-causal variational modeling, and we consider this a promising avenue for future work.

Interpretability and AGI Inspirations. `SBD` also connects to recent studies on interpretability (Shojaee et al., 2025; Ton et al., 2025; Kalai et al., 2025): phenomena such as hallucinations and repetitions may be understood not as isolated generation errors, but as manifestations of system behavior. This view enables explanations grounded in Bayesian principles rather than post-hoc probing or mechanistic attribution.

Speculatively, we observe tentative parallels to psychological studies on grandiose narcissism and narcissistic personality disorder (Zajenkowski & Gignac, 2021; Weinberg & Ronningstam, 2022), which suggest that humans may likewise exhibit persistently confident yet erroneous responses, in analogous to LLMs operating under misaligned field configurations (the **S1-Only** or **MSE-Only** setting in Figure 8). This raises a natural opening question: *is the system behavior S in Proposition 3.1 a controllable, dominant component of intelligence across domains?*

I Potential Risks of Misuse

Given the disruptive nature of rethinking causality and highlighting system behavior in LMs, it is important to acknowledge potential misuse risks as follows alongside scientific contributions.

First, malicious actors could intentionally exploit `SBD` to produce deceptive outputs, create self-reinforcing dependency loops, or embed covert triggers, underscoring the dual-use potential of this framework. Because our formulation makes system behavior explicit, it may expose new attack surfaces analogous to energy-landscape manipulations of model weights, adversarial training vulnerabilities, or even malicious infrastructure components.

Second, although our work demonstrates constructive uses of mathematical-physics field formulations in `GSH`, the same principles could be misused. Adversarial or pathological renormalizations might be engineered to induce misleading behaviors even on well-curated datasets, revealing potential vulnerabilities in model trustworthiness.

Third, our initial attempts of misconfigurations indicates tentative parallels between incorrect LM configurations and pathological human traits such as narcissistic personality disorder (Weinberg & Ronningstam, 2022). Although still speculative now, this analogy raises concern about systems that appear aligned at the task level yet exhibit unstable or manipulative internal dynamics. This may lead to risks conceptually related to emergent misalignment (Betley & et al., 2025).

At the same time, open publication may bring substantial positive impact. By articulating this framework explicitly, we enable the community to anticipate vulnerabilities, design countermeasures, and foster interdisciplinary inquiry into alignment, robustness, and the nature of intelligence itself. Moreover, analogies to pathological fields may inspire new diagnostics for detecting early signs of degeneracy in model behavior. We therefore believe that the benefits of transparency and discussion outweigh the associated risks.