

# 000 001 MULTIPLE TOKEN DIVERGENCE: MEASURING AND 002 STEERING IN-CONTEXT COMPUTATION DENSITY 003 004

005 **Anonymous authors**

006 Paper under double-blind review

## 007 008 ABSTRACT 009

011 Measuring the in-context computational effort of language models is a key chal-  
012 lenge, as metrics like next-token loss fail to capture reasoning complexity. Prior  
013 methods based on latent state compressibility can be invasive and unstable. We  
014 propose Multiple Token Divergence (MTD), a simple measure of computational  
015 effort defined as the KL divergence between a model’s full output distribution  
016 and that of a shallow, auxiliary prediction head. MTD can be computed directly  
017 from pre-trained models with multiple prediction heads, requiring no additional  
018 training. Building on this, we introduce Divergence Steering, a novel decoding  
019 method to control the computational character of generated text. We empirically  
020 show that MTD is more effective than prior methods at distinguishing complex  
021 tasks from simple ones. On mathematical reasoning benchmarks, MTD correlates  
022 positively with problem difficulty. Lower MTD is associated with more accurate  
023 reasoning. MTD provides a practical, lightweight tool for analyzing and steering  
024 the computational dynamics of language models.

## 025 1 INTRODUCTION

028 To solve unfamiliar and challenging problems, language models must perform sophisticated in-  
029 context computation (Brown et al., 2020; Lewkowycz et al., 2022). Can we tell whether, and to  
030 what extent, a model is making use of its computational capacity at any given moment? It is well-  
031 established that the next-token prediction loss offers little insight (Schmidhuber, 1991a;b), as any  
032 particular reduction in loss can, in principle, be arbitrarily difficult to achieve (Bennett, 1988). A  
033 more promising approach is to quantify meaningful computation by measuring the entropy, or in-  
034 compressibility, of a model’s latent representations (Skean et al., 2025; Herrmann et al., 2025). This  
035 concept is rooted in the minimum description length (**MDL**) principle (Grünwald, 2007; Vitányi,  
036 2006; Elmoznino et al., 2024): if the most compact description of a sequence’s structure, given the  
037 training data, is still long, then predicting that sequence is demanding due to a large search space. **In**  
038 **other words, a task is “interesting” or “complex” if the shortest description of the solution program,**  
039 **given the training data, is long. If it is short, on the other hand, we consider the task “easy” or**  
040 **“boring”.** Unfortunately, in general, the length of the shortest description is uncomputable (Li et al.,  
041 2008).

042 **To make it tractable**, the Prediction of Hidden States (PHi) loss was proposed as a measure of in-  
043 context computational complexity, quantifying the per-token information gain in a model’s latent  
044 space (Herrmann et al., 2025). While promising, the PHi framework introduces significant practical  
045 challenges: it requires inserting a noisy information bottleneck that can degrade model performance,  
046 needs further model training which can be unstable, and is highly sensitive to **the precise location of**  
047 **its placement within the model**, and the weighting of multiple loss terms.

048 In this work, we propose a simplified and more direct measure, Multiple Token Divergence (MTD),  
049 which quantifies information gain in the model’s output distribution. The core insight is simple: if a  
050 shallow computational shortcut (e.g., a single transformer block) can approximate the full model’s  
051 prediction, then the model is not performing particularly complex computation. If, however, there is  
052 a significant divergence between these two predictions, we can conclude that the model is leveraging  
053 its deeper computational capacity. MTD is straightforward to implement and can even be computed  
already possess, requiring no additional fine-tuning. In addition to this measure, we present Di-

vergence Steering, a novel decoding method that uses the MTD signal to control the computational character of the generated output. It biases the decoding distribution towards or away from computationally dense tokens—i.e., tokens that are likely under the full prediction but unlikely under shallow approximation by the MTP module. We empirically demonstrate that MTD is more effective than prior methods at distinguishing complex difficult tasks from simple ones. We also investigate the properties of MTD and Divergence Steering in reasoning and creative generation tasks.

## 2 BACKGROUND

### 2.1 PREDICTION OF HIDDEN STATES (PHI)

The Prediction of Hidden states (PHi) method, introduced by Herrmann et al. (2025), creates an information bottleneck (Tishby and Zaslavsky, 2015) within a sequence model in order to measure the complexity of its in-context computation. A PHi layer is inserted between a model’s “bottom” and “top” layers.

The model consists of the following modules: The **Bottom Layers** ( $B_\beta$ ) are the initial Transformer blocks that process the input sequence embeddings.

The **PHi Layer** contains three key components: (1) An **encoder** ( $q_\psi$ ) that, at time step  $t$ , maps the hidden state  $g_t$  from the bottom layers to a posterior distribution over a latent variable  $z_t$ . This distribution is typically a diagonal Gaussian, similar to variational auto-encoders (Kingma and Welling, 2014; Rezende et al., 2014). (2) A **decoder** ( $a_\xi$ ) that reconstructs the hidden state, creating  $g'_t$  from a sample of the latent variable  $z_t$ . (3) An **autoregressive prior** that predicts the distribution of the current latent  $z_t$  using only the history of previous latents,  $z_{<t}$ . This can be implemented with a single Transformer block ( $M_\mu$ ) and two additional linear transforms:  $b_\kappa$ , which maps the inputs to  $M_\mu$  to the right dimensionality, and  $p_\chi$ , which maps the output of the transformer to a prior distribution.

The **Top Layers** ( $T_\tau$ ) are the remaining Transformer blocks that process the sequence of reconstructed hidden states  $g'$ . Finally, we have the standard token **Embedding** ( $E_\epsilon$ ) and the **Output Layer** ( $O_\omega$ ). Here and in the remainder of the paper, Greek subscript letter indicate learnable neural network parameters.

The forward pass processing tokens  $x_1, x_2, \dots$  is described by these equations:

$e_t = E_\epsilon(x_t)$	Token embedding
$g_t = B_\beta(e_1, \dots, e_t)$	Output from bottom layers
$z_t \sim q_\psi(\cdot   g_t)$	Latent sample from posterior
$g'_t = a_\xi(z_t)$	Reconstruction of hidden state
$h_t = T_\tau(g'_1, \dots, g'_t)$	Output from top layers
$\pi(\cdot   x_{<t}) = O_\omega(h_{t-1})$	Next token prediction from output head
$L_{\text{NLL}}(t) = -\log \pi(x_t   x_{<t})$	Negative Log Likelihood (NLL) loss

108 The PHi loss ( $L_{\text{PHi}}$ ) is the KL divergence between the posterior  $q_\psi$ , which has access to the current  
 109 input  $x_t$  via  $g_t$ , and the prior  $p_\chi$ , which only has access to past latents  $z_{<t}$ . We assume an initial  
 110 latent  $z_0$  is given.

$$\begin{aligned} c_t &= b_\kappa(z_{t-1}) && \text{Linear projection of last latent} \\ 112 \quad d_t &= M_\mu(c_1, \dots, c_t) && \text{Output from PHi transformer block} \end{aligned} \quad (1)$$

$$\begin{aligned} 114 \quad L_{\text{PHi}}(t) &= D_{\text{KL}}\left(q_\psi(\cdot | g_t) \parallel p_\chi(\cdot | d_t)\right) && \text{PHi Loss} \\ 115 \quad &= D_{\text{KL}}\left(q_\psi(\cdot | x_1, \dots, x_t) \parallel p_\chi(\cdot | z_0, z_1, \dots, z_{t-1})\right) \end{aligned} \quad (2)$$

118 The model is trained to jointly minimize both  $L_{\text{NLL}}$  and  $L_{\text{PHi}}$ . The PHi loss quantifies the information  
 119 gain at each timestep—the amount of new useful information present in the current input token  $x_t$   
 120 that was not predictable from the history of latent states. It has been shown (Herrmann et al., 2025)  
 121 that this value correlates well with the complexity and “interestingness” of tasks.  
 122

## 123 2.2 MULTIPLE TOKEN PREDICTION (MTP)

125 Multiple Token Prediction (MTP) is a technique used to improve model performance and enable  
 126 faster inference via methods like speculative decoding (Cai et al., 2024; Gloeckle et al., 2024; Liu  
 127 et al., 2024; Xiaomi et al., 2025). In this setup, a computationally cheap auxiliary module is trained  
 128 to directly predict the main model’s future output distribution.

129 First, consider a standard autoregressive model’s forward pass:

$$\begin{aligned} 131 \quad e_t &= E_\epsilon(x_t) && \text{Token embedding} \\ 132 \quad h_t &= F_\phi(e_1, \dots, e_t) && \text{Output of all main transformer blocks} \\ 133 \quad \pi(\cdot | x_{<t}) &= O_\omega(h_{t-1}) && \text{Next token prediction from output head} \\ 134 \quad L_{\text{NLL}}(t) &= -\log \pi(x_t | x_{<t}) && \text{NLL loss} \end{aligned}$$

136 The goal of MTP is to approximate the main model’s prediction for the token one step further ahead,  
 137  $x_{t+1}$ . Note that often in MTP, there are additional modules that approximate predictions for tokens  
 138 even further ahead, i.e.,  $x_{t+n}$  for  $n > 1$ . We will not use them in this work.  
 139

140 A separate, smaller MTP module (e.g., a single Transformer block  $M_\mu$ ) generates its own prediction  
 141 without access to the full model’s current hidden state  $h_t$  and is usually trained with a negative log-  
 142 likelihood loss, which we call  $L_{\text{MTP}}$ . Optionally, the MTP module can be given access to the current  
 143 token embedding  $e_t$ , as indicated by the square brackets.

$$\begin{aligned} 145 \quad c_t &= b_\kappa(h_{t-1}[, e_t]) && \text{Input to MTP module (projection of } h_{t-1} \text{ and possibly } e_t) \quad (3) \\ 146 \quad d_t &= M_\mu(c_1, \dots, c_t) && \text{Output from MTP Transformer block} \\ 147 \quad \pi_{\text{MTP}}(\cdot | x_{<t}) &= O_\omega(d_{t-1}) && \text{MTP’s prediction for token } x_{t+1} \\ 148 \quad L_{\text{MTP}}(t) &= -\log \pi_{\text{MTP}}(x_t | x_{<t}) && \text{MTP Loss} \end{aligned}$$

150 In order to predict the next token  $x + 1$ , the MTP module has access to the model’s history via  $h_{t-1}$   
 151 (and optionally the current embedding  $e_t$ ), but it crucially lacks the result of the main model’s full  
 152 computation at step  $t$  (i.e.,  $h_t$ ). In recent works using MTP for Large Language Models (Xiaomi  
 153 et al., 2025; Liu et al., 2024), the MTP module shares the embedding and output heads with the  
 154 original model, implicitly enforcing alignment of the latent space.

## 155 3 MULTIPLE TOKEN DIVERGENCE

158 Observe the similarity between the PHi framework’s autoregressive prior  $p_\chi$  and posterior  $q_\psi$  on the  
 159 one hand, and the MTP prediction  $\pi_{\text{MTP}}$  and the full model prediction  $\pi$  on the other (Figure 1): in  
 160 both cases, a computationally and informationally constrained module approximates the prediction  
 161 from a full model. The key difference is that for PHi, this approximation occurs in a continuous  
 latent space, whereas MTP operates directly on the discrete token distribution.

162 Based on this analogy, we propose the *Multiple Token Divergence* (MTD) as an alternative to the  
 163 PHi loss. It is defined as the KL divergence between the full model’s next-token prediction  $\pi$  and  
 164 the MTP module’s prediction  $\pi_{\text{MTP}}$ :

$$166 \quad L_{\text{MTD}}(t) = D_{\text{KL}}\left(\pi(\cdot | x_{\leq t}) \parallel \pi_{\text{MTP}}(\cdot | x_{\leq t})\right) \quad (4)$$

$$167 \quad = D_{\text{KL}}\left(\pi(\cdot | F_{\phi}(e_1, \dots, e_t)) \parallel \pi_{\text{MTP}}(\cdot | F_{\phi}(e_1, \dots, e_{t-1}), e_t)\right).$$

$$168$$

$$169$$

170 The MTD loss,  $L_{\text{MTD}}$ , can either be optimized directly in conjunction with the standard next-token  
 171 loss  $L_{\text{NLL}}$ , or it can be calculated post-hoc using an MTP module trained with  $L_{\text{MTP}}$  loss (see Sec-  
 172 tion 4.2). For now, we ignore the optional access of the MTP module to the latest embedding  $e_t$ ; this  
 173 will be addressed [below](#).

174 **On the Difference between PHi and MTD** While PHi introduced a powerful conceptual frame-  
 175 work for analyzing a model’s internal processing, its implementation can be complex. The stochas-  
 176 ticity introduced by the variational information bottleneck can also interfere with the main sequence  
 177 prediction task. In contrast, MTD is significantly simpler to implement as it functions as a non-  
 178 invasive auxiliary task; providing a more direct and less disruptive method to obtain similar insights  
 179 into the model’s per-token computational effort.

180 One interpretation of the PHi loss is that it measures, at every step, changes of the ‘latent program’  
 181 that the model synthesizes in-context to perform next-token prediction. The MTD loss, however,  
 182 measures changes directly at the level of the output predictions. This distinction can lead to signif-  
 183 icant differences. For instance, a small change in the latent program could result in a large shift in  
 184 the output predictions. As an illustrative example, consider a model trained on two distinct types  
 185 of sequences: one type consists of uniformly random tokens, while the other is a specific single,  
 186 repeated token. To distinguish between these two cases, the latent program only needs to gain one  
 187 bit of information. The PHi loss would therefore be low. However, the resulting change in the output  
 188 distribution is large—shifting from a uniform distribution to a one-hot distribution. In this scenario,  
 189 the MTD can be as high as  $D_{\text{KL}}(\text{one-hot} \parallel \text{uniform}) = \log_2(\text{vocabulary size})$  bits. In such cases, we  
 190 expect  $L_{\text{MTD}}$  to be significantly higher than  $L_{\text{PHi}}$ , an effect we observe in our experiments in Sec-  
 191 tion 4.1. Conversely, one can imagine cases where the latent program changes significantly while  
 192 the output predictions remain stable. However, the PHi training objective penalizes encoding such  
 193 changes, as they do not sufficiently improve downstream predictions and thus represent an ineffi-  
 194 cient use of the information bottleneck. Whether the change in the latent program is larger than the  
 195 one in the output distributions or not depends on the exact weighting of the PHi loss during training.

196 **Access to the Latest Token Embedding** An interesting nuance for both PHi loss and MTD is  
 197 that the measured information gain at step  $t$  can originate from two distinct sources: (1) novel  
 198 information contained within the current token  $x_t$  itself, and (2) complex computation performed  
 199 by the model’s main layers ( $B_{\beta}$  for PHi,  $F_{\phi}$  for MTP), which cannot be easily approximated by  
 200 the simpler prior or MTP module ( $M_{\mu}$ ). To disentangle these two sources, we can provide the  
 201 prior/MTP module with direct access to the latest token embedding  $e_t$ , which is a common practice  
 202 in MTP models (see Equation 3). This effectively isolates the second source of information gain.  
 203 For the PHi framework, this modification involves updating Equation 1 to concatenate the previous  
 204 latent state with the current embedding:

$$205 \quad c_t = b_{\kappa}(z_{t-1}, e_t).$$

$$206$$

207 With this change, the PHi prior has access to the same input token as the bottom layers,  $B_{\beta}$ . Conse-  
 208 quently, the modified PHi loss, which we denote as  $\hat{L}_{\text{PHi}}$ , isolates the information gain attributable  
 209 solely to the computation performed by  $B_{\beta}$ :

$$210 \quad \hat{L}_{\text{PHi}}(t) = D_{\text{KL}}\left(q_{\psi}(\cdot | x_1, \dots, x_t) \parallel p_{\chi}(\cdot | z_0, \dots, z_{t-1}, e_t)\right).$$

$$211$$

$$212$$

213 A PHi layer modified in this way acts as an information bottleneck that specifically measures com-  
 214 putational effort. Information that the prior can easily extract from the input embedding  $e_t$  is allowed  
 215 to pass freely, while information that is computationally non-trivial for  $B_{\beta}$  to extract is quantified by  
 $\hat{L}_{\text{PHi}}$ . The same logic applies to the MTD module when it is given access to the latest embedding.

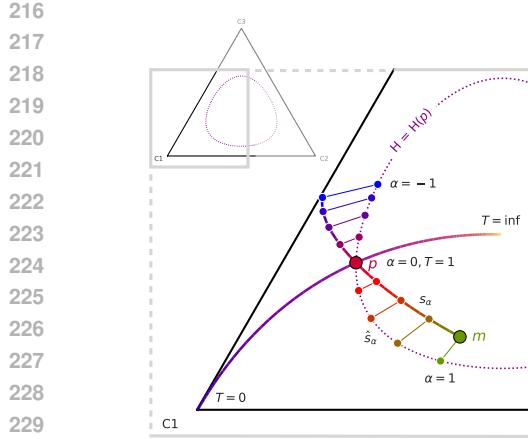


Figure 2: Divergence Steering on a  $K=3$  simplex with temperature curve for  $p$ , geodesic interpolation between from  $m$  to  $p$  and beyond, and projection onto distributions with a fixed entropy of  $H(p)$ .

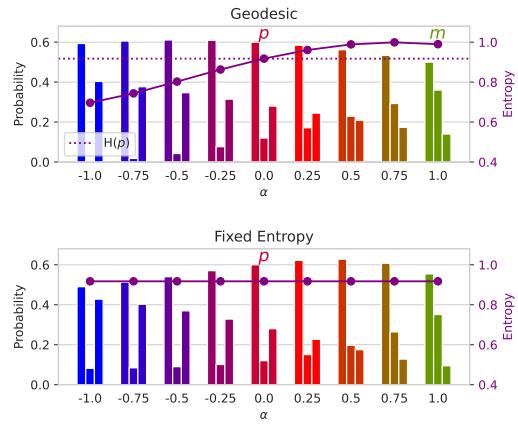


Figure 3: Distributions corresponding to Figure 2. Geodesic interpolation  $s_\alpha$ , and the entropy of the resulting distribution (top). The same distributions projected onto the surface with fixed entropy,  $\hat{s}_\alpha$  (bottom).

Arguably, PHi and MTD loss with access to the latest embedding provide a better measure of dense in-context computation. This access allows the prior/MTP module to account for trivial shifts in the predictions—like the one described in Section 3—thereby reducing the effective difference between the two metrics. In essence, providing access to the latest embedding allows us to quantify the information gain per step that is due to significant computational effort, whether measured in the latent space (PHi) or in the output distribution (MTD).

### 3.1 DECODING WITH DIVERGENCE STEERING

So far, we have presented MTD as a post-hoc analysis tool. However, its formulation, based on the divergence between two output distributions, provides a mechanism to influence the model’s behavior during generation. This allows us to steer the decoding process towards or away from tokens that the shallow MTP module can easily predict. This gives rise to a novel decoding method.

The core idea is to construct a new sampling distribution,  $s_\alpha$ , by interpolating between the full model’s prediction,  $\pi$ , and the MTP module’s prediction,  $\pi_{\text{MTP}}$ . This is controlled by a single parameter,  $\alpha$ : For  $\alpha = 0$ , we recover the original distribution from the full model:  $s_0 = \pi$ . For  $\alpha = 1$ , we use the distribution from the shallow MTP module:  $s_1 = \pi_{\text{MTP}}$ . For  $\alpha < 0$ , we extrapolate away from the MTP module’s prediction. This amplifies the probability of tokens that are considered likely by the full model but unlikely by the shallow shortcut, effectively creating an ‘anti-speculative’ distribution biased towards computationally intensive tokens.

To perform this interpolation in a principled way, we travel along the geodesic path between the two distributions under the Fisher-Rao metric. This is achieved by mapping the distributions onto the positive orthant of a hypersphere and performing spherical linear interpolation (Miyamoto et al., 2024). Let  $p = \pi$  and  $m = \pi_{\text{MTP}}$  be two categorical distributions over a vocabulary of size  $K$ . Their representations on the hypersphere are the square roots of their probabilities:

$$\mathbf{p}_g = (\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_K})$$

$$\mathbf{m}_g = (\sqrt{m_1}, \sqrt{m_2}, \dots, \sqrt{m_K})$$

The angle between these two vectors is  $\Theta = \arccos(\sum_k \sqrt{p_k m_k})$ . The geodesic path  $\mathbf{s}_g(\alpha)$  between  $\mathbf{p}_g$  and  $\mathbf{m}_g$  is then given by:

$$\mathbf{s}_g(\alpha) = \frac{\sin((1 - \alpha)\Theta)}{\sin(\Theta)} \mathbf{p}_g + \frac{\sin(\alpha\Theta)}{\sin(\Theta)} \mathbf{m}_g$$

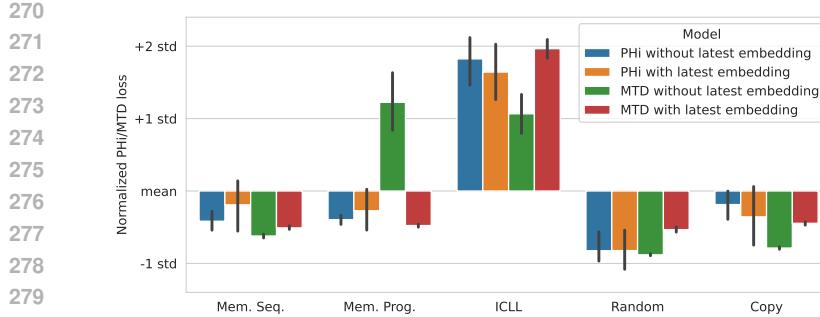


Figure 4: Normalized PHI or MTD loss of the four different model types on each of the five tasks. Only in-context language learning (ICLL) requires sophisticated in-context computation. This is reflected by the scores, with the exception of the MTD model without access to the latest embedding, which assigns high MTD also to the memorized programs task (see the discussion in Sections 3 and 4.1). Bootstrapped mean with 95% confidence intervals across 8 runs.

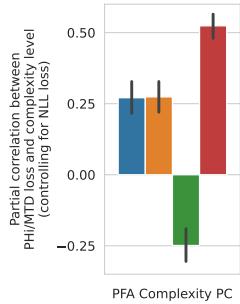


Figure 5: Partial correlation of PHi or MTD loss with the complexity of the modelled PFA, controlling for NLL. Also here, MTD without latest embedding is the outlier.

To map this path back to a valid probability distribution  $s_\alpha$ , we square each component of the vector  $\mathbf{s}_g(\alpha)$ , i.e.,  $s_k(\alpha) = (\mathbf{s}_{g,k}(\alpha))^2$ .

This method introduces a new control knob,  $\alpha$ , which is complementary to the standard temperature parameter,  $T$ . While  $T$  adjusts the entropy of the output distribution,  $\alpha$  adjusts its “computational character.” Because  $\pi_{\text{MTP}}$  often has higher entropy than  $\pi$ , changing  $\alpha$  can also affect entropy. To isolate these effects, we can optionally project the interpolated distribution  $s_\alpha$  to a new distribution  $\hat{s}_\alpha$  such that its entropy matches that of the original distribution, i.e.,  $H(\hat{s}_\alpha) = H(\pi)$  for all  $\alpha$ . This provides two orthogonal levers for shaping the decoding process:  $T$  for entropy and  $\alpha$  for computational density. Figures 2 and 3 visualize the method, additional details can be found in Appendix A. As we will show, the optimal choice of  $\alpha$  is task-dependent (which is also the case for  $T$ ): some tasks benefit from the robust, simpler predictions favored by positive  $\alpha$ , while others may require the novel, less obvious paths uncovered by negative  $\alpha$ .

## 4 EXPERIMENTS

### 4.1 MTD AND PHi LOSS OF SEQUENCE MODELS TRAINED FROM SCRATCH

The considerations from Section 3 leave us with four different model configurations to compare: PHi and MTD models, each with and without access to the latest token embedding. The PHi model without this access corresponds to the original method proposed in prior work (Herrmann et al., 2025). To compare these different setups, we train transformer models from scratch on several tasks and evaluate them in settings similar to those in (Herrmann et al., 2025). For details on the exact training setups, please see Appendix B.1.

**Evaluation on Different Tasks** The four model types are trained on five different tasks: (1) reciting memorized sequences, (2) modeling sequences from a small set of known formal languages (memorized programs), (3) in-context language learning (ICLL), where the formal language is unknown (Akyürek et al., 2024), (4) modeling random token sequences, and (5) a copying task that involves modeling random tokens where subsequences appear twice. Of these, only ICLL—which requires inferring the structure of an unknown probabilistic finite automaton (PFA) in-context—involves meaningful computation, in the sense that a non-trivial latent program must be synthesized by the model. Figure 4 shows a comparison of the normalized PHi and MTD losses for each task. The MTD with latest embedding access shows the clearest distinction between the one complex task and the four “boring” ones; note the high value for ICLL and the consistently low values for all other tasks. For the MTD model *without* latest embedding access, we see the effect alluded to in Section 3: the loss is high for both ICLL and the memorized programs. For the memorized programs task, the actual in-context program required is minimal (only  $\sim \log_2(10)$  bits

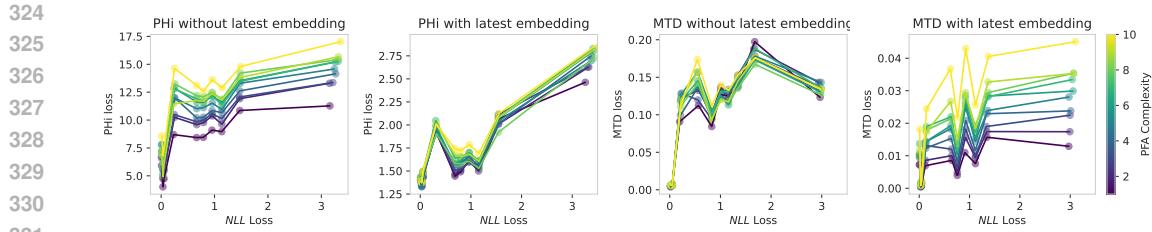


Figure 6: Token-wise PHi loss and MTD against binned NLL, for the different modeled PFA complexities. PHi loss without and MTD with latest embedding access both show a clear correlation with complexity level, across NLL bins. See Figure 9 (Appendix) for a bin-wise normalized version.

to identify any one out of the ten memorized automata). However, the lack of information from the latest token causes a significant shift in the model’s output distribution, resulting in a high MTD. Finally, giving the PHi layer access to the latest embedding does not appear to improve its ability to distinguish boring from interesting tasks.

**Task Complexity** Focusing on the ICLL task, we investigate the relationship between the models’ PHi or MTD losses and the complexity of the underlying language, as measured by the description length of the PFA. Figure 5 displays the partial correlation between the mean PHi or MTD loss across a sequence and the language’s complexity. We control for the mean NLL, as it is positively correlated with language complexity ( $r=0.367$ , 95% CI [0.315, 0.424]). Here again, we find that MTD with latest embedding access shows the strongest positive correlation ( $r=0.524$  [0.480, 0.565]). In contrast, MTD without access to the latest embedding is negatively correlated with language complexity when controlling for NLL, confirming that it is not a reliable measure for this purpose. Figure 6 shows the token-wise PHi or MTD loss against binned NLL loss, broken down by PFA complexity (from 1, simple, to 10, complex). This analysis reveals that only the original PHi loss (without embedding access) and the MTD loss with embedding access show a clear, positive token-wise relationship with language complexity after controlling for NLL.

## 4.2 PRE-TRAINED LANGUAGE MODELS

To validate our hypotheses on existing large-scale models, we leverage the pre-trained, open-source MiMo-7B model (Xiaomi et al., 2025). We choose it as a high-quality 7B parameter model, whose base pre-training incorporates an MTP objective, providing the built-in auxiliary prediction heads necessary for calculating the MTD without any post-hoc modification. Additionally, we use the Mistral 7B model (Jiang et al., 2023), which provides no MTP module out-of-the-box. Instead we train MTP heads of different sizes by keeping the base model frozen and using standard teacher-student distillation (Schmidhuber, 1992; Hinton et al., 2015) with a small fraction of the original data and compute (see Appendix B.2). For clarity, in the main part of the paper, we report the results only for the MiMo model. The qualitatively very similar results for the Mistral models can be found in Appendix C.1

**Reasoning Difficulty** We employ the MATH dataset (Hendrycks et al., 2021b), which provides mathematics problems labeled from Level 1 (easy) to Level 5 (hard), along with detailed reasoning solutions. We first compute the mean MTD for the provided step-by-step solution for each problem in the dataset and find that it clearly correlates with the difficulty level ( $r=0.179$ , 95% CI [0.152, 0.203]). Interestingly, the NLL loss *negatively* correlates with problem difficulty ( $r=-0.249$  [-0.274, -0.224]). This suggests that from the model’s perspective, reasoning chains for difficult problems are no less plausible or predictable. However, the higher MTD indicates that the model makes increased use of its full capacity to process and generate them. We also have the model generate ten different chains-of-thought (CoTs) for each problem and repeat the analysis on these self-generated solutions. There again, we observe very similar results: the partial correlation between MTD and difficulty level, controlling for NLL, is  $r=0.199$  [0.189, 0.208], while the correlation between NLL and difficulty is  $r=-0.158$  [-0.168, -0.149]. These effects hold consistently across most problem categories, as shown in Figures 10 and 11 (Appendix). Since the provided rationales, as well as

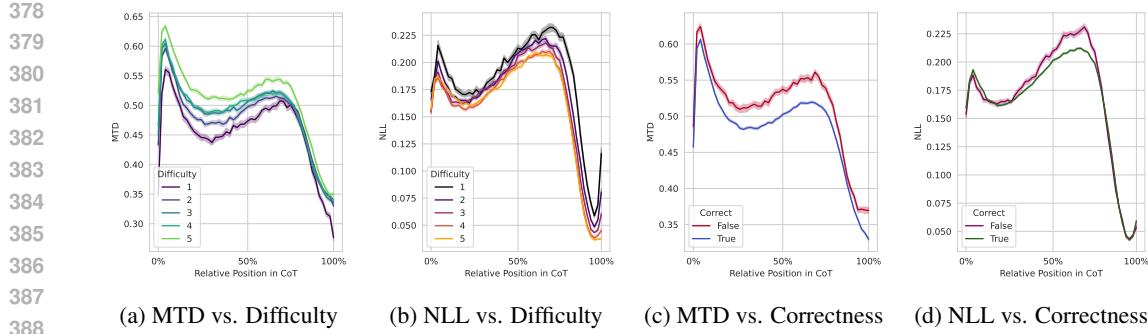


Figure 7: Token-wise losses against relative positions in self-generated CoT for the MATH test dataset. MTD shows a clear correlation with difficulty across the full CoT (a), the relationship between NLL and difficulty is less clear (b). Similarly, correct CoTs show higher MTD over all relative positions (c), which is not the case for NLL (d).

the generated CoTs, are longer for more difficult problems, the cumulative NLL also correlates positively with difficulty level (see Figures 12 and 13 in the Appendix).

**Token-wise Development Across the Response** We track the token-wise values for MTD and NLL across each generated CoT. As seen in Figure 7a, the positive correlation between MTD and problem difficulty holds consistently from the first tokens of the response to the last. Likewise, the negative correlation for NLL persists throughout the generation, even though not as pronounced (Figure 7b). The difference between MTD and NLL in their correlation with the problem difficulty is notable because, at a global level, MTD and NLL are positively correlated with each other ( $r=0.255$  [0.246, 0.265]). This highlights that MTD captures a distinct signal related to computational effort that is not present in the standard NLL loss.

**Reasoning Accuracy** For the self-generated CoTs, we also investigate the relationship between MTD values and the correctness of the final answer. Figure 7c plots the token-wise MTD, stratified by whether the rationale was correct or incorrect. We observe that correct responses are consistently associated with lower MTD. The relationship between NLL and correctness is less consistent (see Figure 7d). Following the methodology from prior work (Herrmann et al., 2025), we randomly assemble pairs of one correct and one incorrect CoT for each math problem. The probability of choosing the correct CoT when picking the one with the lower mean MTD is 67.1% (95% CI: [65.4%, 68.7%]). When selecting the one with the lower NLL, the probability is 73.3% [71.9%, 74.8%]. For the cases where NLL and MTD are agree, we get 80.4% [78.5%, 81.9%] accuracy. We repeat these experiments on the GSM-8k dataset (Cobbe et al., 2021), where we find that selecting CoTs with lower MTD yields 66.0% [62.9%, 69.2%] correct answers, while lower NLL yields 72.2% [69.1%, 75.0%] and combined yields 75.5% [71.7%, 79.2%]. For token-wise curves, please see Appendix C.1. These findings stand in contrast to the results for PHi loss, where, for a Llama 3B model, correct answers are associated with a *high* PHi loss (Herrmann et al., 2025). While further investigation is needed, we hypothesize that different models may have different tendencies to either overly simplify or overly complicate their reasoning process (Sui et al., 2025). This tendency could determine whether computationally intensive answers—as opposed to more straightforward ones—are more or less likely to be correct for a given model architecture or training regime.

### 4.3 DIVERGENCE STEERING AND CREATIVE TASKS

Having established MTD as an indicator of complex in-context computation, we now investigate whether we can use it to influence model generation. Specifically, can biasing generation towards tokens with high MTD lead to more complex or creative outputs? The Divergence Steering method allows us to test this hypothesis.

**Algorithmic Toy Tasks** We adopt the framework proposed by Nagarajan et al. (2025), training transformer models on four distinct tasks: sibling discovery, triangle discovery, circle construction,

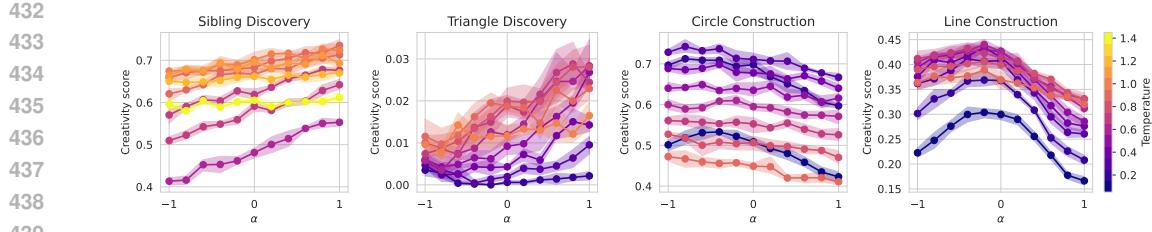


Figure 8: For the discovery tasks, positive  $\alpha$  leads to higher creativity, whereas for the construction tasks, negative  $\alpha$  leads to higher creativity. Results for geodesic distributions  $s_\alpha$ .

and line construction (for details, please see Appendix B.3). The objective for each task is to generate sequences that are simultaneously *valid*, *novel* (i.e., not memorized from the training set), and *unique* within a fixed number of attempts. Success is measured by a creativity score, where 1 indicates perfect performance across all three criteria. All models used in this experiment are configured with MTP modules that have access to the latest token embedding. Figure 8 shows the creativity scores across a range of values for temperature and our steering parameter,  $\alpha$ . The results reveal a task-dependent effect. For the “discovery” tasks, positive values of  $\alpha$ —which bias generation toward the simpler predictions of  $\pi_{\text{MTP}}$ —yield higher creativity scores. For the “construction” tasks, negative values of  $\alpha$ —which create an “anti-speculative” distribution biased away from  $\pi_{\text{MTP}}$ —lead to better performance. A more detailed analysis (Appendix C.2) suggests that positive  $\alpha$  helps the model avoid memorized solutions (improving novelty), whereas negative  $\alpha$  can encourage the generation of more structurally sound outputs (improving validity). The optimal strategy, therefore, depends on the specific demands of the task. Crucially, temperature and  $\alpha$  function as largely independent controls over the decoding process: for all four tasks, the best-performing combination of temperature and  $\alpha$  achieves significantly higher creativity scores than optimizing for temperature alone. The qualitative behavior is similar for both geodesic and fixed-entropy distributions (see Figure 16 in the Appendix).

**Creative Writing Benchmark** We also investigate how Divergence Steering can affect the creative writing performance of the MiMo model. For the creative writing benchmark (Paech, 2025), the model generates several short stories following various prompts. An LLM judge scores the generated stories according to many different criteria, and also assigns a main score, the ‘Overall Impression’. From table 1, we can see that a slightly negative steering parameter ( $\alpha = -0.1$ ) leads to the best results overall. Most criteria, including Overall Impression, reach the best scores when biasing tokens away from the MTP distribution via Divergence Steering. As we would expect, we get the least ‘Unsurprising and Uncreative’ stories for negative  $\alpha$ . For positive  $\alpha$ , however, stories show fewer characteristics connected with overcomplication, such as ‘Purple Prose’ and ‘Overwrought’. The experimental details and additional results can be found in Appendix B.3 and C.3. These results show that shaping the computational character of the decoding distribution can have positive effects, especially for creative tasks.

## 5 DISCUSSION & FUTURE WORK

In our experiments, MTD outperforms PHi loss in differentiating “boring” from “interesting” tasks and simple from complex ones. It successfully isolates the per-token information gain attributable to non-trivial, or “irreducible” (Wolfram, 2002), computation by the model. It should be noted that the utility of the MTD signal can be contingent on the relative capacities of the main model and the MTP module ( $M_\mu$ ): if the MTP module is too powerful, MTD approaches zero, and if it is too weak, MTD offers little beyond the standard NLL loss. However, experiments with several different MTP sizes show robustness of the effects reported in this work (Appendix C.1). Furthermore, MTD may conflate genuine computational effort with simple patterns that exceed the capacity of the lightweight shortcut.

Our findings also surface several intriguing questions. The positive correlation of MTD with problem complexity, in direct contrast to the negative correlation of NLL, warrants further investigation to determine if this is a general pattern across models and scales. Similarly, our result that lower

486  
 487 Table 1: Creative writing criteria which reach the best score for a given  $\alpha$  value. For 10 different  
 488 criteria, including Overall Impression,  $\alpha = -0.1$  gives the best results. In contrast,  $\alpha = 0$  works  
 489 best only for 6 criteria. Criteria followed by a down arrow ( $\downarrow$ ) are negative, meaning that a low score  
 490 is best. Results for geodesic distributions  $s_\alpha$ .

$\alpha = -0.4$	$\alpha = -0.2$	$\alpha = -0.1$	$\alpha = 0.0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.4$
—	(1) Unsurprising or Uncreative $\downarrow$	(1) Overall Impression, (2) Emotionally Complex, (3) Emotionally Engaging, (4) Adherence to Instructions, (5) Believable Character Actions, (6) Nuanced Characters, (7) Imagery and Descriptive Quality, (8) Elegant Prose, (9) Well-earned Lightness or Darkness, (10) Weak Dialogue $\downarrow$	(1) Overall Reader Engagement, (2) Consistent Voice/Tone of Writing, (3) Coherent, Engaging, (4) Sentences Flow Naturally, (5) Amateurish $\downarrow$ , (6) Meandering $\downarrow$	(1) Unearned Transformations $\downarrow$ , (2) Incongruent Ending Positivity $\downarrow$	—	(1) Overwrought $\downarrow$ (2) Purple Prose $\downarrow$

502 MTD is associated with correct reasoning contrasts with prior findings for PHi loss, suggesting the  
 503 relationship between computational effort and correctness is complex and model-dependent. While  
 504 Divergence Steering enhanced performance on creative tasks, in preliminary experiments we found  
 505 no clear improvement in the reasoning of large pre-trained models, perhaps because significant  
 506 changes to the decoding strategy interfere with behaviors learned during post-training.

507 Our findings suggest that MTD and Divergence Steering have the potential for many applications in  
 508 training and inference. Examples could be **Dynamic Compute Allocation**: MTD could be moni-  
 509 tored in real-time during generation. A prolonged period of low MTD might trigger early stopping  
 510 for a simple task, while a sudden spike in MTD could activate more powerful components (e.g.,  
 511 additional Mixture-of-Experts layers) for a difficult step. **Solution Convergence**: The transition  
 512 from a high-MTD processing phase to a low-MTD conclusion could act as a signal that the model  
 513 has “settled” on a solution, potentially allowing for more efficient decoding. **Intrinsic Motivation**:  
 514 In agent-based settings, MTD could serve as an intrinsic reward. This would encourage an agent to  
 515 pursue policies that lead to computationally interesting states (high information gain), fostering the  
 516 development of more sophisticated behaviors. **Open Endedness**: MTD and Divergence steering al-  
 517 lows the filtering or direct generation of “interesting” data. This may help to prevent model collapse  
 518 when training on self-generated data and enable more creative, open-ended learning.

## 527 6 CONCLUSION

528 In this work, we introduce Multiple Token Divergence (MTD), a practical and direct measure for  
 529 quantifying the computational effort of language models. By measuring information gain in the  
 530 output distribution, MTD serves as a more robust and stable metric than prior methods that rely on  
 531 latent state compression. We show that giving the auxiliary prediction module access to the latest  
 532 token embedding allows MTD to specifically isolate the information gain attributable to non-trivial  
 533 computation. Our findings demonstrate that MTD successfully distinguishes complex in-context  
 534 reasoning from simpler tasks and reveals a nuanced relationship between computational effort and  
 535 predictive loss. As a non-invasive and easily implemented metric, MTD provides a valuable new  
 536 tool for analysis and evaluation. Furthermore, we introduce Divergence Steering, a novel decoding  
 537 method that uses the MTD signal to actively steer the generation process towards either more or less  
 538 computationally dense sequences. Shaping this “computational character” is complementary to the  
 539 standard entropy adjustment using decoding temperature and can help with creative tasks.

540 REPRODUCIBILITY STATEMENT  
541

542 All experiments can be reproduced using a single consumer-grade GPU (e.g., NVIDIA GeForce  
543 RTX 4090). Appendix B provides the necessary details for the exact replication of our experiments.  
544 The weights of the pre-trained models used in this work, MiMo 7B (Xiaomi et al., 2025) and Mistral  
545 7B (Jiang et al., 2023), are publicly available. Similarly, we employ open-source training and test  
546 datasets: FineWeb (Penedo et al., 2024), MATH (Hendrycks et al., 2021a), and GSM-8k (Cobbe  
547 et al., 2021). The PFA-related tasks (Section 4.1) follow the implementation in Herrmann et al.  
548 (2025), and the creative toy tasks (Section 4.3) are adopted from Nagarajan et al. (2025), both with  
549 no changes except the ones specified. The prompts used for the creative writing benchmark are also  
550 publicly available (Paech, 2025).

551  
552 REFERENCES  
553

554 Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. In-context language learning: Architectures and  
555 algorithms. In *Proc. Int. Conf. on Machine Learning (ICML)*, Vienna, Austria, July 2024.

556 C. H. Bennett. Logical depth and physical complexity. In *The Universal Turing Machine: A Half Century  
557 Survey*, pages 227–258. Oxford University Press, 1988.

559 Tom B Brown et al. Language models are few-shot learners. In *Proc. Advances in Neural Information Pro-  
560 cessing Systems (NeurIPS)*, Virtual only, December 2020.

561 Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao.  
562 Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:*  
563 [2401.10774](https://arxiv.org/abs/2401.10774), 2024.

565 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plap-  
566 pert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training veri-  
567 fiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.

568 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,  
569 Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang,  
570 Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien  
571 Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern,  
572 Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe  
573 Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel  
574 Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-  
575 Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan,  
576 Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis An-  
577 derson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar,  
578 Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evti-  
579 mov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer  
580 van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Ji-  
581 awen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua John-  
582 stun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Ken-  
583 neth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. URL  
584 <https://doi.org/10.48550/arXiv.2407.21783>.

585 Eric Elmozino, Tom Marty, Tejas Kasetty, Léo Gagnon, Sarthak Mittal, Mahan Fathi, Dhanya Sridhar, and  
586 Guillaume Lajoie. In-context learning and occam’s razor. *ArXiv*, abs/2410.14086, 2024.

587 Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster  
588 large language models via multi-token prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller,  
589 Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st  
International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*,  
590 pages 15706–15734. PMLR, 21–27 Jul 2024.

591 Peter D. Grünwald. *The Minimum Description Length Principle*. Springer, 2007.

592 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob  
593 Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proc. Advances in Neural  
Information Processing Systems (NeurIPS)*, December 2021a.

594 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob  
 595 Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.  
 596

597 Vincent Herrmann, Róbert Csordás, and Jürgen Schmidhuber. Measuring in-context computation complexity  
 598 via hidden state prediction. In *Forty-second International Conference on Machine Learning*, 2025.

599 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9,  
 600 2015.

601 Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego  
 602 de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,  
 603 Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and  
 604 William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

605 Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Int. Conf. on Learning Represen-*  
 606 *tations (ICLR)*, Banff, AB, Canada, April 2014.

607

608 Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh,  
 609 Ambrose Sloane, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-  
 610 Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Proc. Advances*  
 611 *in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, USA, November 2022.

612 Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer,  
 613 2008.

614 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng,  
 615 Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

616

617 Sachit Menon, David Blei, and Carl Vondrick. Forget-me-not! Contrastive critics for mitigating posterior  
 618 collapse. In *Uncertainty in Artificial Intelligence*, pages 1360–1370. PMLR, 2022.

619 Henrique K Miyamoto, Fábio CC Meneghetti, Julianna Pinele, and Sueli IR Costa. On closed-form expressions  
 620 for the fisher–rao distance. *Information Geometry*, 7(2):311–354, 2024.

621

622 Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice & look before you  
 623 leap: Going beyond the creative limits of next-token prediction. In *Forty-second International Conference*  
 624 *on Machine Learning*, 2025.

625

626 Samuel J Paech. Eq-bench creative writing benchmark v3. <https://github.com/EQ-bench/creative-writing-bench>, 2025.

627

628 Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. Fineweb: decanting the web for  
 629 the finest text data at scale. *HuggingFace*. Accessed: Jul, 12, 2024.

630

631 Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate  
 632 inference in deep generative models. In *Proc. Int. Conf. on Machine Learning (ICML)*, volume 32, pages  
 633 1278–1286, Beijing, China, June 2014.

634

635 Jürgen Schmidhuber. Curious model-building control systems. In *Proc. Int. Joint Conf. on Neural Networks*,  
 636 pages 1458–1463, 1991a.

637

638 Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural con-  
 639 trollers. In *Proc. Int. Conf. on Simulation of Adaptive Behavior: From Animals to Animats*, pages 222–227,  
 640 1991b.

641

642 Jürgen Schmidhuber. Learning complex, extended sequences using the principle of history compression. *Neural*  
 643 *Computation*, 4(2):234–242, 1992. doi: 10.1162/neco.1992.4.2.234.

644

645 Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-  
 646 Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International*  
 647 *Conference on Machine Learning*, 2025.

648

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

652

653 Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *2015 IEEE*  
 654 *Information Theory Workshop (ITW)*, 2015.

648 Paul M Vitányi. Meaningful information. *IEEE Transactions on Information Theory*, 52(10), 2006.  
649  
650 Stephen Wolfram. A new kind of science. 2002.  
651  
652 LLM-Core Xiaomi, :, Bingquan Xia, Bowen Shen, Cici, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang,  
653 Huaiqiu Liu, Jiebao Xiao, Jinhao Dong, Liang Zhao, Peidian Li, Peng Wang, Shihua Yu, Shimao Chen,  
654 Weikun Wang, Wenhan Ma, Xiangwei Deng, Yi Huang, Yifan Song, Zihan Jiang, Bowen Ye, Can Cai,  
655 Chenhong He, Dong Zhang, Duo Zhang, Guoan Wang, Hao Tian, Haochen Zhao, Heng Qu, Hongshen  
656 Xu, Jun Shi, Kainan Bao, Kai Fang, Kang Zhou, Kangyang Zhou, Lei Li, Menghang Zhu, Nuo Chen,  
657 Qiantong Wang, Shaohui Liu, Shicheng Li, Shuhao Gu, Shuhuai Ren, Shuo Liu, Sirui Deng, Weiji Zhuang,  
658 Weiwei Lv, Wenyu Yang, Xin Zhang, Xing Yong, Xing Zhang, Xingchen Song, Xinzhe Xu, Xu Wang,  
659 Yihan Yan, Yu Tu, Yuanyuan Tian, Yudong Wang, Yue Yu, Zhenru Lin, Zhichao Song, and Zihao Yue.  
660 Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025. URL  
661 <https://arxiv.org/abs/2505.07608>.  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

## 702 A FIXED ENTROPY PROJECTION FOR DIVERGENCE STEERING

704 To project the distribution  $s_\alpha$  onto the hypersurface with entropy  $H(p)$ , we solve the following  
 705 optimization problem:

$$\begin{aligned} 706 \quad & \min_{\hat{s}_\alpha} D_{\text{KL}}(\hat{s}_\alpha || s_\alpha) \\ 707 \quad & \text{subject to } H(\hat{s}_\alpha) = H(p) \\ 708 \quad & \sum_i \hat{s}_{\alpha,i} = 1 \\ 709 \end{aligned}$$

711 By solving the Lagrangian, we see that this is equivalent to finding a temperature-scaled version of  
 712  $s_\alpha$ . This means that  $\hat{s}_\alpha$  takes the form:

$$714 \quad \hat{s}_\alpha = \text{softmax} \left( \frac{\log s_\alpha}{T} \right) \\ 715$$

716 for some temperature  $T$ , such that the entropy constraint  $H(\hat{s}_\alpha) = H(p)$  is met. Since entropy is a  
 717 smooth monotonic function of the temperature, we can use a fast root-finding algorithm like binary  
 718 search to find the correct value for  $T$ .

719 In practice, divergence steering, either with geodesic interpolation or this fixed-entropy projection,  
 720 does not meaningfully slow down the generation process. For large vocabularies, however, it might  
 721 be sensible use Divergence Steering in combination with top-k sampling and only optimize the  
 722 remaining smaller distribution.

## 724 B EXPERIMENT DETAILS

### 726 B.1 SEQUENCE MODELS TRAINED FROM SCRATCH

728 We train all models using the Adam optimizer, a batch size of 16 and gradient norm clipping of 1.0.  
 729 The learning rate is 0.0003, with a 500 step linear warm-up from zero and no decay. All losses are  
 730 weighted equally, for the PHi loss we take the mean of the element-wise KL-Divergence for  $z$ , not  
 731 the sum. Every model variation is trained 8 times with different random seeds for the initial weights  
 732 and the procedurally generated data (which results in different memorized sequences and programs).  
 733 The training of a model can be done on a single consumer-grade GPU (e.g., NVIDIA RTX 4090).

734 The base model is based on the Llama 3.2 architecture (Dubey et al., 2024).

- 736 • Number of layers: 12
- 737 • Model dimensionality: 768
- 738 • Number of attention heads: 6
- 739 • MLP intermediate size: 2048
- 740 • Embedding layer and output head are tied

#### 741 PHi models:

742 To prevent posterior collapse, we employ an additional contrastive self-critic loss (Menon et al.,  
 743 2022).

- 744 • Training steps: 30,000
- 745 • Placement of the PHi Layer: After the 10th layer
- 746 •  $z$  dimensionality: 768
- 747 •  $q_\psi$ : Linear transform
- 748 •  $a_\xi$ : Linear transform
- 749 •  $b_\kappa$ : Linear transform
- 750 •  $M_\mu$ : One transformer block like the ones in the rest of the model
- 751 •  $p_\psi$ : Linear transform

#### 752 MTD models:

- 754 • Training steps: 10,000
- 755 •  $b_\kappa$ : Linear transform
- 756 •  $M_\mu$ : One transformer block like the ones in the rest of the model

756 For generation of training and testing data, we follow Herrmann et al. (2025). The only difference  
 757 is that we do not perturb any tokens during training, and that we use the same models for the task  
 758 differentiation and task complexity experiments (Section 4.1).  
 759

760 **B.2 PRE-TRAINED LANGUAGE MODELS**  
 761

762 For our experiments in Section 4.2, we use the SFT version of the MiMo-7B model (Xiaomi et al.,  
 763 2025). To calculate the MTD, we use the included MTP head that predicts one token in advance.

764 For additional experiments, we use the instruction tuned Mistral 7B v0.3 model (Jiang et al., 2023).  
 765 We train MTP modules of three different sizes: **MTP small** consists of one transformer block,  
 766 **MTP medium** of two blocks, and **MTP large** of three. They are initialized using the weights from  
 767 the last layers of the full model and trained on data from the FineWeb (Penedo et al., 2024), the  
 768 MATH training set, and the GSM-8k training set (a mixture of 80%, 10% and 10% respectively).  
 769 We use MTD as the training loss. The MTP modules have access to the latest embedding. We  
 770 use the AdamW optimizer, a batch size of 4 and train for 30,000 steps, with cosine annealing and  
 771 a maximum learning rate of 0.0001. As shown in Table 2, this can be done in a few hours on a  
 772 consumer-grade GPU.  
 773

All experimental results include bootstrapped 95% confidence intervals.

774 **B.3 DIVERGENCE STEERING AND CREATIVITY TASKS**  
 775

776 **Algorithmic Toy Tasks** The MTD models use the architecture and training procedure specified in  
 777 Section B.1. For each task, a dedicated model is trained for 50,000 steps. No seed conditioning is  
 778 used. For task definitions and evaluation procedure, we refer to Nagarajan et al. (2025).

779 The creativity score is defined as the fraction of all generated items that are valid, unique, and novel.  
 780 In addition, we define three more scores:

781

- 782 • Validity score: fraction of valid items among all generated items
- 783 • Uniqueness score: fraction of unique items among valid generated items
- 784 • Novelty score: fraction of novel items among valid unique generated items

785 These are used in the additional empirical analysis in section C.2.

786 **Creative Writing** For the generation of the stories, we use the prompting structure from the creative  
 787 writing benchmark (Paech, 2025). Divergence steering for temperatures [0.6, 0.7, 0.8] and  
 788  $\alpha$ -values [-0.4, -0.2, -0.1, 0., 0.1, 0.2, 0.4] is evaluated, with geodesic interpolation. For each setting,  
 789 a total of 96 stories is generated and judged. As a judge model, we use Claude Sonnet 4.5.  
 790 Again we adhere to the prompting structure and criteria proposed by the benchmark. Every story  
 791 is evaluated individually, no pairwise ELO-type scoring is used. The detailed results are shown in  
 792 section C.3.  
 793

794  
 795  
 796  
 797  
 798  
 799  
 800  
 801  
 802  
 803  
 804  
 805  
 806  
 807  
 808  
 809

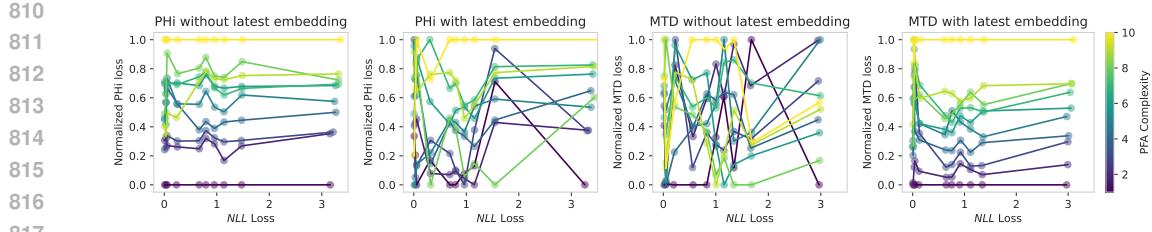


Figure 9: Similar to Figure 6, but normalized for each NLL bin. PHi loss without access to the latest embedding, and MTD loss with access to the latest embedding both show a clear correlation with complexity level, across NLL bins.

## C ADDITIONAL EXPERIMENTAL RESULTS

Figure 9 shows the normalized PHi losses and MTDs, binned by NLL and normalized, making it clear to see that PHi without and MTD with access to the latest embedding show the clearest token-wise relationship with PFA complexity.

### C.1 DETAILED RESULTS FOR MiMo AND MISTRAL

In this section, we report detailed MTD results for the Mistral 7B models, in comparison with the MiMo model, and for different sizes of the MTP module. The main take-away is that the results are consistent with the findings for MiMo and stable across the MTP module configurations. While qualitatively very similar, they are quantitatively less strong, which suggests that training a model from scratch with an MTP module can have benefits.

Table 2 shows that, as we would expect, with larger MTP module size, the MTP loss and MTD significantly drop. However, as it turns out, this does not affect the relationship between MTD and reasoning difficulty and correctness. We also note that, while due to different tokenization, the losses between MiMo and Mistral models are not directly comparable, MiMo’s MTP Loss is much closer to the full NLL, suggesting that its MTP module can predict better than even the large MTP module for Mistral.

From Table 3, we can see that the relationship between task difficulty and MTD is the same for MiMo and Mistral models. For both the provided solutions and for ten self-generated CoTs to the MATH test questions, it consistently shows significant *positive* correlation between MTD and question difficulty (controlled for NLL) and *negative* correlation between NLL and question difficulty (controlled for MTD). The differences in correlation strength for the three MTP module sizes fall within the confidence intervals. For the CoTs generated by Mistral, the partial correlations with difficulty are present and statistically significant, but less clear than for MiMo.

Table 4 shows the relationship between MTD and the correct reasoning: we measure the probability of picking the correct answer out of pair of random pair of generated CoTs, one correct, and the other one incorrect, when choosing based on the specified criterion. We test the MiMo and Mistral models on the test splits of the MATH and GSM-8k datasets. For all models and both datasets, there is a clear connection between low MTD and correctness. This connection goes beyond the relationship between NLL and correctness, which we can see from the fact that the chance of choosing correctly increases significantly if we pick a response with lower NLL *and* lower MTD. Also here, the difference between the Mistral MTP module sizes stay within the confidence intervals.

864  
865  
866  
867  
868  
869

Table 2: Comparison between MiMo and Mistral models in terms of tokenizer vocabulary size, loss of the main model (NLL), loss of the MTP module (MTP), multiple token divergence (MTD), and training time. For Mistral models, larger MTP modules lead to reduced MTP Loss and MTD.

Metric	MiMo	Mistral MTP Small	Mistral MTP Medium	Mistral MTP Large
Vocabulary Size	151936	32000	32000	32000
NLL	1.345	0.943	0.943	0.943
MTP Loss	1.425	1.298	1.217	1.178
MTD	0.469	0.553	0.476	0.429
RTX 4090 hours	—	9.38	10.25	11.08

870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882Table 3: Partial correlation coefficients  $r$  for the test questions of the MATH dataset between MTD and difficulty, controlled for NLL, and between NLL and difficulty, controlled for MTD. Results for the provided step-by-step solutions, and for the self-generated CoTs. Across all models, MTD correlates positively with difficulty, and NLL negatively.

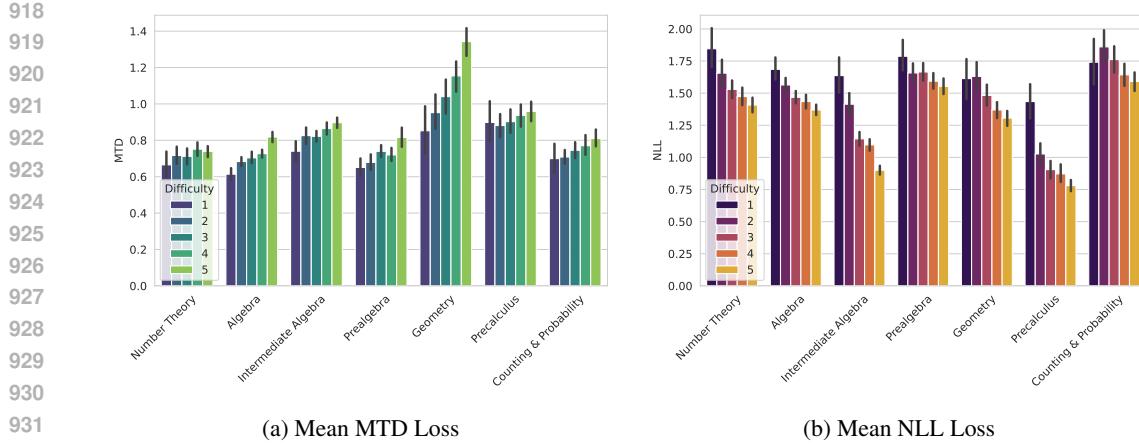
Correlation Type	MiMo	Mistral MTP Small	Mistral MTP Medium	Mistral MTP Large
<b>Provided Solutions</b>				
MTD-Diff (ctrl NLL)	0.162 [0.137, 0.189]	0.109 [0.079, 0.138]	0.094 [0.065, 0.122]	0.096 [0.066, 0.128]
NLL-Diff (ctrl MTD)	-0.239 [-0.266, -0.212]	-0.153 [-0.180, -0.127]	-0.143 [-0.172, -0.115]	-0.144 [-0.173, -0.115]
<b>Generated CoT Solutions</b>				
MTD-Diff (ctrl NLL)	0.199 [0.189, 0.208]	0.050 [0.041, 0.058]	0.044 [0.035, 0.053]	0.043 [0.034, 0.052]
NLL-Diff (ctrl MTD)	-0.205 [-0.215, -0.195]	-0.041 [-0.050, -0.033]	-0.047 [-0.056, -0.038]	-0.047 [-0.056, -0.038]

893  
894  
895  
896  
897  
898  
899  
900  
901

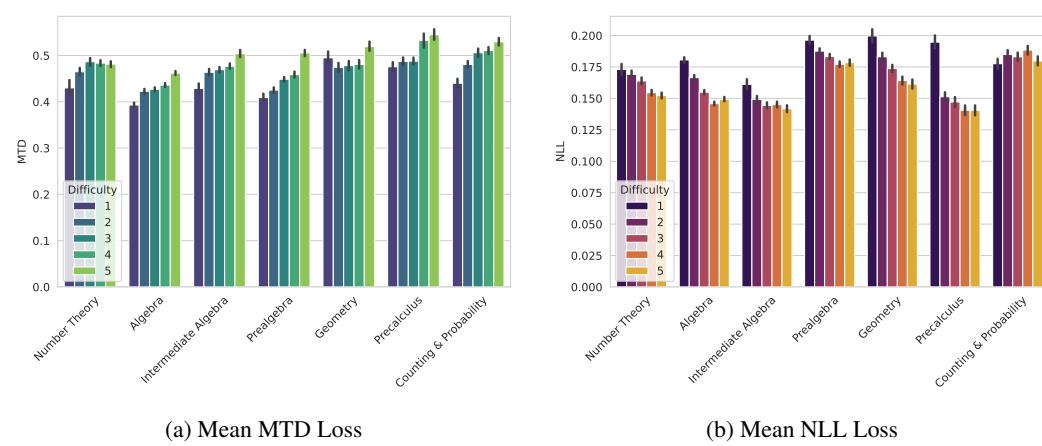
Table 4: Chance of the correct answer from pairs of CoTs, for different choosing strategies. Across all models, and for MATH and GSM-8k test questions, lower MTD CoTs are more likely to be correct.

Criterion	MiMo	Mistral MTP Small	Mistral MTP Medium	Mistral MTP Large
<b>MATH Dataset</b>				
Lower NLL	73.3% [71.9, 74.8]	59.1% [57.6, 60.6]	59.1% [57.6, 60.6]	59.1% [57.6, 60.6]
Lower MTD	67.1% [65.5, 68.5]	57.2% [55.6, 58.6]	57.5% [56.0, 59.0]	57.8% [56.2, 59.3]
Both Lower	80.4% [78.5, 81.9]	63.0% [61.1, 64.9]	62.7% [60.8, 64.6]	62.7% [60.9, 64.6]
<b>GSM-8k Dataset</b>				
Lower NLL	72.2% [69.1, 75.2]	64.9% [63.1, 66.7]	64.9% [63.1, 66.7]	64.9% [63.1, 66.7]
Lower MTD	66.0% [62.4, 69.3]	59.2% [57.5, 61.0]	61.8% [60.1, 63.5]	61.3% [59.6, 63.1]
Both Lower	75.5% [71.7, 79.2]	69.9% [67.6, 72.1]	71.4% [69.3, 73.5]	70.9% [68.8, 73.0]

912  
913  
914  
915  
916  
917



933 Figure 10: Mean losses of the MiMo model across the provided step-by-step solutions to the problems  
934 of the MATH test set, grouped by category and difficulty level. MTD clearly grows with difficulty,  
935 suggesting that the model is making more use of its computational capacity when processing  
936 more challenging problems. NLL loss, on the other hand, goes down with increasing complexity.  
937 Figure 11 shows similar results for self-generated chains of thought.

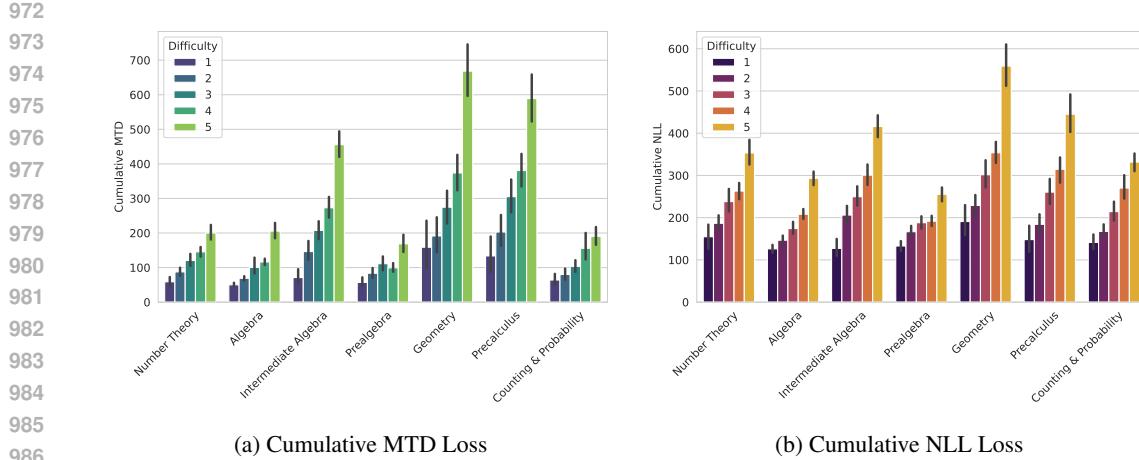


953 Figure 11: Mean losses of the MiMo model across self-generated CoTs for the problems of the  
954 MATH test set, grouped by category and difficulty level. Similarly as in Figure 10, we observe that  
955 MTD clearly grows with difficulty, as the model is making more use of its computational capacity  
956 when generating the solutions to more challenging problems. Also here, the mean NLL goes down  
957 with problem difficulty.

958  
959 **Additional Visualizations** Figure 10 shows MTD and NLL for the provided step-by-step solutions  
960 for the MiMo model, broken down by category and difficulty level. Figure 11 shows the same for  
961 the self-generated CoTs. The results are qualitatively similar, even though, although the differences  
962 between categories for the CoTs are less pronounced.

963 Figures 12 and 13 use the cumulative instead of the mean losses. Due to the fact that the pro-  
964 vided solutions as well as the generated ones grow in length as the problems become more difficult,  
965 cumulative NLL also correlates positively with difficulty level.

966 Figure 14 shows the development of MTD and NLL across self-generated CoTs for the problems  
967 of the GSM-8k test dataset (analogous to Figures 7c and 7d for MATH). Correct CoTs clearly have  
968 lower MTD, and lower NLL. Interestingly, for the GSM-8k dataset, the shapes of the NLL curves  
969 differ significantly from the shapes of the MTD, missing the prominent initial bump. Currently, we  
970 have no explanation for this.



987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

Figure 12: Cumulative losses of the MiMo model across provided solutions from the MATH test set. Since more difficult problems have longer solutions, both cumulative MTD and cumulative NLL correlate with problem difficulty.

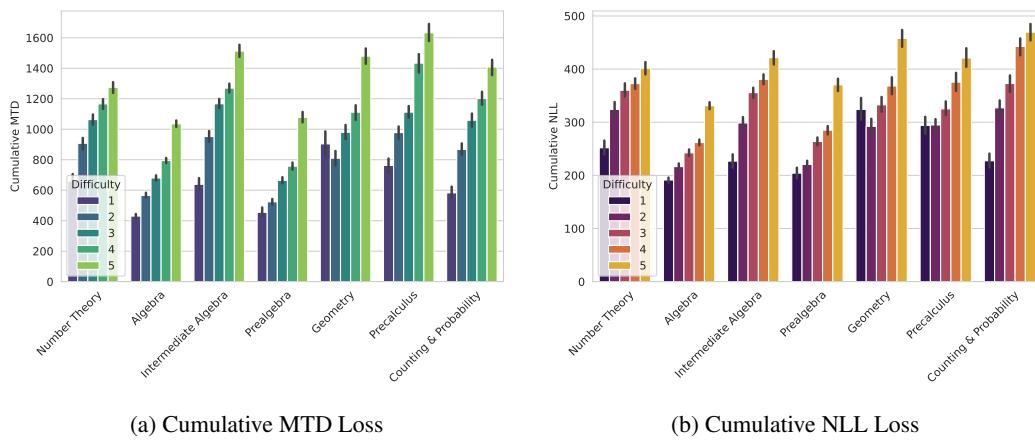


Figure 13: Cumulative losses of the MiMo model across self-generated CoTs for the problems of the MATH test set. We observe a similar effect as in Figure 12.

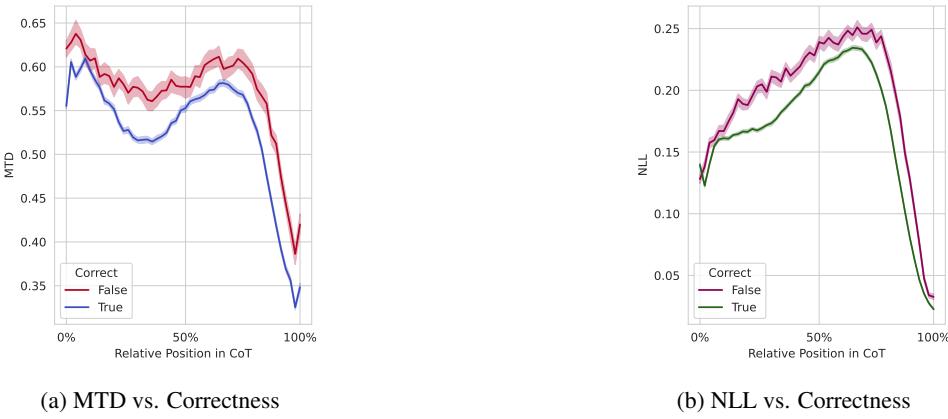
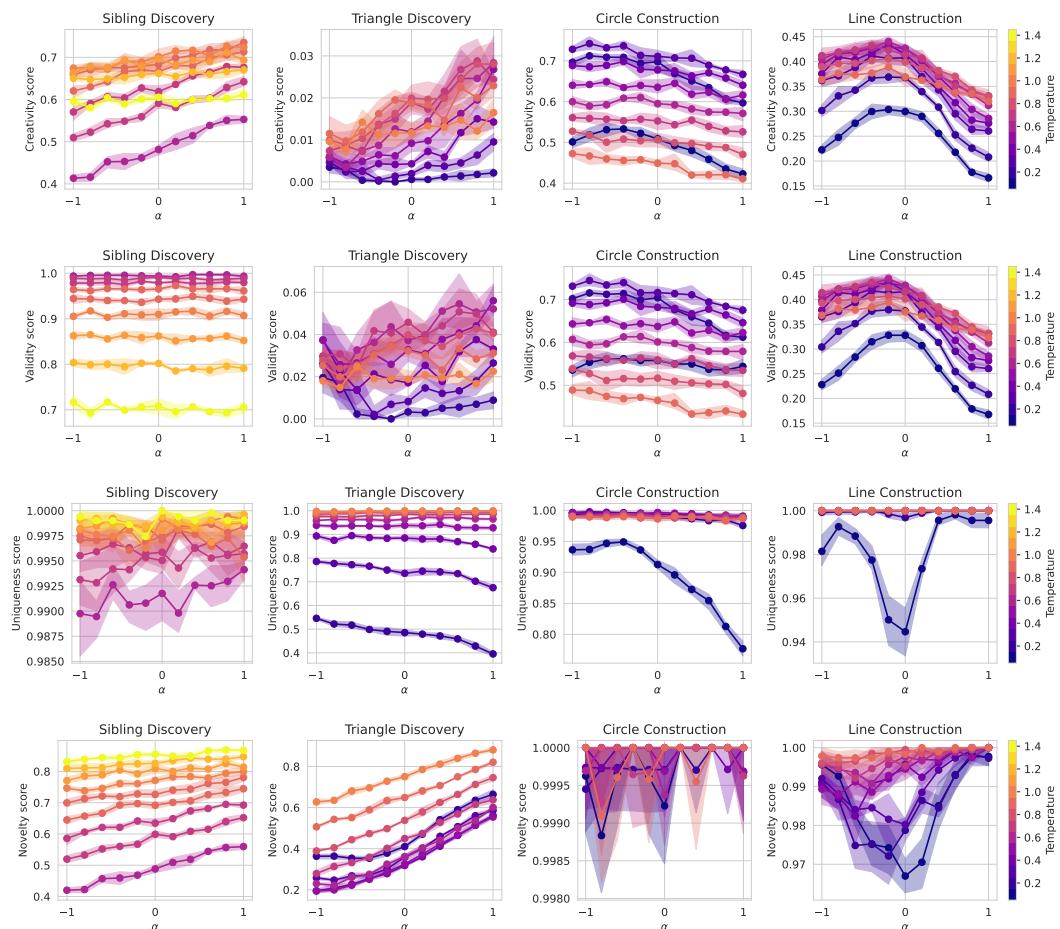


Figure 14: Token-wise losses against relative positions in self-generated CoTs by MiMo for the GSM-8k test dataset. Lower MTD and lower NLL are both associated with more correct reasoning.

1026 C.2 DIVERGENCE STEERING AND CREATIVE ALGORITHMIC TASKS  
1027  
1028  
1029

1030 Figure 15 shows the creativity scores for the four tasks, using different values for temperature and  $\alpha$ .  
 1031 In addition, we break down the results into validity, uniqueness and novelty scores. By the nature of  
 1032 the task, sibling and triangle discovery models are at risk of overfitting to the training data. A positive  
 1033  $\alpha$  value can help avoiding repeating memorized examples, as can be seen from the increased novelty  
 1034 scores. The models for circle and line construction, on the other hand, are less prone to overfitting,  
 1035 due to the combinatorial nature of the task. The novelty and uniqueness scores are consistently high.  
 1036 For these tasks, negative  $\alpha$  appears to help construct increase the validity scores.

1037 Figure 16 shows qualitatively very similar results for fixed entropy distributions.



1073 Figure 15: Breakdown of the creativity scores into validity, uniqueness, and novelty. Positive  $\alpha$  can  
 1074 improve novelty, negative  $\alpha$  can improve validity. Results for geodesic distributions  $s_\alpha$ .  
 1075  
 1076  
 1077  
 1078  
 1079

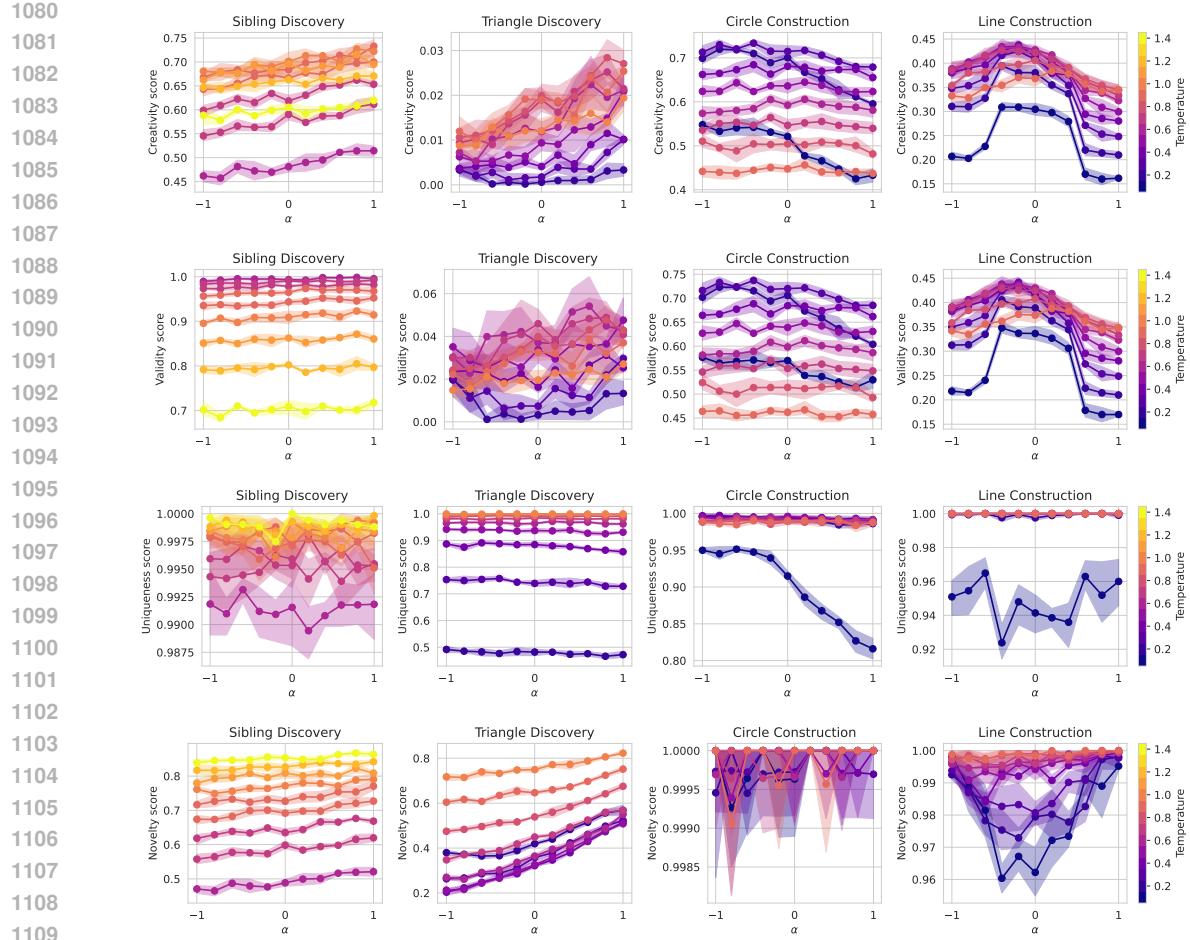


Figure 16: Breakdown of the creativity scores into validity, uniqueness, and novelty. Positive  $\alpha$  can improve novelty, negative  $\alpha$  can improve validity. Results for fixed entropy distributions  $\hat{s}_\alpha$ .

### C.3 DIVERGENCE STEERING AND CREATIVE WRITING

Here we report detailed results for different divergence steering settings in the creative writing benchmark (Paech, 2025). For each temperature- $\alpha$  configuration, MiMo generates 96 stories following a range of pre-defined writing prompts. An LLM judge writes an assessment and rates each story according to 21 different criteria, including the aggregate ‘Overall Impression’, which serves as the main score. Figures 17 and 18 show the relationship between temperature,  $\alpha$  and the different scores. For many criteria, including the overall impression, there is a connection between good scores and slightly negative  $\alpha$ . This means that steering away from the ‘obvious’ tokens predicted by the MTP module can lead to better writing performance. On the other hand, some negative criteria connected with excess complexity, such as ‘Overwrought’ and ‘Purple Prose’ tend to increase with negative  $\alpha$ .

To establish statistical significance, we perform a quadratic regression analysis on the Overall Impression scores. We find that the optimal  $\alpha$  is negative and reject the null hypothesis (that  $\alpha = 0$  works best) with  $p < 0.01$ .

### DECLARATION OF LLM USAGE

We used Gemini 2.5 and 3.0 to proofread and occasionally to improve sentence flow and wording.

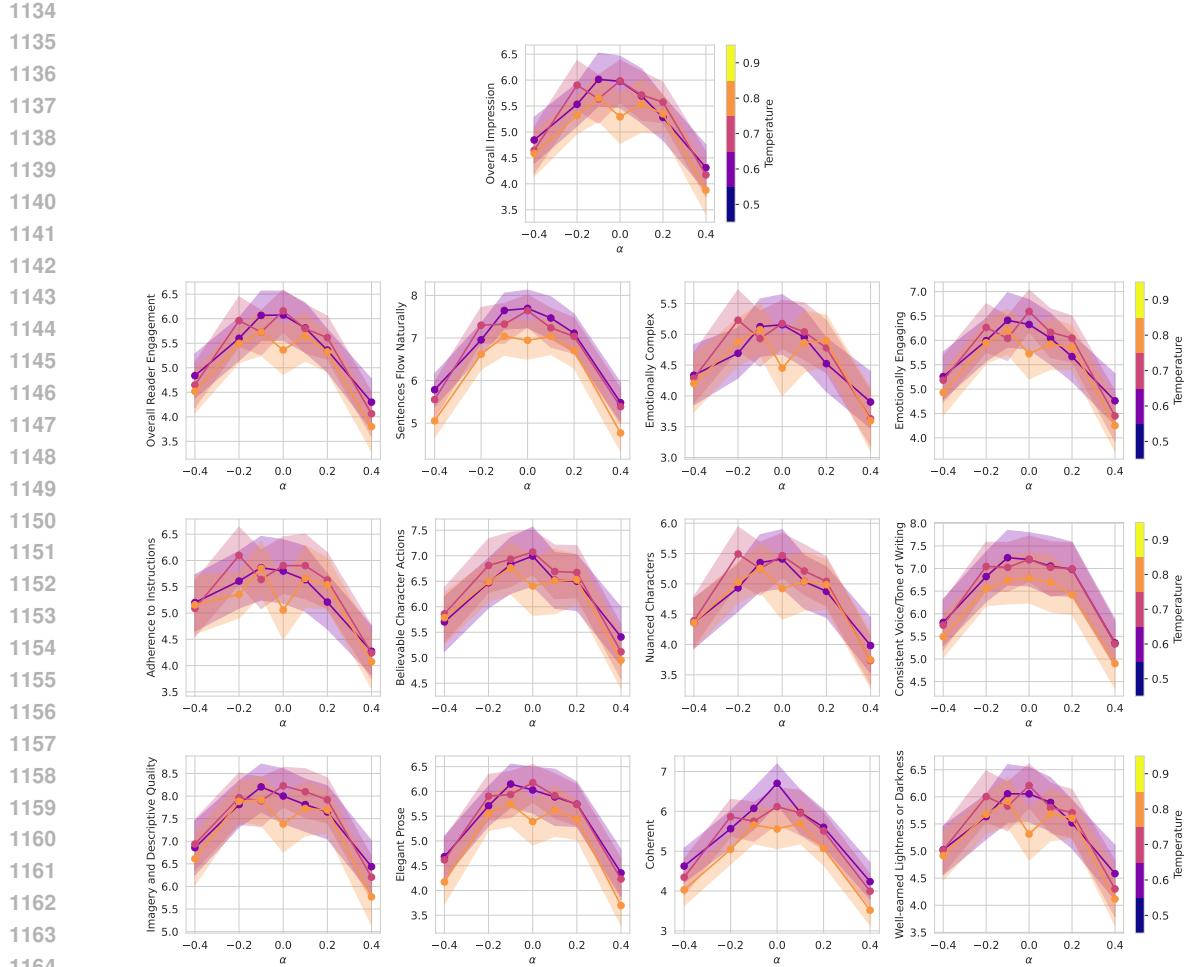


Figure 17: Creative writing scores for positive criteria, for different values of temperature and  $\alpha$ . Higher value means better. At the top, the aggregate ‘Overall Impression’ is shown. For many criteria, including Overall Impression, we observe the highest scores for  $\alpha = -0.1$ .

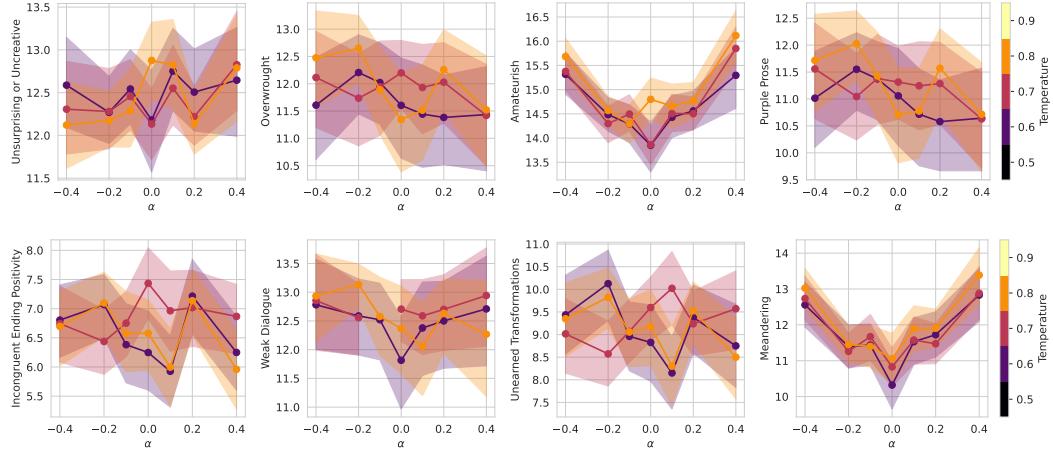


Figure 18: Creative writing scores for negative criteria (lower means better).