TYPICALITY OF REPRESENTATIONS PREDICTS INPUT-LEVEL PERFORMANCE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting whether a model will perform well on a set of inputs remains a topic of central importance in machine learning. Motivated by the information-theoretic notion of typicality, we address this open problem by introducing representational typicality: a framework for measuring how well the intermediate representations induced by an input conform to a model's expected representational regularities, and consequently of the model's performance on the input. Without requiring any dataset labels or model retraining, the proposed method first derives the characteristic statistics of a model's intermediate representations using only readily accessible data. The method then evaluates whether the representations induced by new data are compatible, in the sense of being typical, with these characteristic representation statistics to predict input-level performance. Across an array of widely used language models and downstream tasks, our experiments comprehensively demonstrate that typicality is highly predictive of input-level performance, outperforms baseline measures, remains predictive across varying experimental configurations, and is robust to a suite of hypothetical confounding factors.

1 INTRODUCTION

What characterizes when a model will be performant on some input data? This is a central and open problem in machine learning, yet there remains no conclusive answer to the question.

To make progress towards this goal, consider the notion of typicality from information theory. In this setting, the set of all possible sequences can be split into two groups: the typical set, and the non-typical set. The typical set (Shannon, 1948) consists of the sequences which match the empirical statistics of the underlying source data. These sequences usually form just a small fraction of all the possible sequences, yet any draw from the source will almost surely yield a sequence that is a member of this set (Cover & Thomas, 2006).

Importantly, the single most likely sequences are often not members of the typical set, because they are uncharacteristic of naturally-occurring sequences. This is because in high dimensional spaces, the estimate which maximizes pointwise likelihood (i.e. the maximum likelihood estimate, or MLE), may lie far from the set containing most of the probability mass (Vershynin, 2026). Taking the case of the multivariate gaussian distribution as an example, the MLE is the mean; yet almost all the probability mass concentrates in a shell of characteristic radius away from the mean.

To relate this to the learning setting, first consider that in the context of neural networks, the successive activations induced by an input can be viewed as a model's internal representation of the input. Thus typicality implies that the inputs which induce the most likely intermediate activation patterns (i.e. those which maximize likelihood), are not necessarily representative of the family of activation patterns which characterize a model's reference distribution. This also matches the observed natural world – there is not one canonical sentence or input pattern in general, but rather a vast set of inputs, all of which can be considered approximately equally typical.

Applying this notion of typicality to the learning setting, the central hypothesis of this work is:

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.

Central Hypothesis

A model will be performant on inputs which induce representations that are compatible, in the sense of being typical, with the model's expected representation statistics.

In this work we ask the question of what specific aspect of a model-input pair is predictive of performance. We answer this question using representational typicality: a framework we introduce for quantifying the typicality of the intermediate representations induced by an input.

In summary, our contributions are as follows:

1. We introduce representational typicality, a framework for quantifying how well an input conforms to a model's expected activation regularities. We then define the typicality distance metric as a measure which quantifies the extent of an input's deviation from typicality.
2. We instantiate representational typicality in the language model setting by estimating the expected intermediate representation statistics, taken over token-level hidden representations, for a calibration set of data. At test time, we measure an input's aggregate deviation from typicality across the model layers.
3. We empirically evaluate and verify typicality as a predictor of input-level performance across several models and downstream tasks, and conduct ablations and control experiments which serve to distinguish the predictive power of typicality from hypothetical confounding factors.

The rest of the paper is organized as follows. In Section 2 we provide a brief background on typical sequences before presenting our methods in Section 3. In Section 4 we describe our experiments, analyze the results, and explore the proposed method's predictive power across varying experimental configurations. In Section 5 we comment on related work, and finally in Section 6 we provide a few concluding remarks.

2 Background: Typical Sequences

Here we provide a brief outline of the information-theoretic notion of typicality.

Typical Sequences Let X denote a random variable with probability density function $f(x)$, and let $x^N = \{x_1, \ldots, x_N\}$ be a sequence drawn from X. Let the differential entropy of X, a measure of its information content (Cover & Thomas, 2006), be denoted by $h(X)$.

A sequence is considered typical when its negative log-likelihood (NLL) is close to $h(X)$, i.e.:

$$-\frac{1}{N}\log f\left(x^N\right) \approx h(X). \tag{1}$$

Thus the typical sequences are the sequences which are likely to have been sampled from the source distribution, and which have representative information content.

Typical Set of the Normal Distribution To build intuition, suppose X is normally-distributed with mean μ and variance σ^2, i.e. $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Its differential entropy is $h(X) = \frac{1}{2}\log\left(2\pi e\sigma^2\right)$, while the NLL of a sequence $x_1, \ldots, x_N \in \mathbb{R}$ is

$$-\frac{1}{N}\log\prod_{i=1}^{N} f(x_i) = \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2N\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right)^2. \tag{2}$$

As the sample size N grows asymptotically large, $\frac{1}{N\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 \to 1$, and consequently the NLL approaches $h(X)$. For finite sample size N, a sequence is thus considered typical when

$$\frac{1}{N\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 \approx 1. \tag{3}$$

Geometric Perspective on Typical Sequences Equation 3 implies that the average squared value of a (standardized) typical sequence has a characteristic value of 1. Equivalently, a sequence is typical if its corresponding vector lies within a thin shell concentrated around the unit hypersphere.

3 METHODS

We begin by defining our typicality distance metric in Subsection 3.1. We then instantiate the method in the language model setting, describe the derivation of the calibration set statistics, and explore a few facets of the method in Subsection 3.2. We subsequently describe how the method is used at test time in Subsection 3.3.

3.1 TYPICALITY DISTANCE METRIC

In Section 2 we examined what characterizes a typical sequence in the univariate setting. However, the activations of neural networks are better represented by multivariate random variables, since each network layer consists of multiple units/channels. Thus we consider the multivariate setting here.

Consider $(X_1, \ldots, X_K) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for which $h(X_1, \ldots, X_K) = \frac{1}{2} \log\left((2\pi e)^K |\boldsymbol{\Sigma}|\right)$. The NLL of a sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^K$ is

$$-\frac{1}{N} \log \prod_{i=1}^{N} f(\boldsymbol{x}_i) = \frac{K}{2} \log(2\pi) + \frac{1}{2} \log|\boldsymbol{\Sigma}| + \frac{1}{2N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}). \tag{4}$$

As the sample size N grows asymptotically large, $\frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \to K$, and consequently the NLL approaches $h(X_1, \ldots, X_K)$. For finite sample size N, a sequence can thus be considered typical when

$$\frac{1}{NK} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \approx 1. \tag{5}$$

Therefore, we define the typicality distance metric as:

$$D \coloneqq \left| \frac{1}{NK} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) - 1 \right|, \tag{6}$$

which precisely quantifies deviation from typicality.

3.2 CALIBRATION SET STATISTICS

Let $\mathcal{A}$ denote a set of selected affine layers in a language model. Consider a calibration set of data containing P prompts, where the p-th prompt consists of N_p tokens. For layer $\ell \in \mathcal{A}$, let $\boldsymbol{x}_{\ell,p,t} \in \mathbb{R}^{K_\ell}$ denote the activation vector of the t-th token in the p-th prompt, where K_ℓ is the representation dimension of the layer.

For each layer $\ell \in \mathcal{A}$ and each calibration prompt p, we estimate the coordinate-wise mean $\boldsymbol{m}_{\ell,p} = \frac{1}{N_p} \sum_{t=1}^{N_p} \boldsymbol{x}_{\ell,p,t}$ and variance $\boldsymbol{s}^2_{\ell,p} = \frac{1}{N_p} \sum_{t=1}^{N_p} (\boldsymbol{x}_{\ell,p,t} - \boldsymbol{m}_{\ell,p})^2$ across the N_p tokens. We then standardize each calibration prompt's coordinate-wise activations according to its respective statistics, i.e. we evaluate: $\boldsymbol{z}_{\ell,p,t} = \mathrm{diag}(\boldsymbol{s}_{\ell,p})^{-1} (\boldsymbol{x}_{\ell,p,t} - \boldsymbol{m}_{\ell,p})$.

We subsequently pool the token representations over all of the calibration prompts, for a total of $N_P = \sum_{p=1}^{P} N_p$ activation vectors. Specifically, for each layer $\ell \in \mathcal{A}$, we estimate the global coordinate-wise mean $\boldsymbol{\mu}_\ell = \frac{1}{N_P} \sum_{p=1}^{P} \sum_{t=1}^{N_p} \boldsymbol{x}_{\ell,p,t}$ and variance $\boldsymbol{\sigma}^2_\ell = \frac{1}{N_P} \sum_{p=1}^{P} \sum_{t=1}^{N_p} (\boldsymbol{x}_{\ell,p,t} - \boldsymbol{\mu}_\ell)^2$, and a global covariance matrix $\boldsymbol{R}_\ell = \frac{1}{N_P} \sum_{p=1}^{P} \sum_{t=1}^{N_p} \boldsymbol{z}_{\ell,p,t} \boldsymbol{z}^\top_{\ell,p,t}$ across all of the N_P standardized vectors. These statistics are computed over the calibration set once and thereafter held fixed.

We next discuss a few facets of the methods presented in this subsection.

Importance of Standardizing Before Measuring the Covariance Matrix Standardizing the activations before measuring the covariance matrix prevents coordinates with large activation scales from dominating the statistics. This is of particular relevance in the language model setting due to the presence of activations exhibiting significantly larger values relative to others (Sun et al., 2024). Thus the covariance matrix $\boldsymbol{R}_\ell$ retains information about the dependence between coordinates, without becoming sensitive to coordinate-wise scales (this is especially valuable because large activation scales can be mitigated by the scale of the model's subsequent weights).

Use of Affine Layers Applying the method to the outputs of affine layers mitigates the need to adapt the method to account for the various types of sub-modules in a network. This is because each affine layer transforms the previous layer's outputs using a linear projection. Furthermore, the outputs of affine layers provide a reliable approximation to normally-distributed data (Wolinski & Arbel, 2025) (although as is shown in the experimental results, having normally-distributed activations is not a requirement for the method to work). Altogether, using affine layers facilitates the wide adoption of the method across various model and layer types.

No Auxiliary Information and No Model Retraining Are Required The method only requires the network's forward pass activations; no labels, no auxiliary model, and no model retraining is required. For this reason, the method is simple to use and is widely applicable.

3.3 Test Time

Consider a test prompt containing N tokens. Let $\boldsymbol{x}_{\ell,t}$ denote the activations corresponding to token t, in layer ℓ.

Using the calibration statistics derived in Subsection 3.2, we begin by standardizing $\boldsymbol{x}_{\ell,t}$ to obtain $\boldsymbol{z}_{\ell,t} = \operatorname{diag}(\boldsymbol{\sigma}_\ell)^{-1}(\boldsymbol{x}_{\ell,t} - \boldsymbol{\mu}_\ell)$. We then evaluate the quantity $\boldsymbol{z}_{\ell,t}^\top \boldsymbol{R}_\ell^{-1} \boldsymbol{z}_{\ell,t}$, and subsequently average across all of the tokens in the layer and normalize by the layer dimension to obtain the expression $\frac{1}{NK_\ell} \sum_{t=1}^{N} \boldsymbol{z}_{\ell,t}^\top \boldsymbol{R}_\ell^{-1} \boldsymbol{z}_{\ell,t}$.

Following our derivation from Subsection 3.1, we then evaluate each layer's typicality distance as

$$D_\ell = \left| \frac{1}{NK_\ell} \sum_{t=1}^{N} \boldsymbol{z}_{\ell,t}^\top \boldsymbol{R}_\ell^{-1} \boldsymbol{z}_{\ell,t} - 1 \right|. \tag{7}$$

A small value for D_ℓ indicates that the prompt induces activations vectors with the correlation structure expected under the reference calibration distribution, whereas a large value indicates a representational mismatch.

Finally, we average across the layer-wise typicality distances to obtain a model-level value:

$$D_M = \frac{1}{|\mathcal{A}|} \sum_{\ell \in \mathcal{A}} D_\ell. \tag{8}$$

In summary, we first use a calibration set to obtain the layer-wise statistics $\boldsymbol{\mu}_\ell$, $\boldsymbol{\sigma}_\ell$, and $\boldsymbol{R}_\ell$. At test time, we standardize the representations of each token across the layers of the model, compute each layer's corresponding typicality distance value using Equation 7, and finally aggregate the layer-wise values using Equation 8 to obtain a model-level typicality distance value.

4 Experimental Results & Analysis

4.1 Experimental Setup

Details on the models and datasets employed are outlined in Appendix A.

Each result shows the absolute value of the Spearman rank correlation coefficient ρ (Spearman, 1904) (a measure of rank correlation) associated with it. Wherever a result is shown graphically, it is displayed using the mean and its 95% confidence interval.

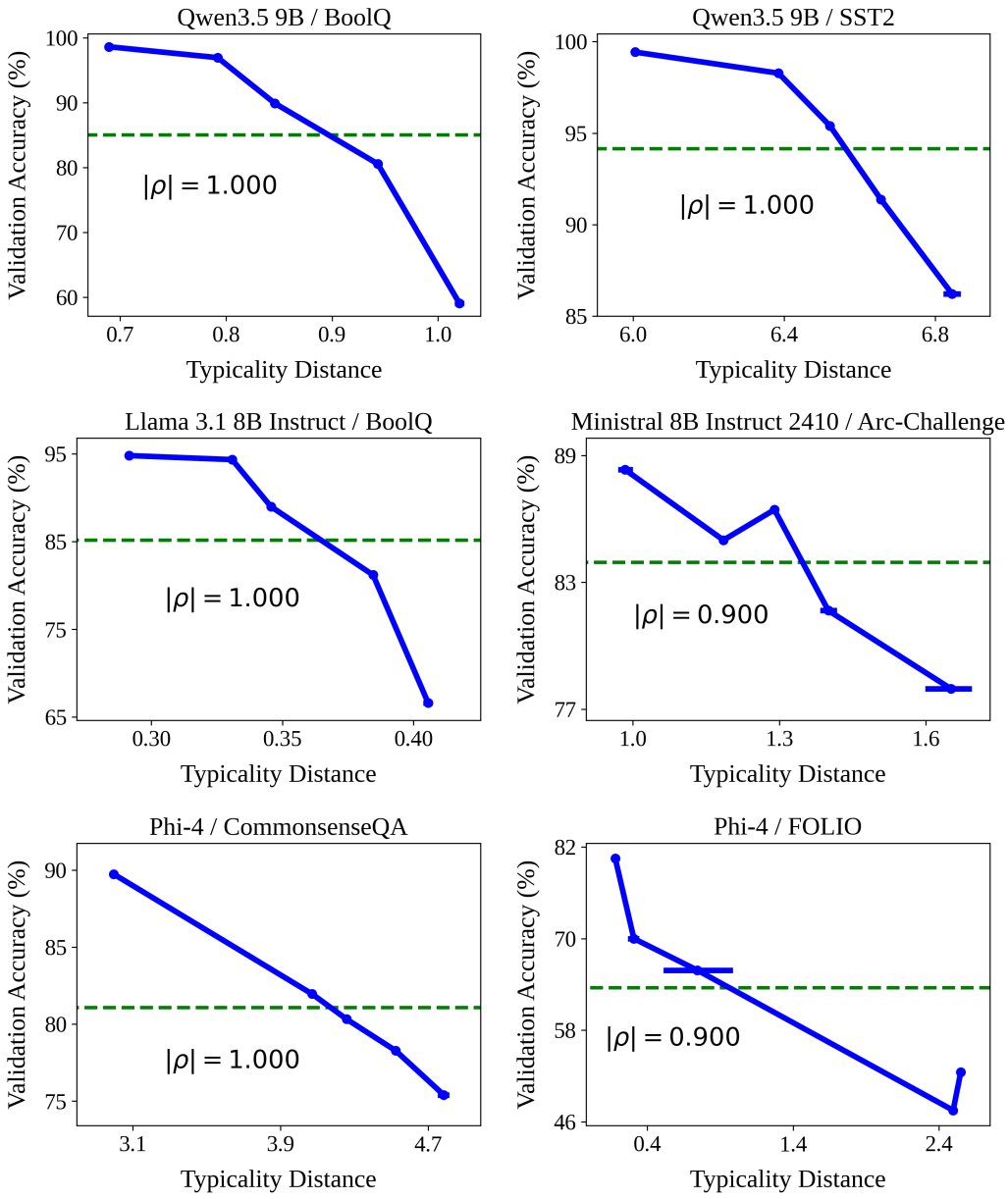


Figure 1: **Typicality is predictive of performance for various models and downstream tasks.** Each plot shows a strong trend of performance predictiveness, with $|\rho|$ providing a quantitative measure for the trend. For each plot, the green dotted line represents the average validation set accuracy, and serves to underscore that lower typicality distance inputs tend to have higher validation accuracy, while higher typicality distance inputs tend to have lower validation accuracy.

4.2 Typicality is Predictive of Input-Level Performance

Figure 1 shows that typicality is highly predictive of input-level performance across a variety of language models and downstream tasks. The plots show a consistent downward trend, showing that lower values of typicality are associated with higher performance, and higher values of typicality are associated with lower performance. This is also supported by the high values for $|\rho|$ obtained across the experiments.

4.3 Comparison of Typicality Distance to Various Baselines

Here we contrast typicality with other baseline measures for predicting input-level performance.

Given an input prompt, let the output entropy for the response produced by a language model be the entropy computed over the output softmax probabilities of the vocabulary.

Furthermore, following the notation of Subsection 3.3, let the MLE distance of an input prompt be given by $\frac{1}{|\mathcal{A}|}\sum_{\ell\in\mathcal{A}}\left|\frac{1}{NK_\ell}\sum_{t=1}^{N}\boldsymbol{z}_{\ell,t}^{\top}\boldsymbol{R}_\ell^{-1}\boldsymbol{z}_{\ell,t}\right|$.

We also contrast the proposed method against text-based self-evaluation (Kadavath et al., 2022) and supervised internal state probing (Azaria & Mitchell, 2023).

Figure 2 shows that typicality is more predictive of input-level performance than the alternative metrics across several models and downstream tasks. This gives further evidence for the utility and novelty of the proposed method.

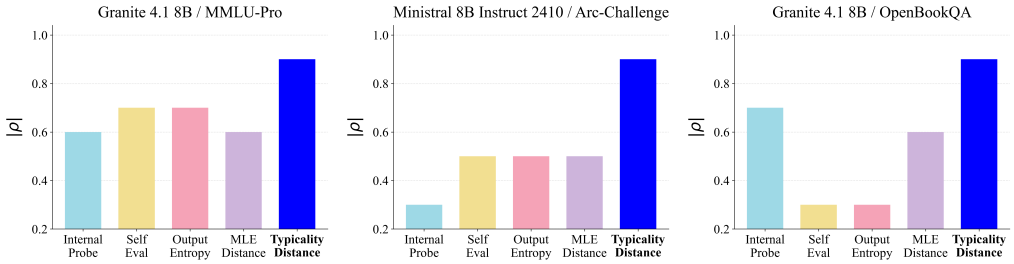


Figure 2: **Typicality is more effective than baseline metrics.** Values for $|\rho|$ for measuring input-level performance predictiveness across various baselines, and for varying models and downstream tasks. Higher is better.

4.4 Ablations & Control Experiments

We next show that the method remains predictive across various experimental configurations, and is robust to a suite of hypothetical confounding factors.

Controlling for Input Sequence Length Figure 3 shows that typicality is predictive of input-level performance irrespective of the input's sequence length. Moreover, this is of value because input prompts to language models can vary significantly in their sequence lengths.

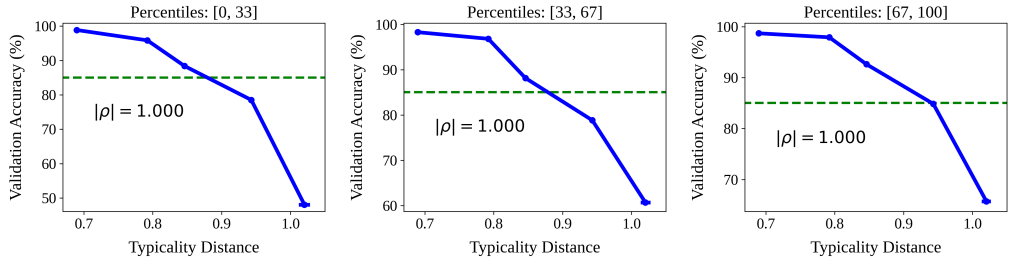


Figure 3: **Typicality remains predictive of input-level performance after controlling for input sequence length.** As in Figure 1, except the validation set was first split into three groups based on input sequence length percentiles (see plot titles). (Qwen3.5 9B / BoolQ.)

Impact of Number of Calibration Samples Used Figure 4 shows that various calibration sizes reliably lead to the same predictive power. Moreover, the plots show that the method only requires a relatively small number of samples to work, making the method applicable even when calibration data is difficult to obtain.

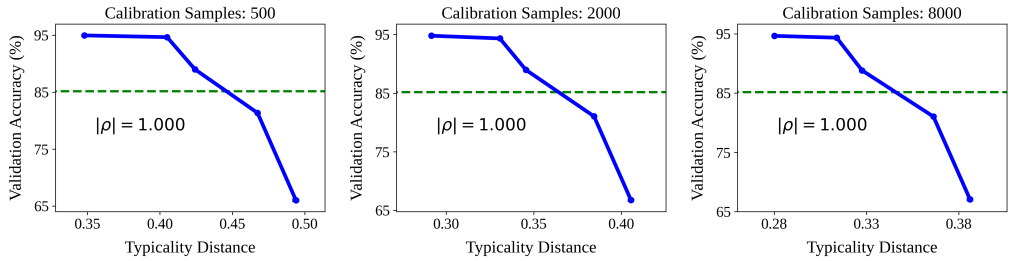


Figure 4: **The proposed method is robust to differences in calibration set size.** Llama 3.1 8B Instruct / BoolQ. As in Figure 1, but for differing sizes for the calibration set (see plot titles).

Impacting of Using Different Layer Types Figure 5 shows that different layer types can be used while still reliably leading to the same predictive power. This also indicates that the method is widely applicable, since the typicality distance of different affine layers can be used in isolation (as in Figure 5) or in aggregation (as demonstrated in all of the other experimental results).

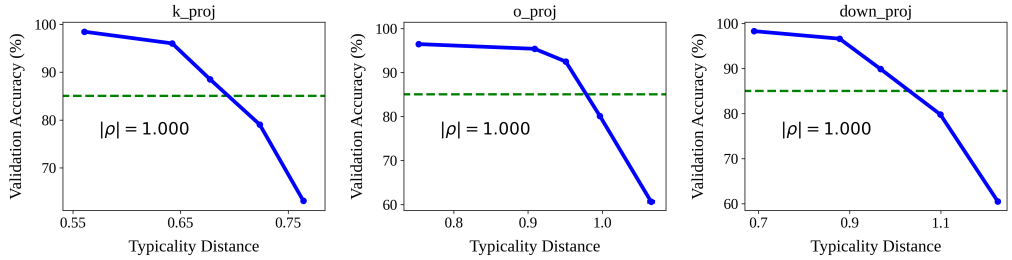


Figure 5: **The proposed method is robust to the use of different layer types.** Qwen3.5 9B / BoolQ. As in Figure 1, but when different layer types are employed (see plot titles).

Impact of Using Subsets of Layers Figure 6 shows that different subsets of the model layers can be used while still leading to the same predictive power, providing further evidence of the method's general effectiveness.

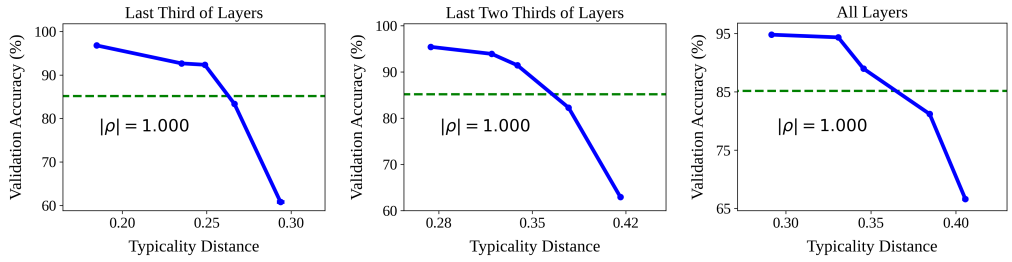


Figure 6: **The proposed method is robust to differing subsets of the model layers being employed.** Llama 3.1 8B Instruct / BoolQ. As in Figure 1, but when differing subsets of the model layers are employed (see plot titles).

4.5 The Assumption of Gaussian Activations Does Not Hinder the Method

We consider what effect gaussianizing activations, in order to better suit the assumptions of the proposed method, might have on the method's utility; and consequently whether the method is robust to non-gaussianity in the activations.

We apply normality normalization (Eftekhari & Papyan, 2025) to gaussianize the activations of each layer before measuring the layer's calibration statistics. We also apply normality normalization at test

time, i.e. before comparing the activations to the calibration statistics. Outside of these measurements, we let the activations of the model proceed through the model unaltered.

Figure 7 shows that the method is not sensitive to deviations from gaussianity in the activations, because explicitly gaussianizing the activations does not impact the method's predictiveness.

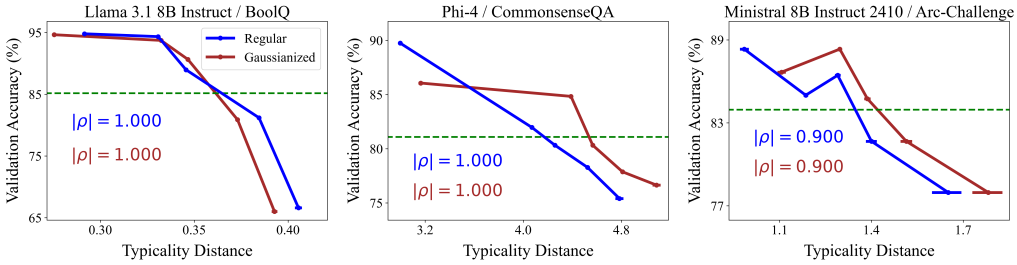


Figure 7: **The assumption of gaussianity does not hinder the method.** As in Figure 1, but when evaluating the method without and with gaussianizing the activations (see plot legend).

4.6 Effect of Using Calibration Set Labels

Here we consider augmenting the method to make use of calibration set labels.

Consider only using correctly classified calibration samples when computing and aggregating layer-wise activation statistics. The motivation for this experiment is to investigate whether there is any benefit in calibrating on correctly classified samples alone, since the incorrectly classified calibration samples would presumably no longer skew the representation statistics.

However, Figure 8 shows that the method works equally well without and with the use of calibration set labels, and thus reinforces the method's strength in not needed calibration set labels.

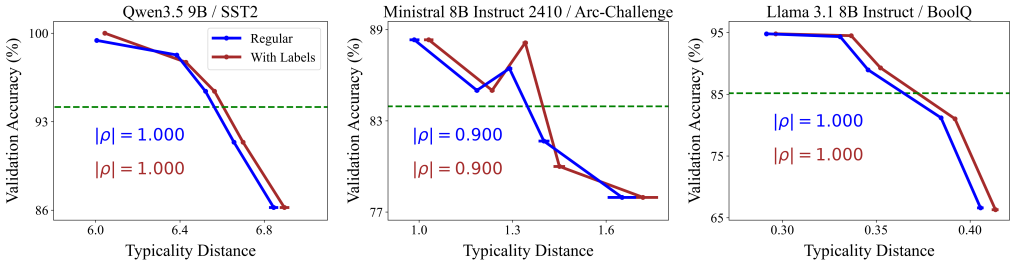


Figure 8: **The proposed method is agnostic to the need for calibration set labels.** As in Figure 1, but without and with conditioning the calibration statistics on cases which were correctly classified (see plot legend).

5 Related Work

Representation-Space Reliability and Generalization There has been an interest in studying whether the outputs or intermediate representations of models contain information about model reliability and generalization. Maximum softmax probability (Hendrycks & Gimpel, 2017) and predictive entropy (Malinin & Gales, 2021) use the confidence of model outputs for predicting model reliability. However, high confidence or low entropy does not necessarily imply correctness, and these signals do not make use of the intermediate representations to inform their predictions. Lee et al. (2018) fit class-conditional Gaussian distributions to hidden features and use Mahalanobis distance to detect out-of-distribution and adversarial inputs, and Sastry & Oore (2020) used Gram matrices to identify inconsistencies between a model's predicted class and its intermediate feature correlations. These methods demonstrate the utility of representation-space statistics for reliability estimation, but they rely on dataset labels, do not address the distinction between pointwise likelihood and

typicality that motivates our approach, and have not been demonstrated to work at the scale of modern large language models. Other work relates performance to local training-sample density (Ji et al., 2022), latent agreement across independently trained models (Deng et al., 2023), and consistency of downstream behavior across learned representation spaces (Park et al., 2024). For language models, Kuhn et al. (2023) proposed semantic entropy, which estimates uncertainty over distinct meanings rather than token sequences, while Kadavath et al. (2022) showed that large language models can often estimate whether their own answers are correct. In contrast, our work predicts input-level performance from the typicality of the intermediate activations induced by each input, and moreover requires no dataset labels, auxiliary models, or model retraining.

Information Content of Intermediate Representations A related area of investigation has examined the geometry and information content of intermediate representations. Intermediate layers can provide strong downstream embeddings (Skean et al., 2025), while intrinsic dimension and effective rank have been connected to fine-tuning dynamics, uncertainty, representation quality, and model performance (Ruppik et al., 2025; Wei et al., 2024; Valeriani et al., 2023; Viswanathan et al., 2025; Cheng et al., 2025). Li & Xu (2026) further connect the spectral geometry of intermediate representations to generalization through effective-rank-based complexity measures. In contrast, our work considers whether a specific input induces activations that are statistically typical for a fixed model, and uses this to predict input-level performance.

Typicality, Likelihood, and OOD Detection Nalisnick et al. (2020) detect out-of-distribution inputs by testing whether their average negative log-likelihood is close to the expected information content under a generative model, and Choi et al. (2019) emphasize that high likelihood need not coincide with regions containing most of the probability mass. In contrast, our approach leverages the distinction between likelihood and typicality, and additionally transfers it from the input distribution to the model's internal representation distribution. Our resulting typicality distance metric therefore measures whether a particular input induces a statistically characteristic computation, rather than whether the input itself is typical under a generative model.

Typicality in Language Modeling Meister et al. (2022) propose the expected information hypothesis to explain why maximum probability decoding can produce less natural text, i.e. human-like text should have information content close to the model distribution's entropy. Building on this idea, Meister et al. (2023) introduce locally typical sampling, which restricts generation to tokens whose surprisal is close to the model's conditional entropy, and Sen et al. (2026) further explore decoding strategies in this setting. Importantly, these works only study the typicality of the output sequence produced by the model. In contrast, we investigate typicality in the intermediate representations of models, and use it as an input-conditioned measure of model-input compatibility, rather than as a decoding rule.

6 Conclusion

We began with the question of what characterizes when a model will be performant for a given set of inputs. Our results support a simple conclusion: performance is strongly related to whether those inputs induce representations that are typical of the model's expected activation regularities. Motivated by the information-theoretic notion of typicality, we introduced representational typicality and the associated typicality distance metric. The metric compares the correlation structure of token-level representations induced by an input with the characteristic statistics estimated from calibration data, and aggregates this deviation across model layers.

Across an array of widely used models and downstream tasks, our experiments comprehensively showed that representational typicality is highly predictive of input-level performance, outcompetes baseline measures, remains predictive across varying experimental configurations, and is robust to a suite of hypothetical confounding factors. Thus, without requiring any model retraining or dataset labels, representational typicality offers an effective framework for predicting input-level performance.

AI Use Statement We did not use any generative AI tools for the following: the ideation phase of the work, motivating the work, the conceptual frameworks used in the work, the methodological development of the work, the experimental setup and design, and the writing of the paper. We only used AI tools to supplement our own efforts in finding and analyzing related work in the literature, and have carefully reviewed all AI-assisted findings pertaining to this.

Reproducibility Statement We outline the representational typicality framework in full in Section 3; including the derivation of the typicality distance metric, the aggregation of the calibration set's representation statistics, and the method's use at test time. Section 4 describes the evaluation protocol, the baseline comparisons, and the ablations and control experiments used for obtaining the experimental results. Appendix A furthermore specifies: the model checkpoints and downstream benchmarks used (including the exact model layers employed), the construction of the calibration and evaluation sets, and the filtering of model refusal cases. Code for reproducing the experimental setup is provided in the Supplementary Materials.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hyunsun Choi, Eric Jang, and Alexander A. Alemi. Waic, but why? generative ensembles for robust anomaly detection, 2019.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Ailin Deng, Miao Xiong, and Bryan Hooi. Great models think alike: Improving model reliability via inter-model latent agreement. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7675–7693. PMLR, 23–29 Jul 2023.

Daniel Eftekhari and Vardan Papyan. On the importance of gaussianizing representations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15018–15037. PMLR, 13–19 Jul 2025.

Granite Team, IBM. Granite-4.1-8b. Hugging Face model repository, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Xu Ji, Razvan Pascanu, R. Devon Hjelm, Balaji Lakshminarayanan, and Andrea Vedaldi. Test sample accuracy scales with training sample density in neural networks. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 629–646. PMLR, 22–24 Aug 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Shaojie Li and Yunbei Xu. Pointwise generalization in deep neural networks, 2026.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. On the probability-quality paradox in language generation, 2022.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 01 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00536.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260.

Mistral AI. Ministral-8b-instruct-2410. Hugging Face model repository, 2024.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality, 2020.

Young-Jin Park, Hao Wang, Shervin Ardeshir, and Navid Azizan. Quantifying representation reliability in self-supervised learning models. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2835–2860. PMLR, 15–19 Jul 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and Milica Gasic. Less is more: Local intrinsic dimensions of contextual language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 67795–67828. Curran Associates, Inc., 2025. doi: 10.52202/085713-2276.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8491–8501. PMLR, 13–18 Jul 2020.

Jaydip Sen, Saptarshi Sengupta, and Subhasis Dasgupta. *Advancing Decoding Strategies: Enhancements in Locally Typical Sampling for LLMs*, pp. 425–484. Springer Nature Switzerland, Cham, 2026. ISBN 978-3-032-13497-4. doi: 10.1007/978-3-032-13497-4_17.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The geometry of tokens in internal representations of large language models, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 39501–39521. Curran Associates, Inc., 2024. doi: 10.52202/079017-1248.

Pierre Wolinski and Julyan Arbel. Gaussian pre-activations in neural networks: Myth or reality? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

A Model & Dataset Details

Models and Downstream Tasks We evaluate representational typicality on a collection of fixed language model checkpoints and downstream benchmarks, chosen to span several model families and task types. The models employed are: Llama 3.1 8B Instruct (Grattafiori et al., 2024), Phi-4 (Abdin et al., 2024), Qwen3.5 9B (Qwen Team, 2026), Ministral 8B Instruct 2410 (Mistral AI, 2024), and Granite 4.1 8B (Granite Team, IBM, 2026). The datasets employed are: BoolQ (Clark et al., 2019), CommonsenseQA (Talmor et al., 2019), ARC-Challenge (Clark et al., 2018), MMLU-Pro (Wang et al., 2024), OpenBookQA (Mihaylov et al., 2018), SST2 (Socher et al., 2013), and FOLIO (Han et al., 2024). For all of our experiments, we bucket the validation set into five equally-sized buckets according to the validation set's typicality distance percentiles, and measure accuracy across each of these buckets.

Affine Layers Used We compute the typicality distance metric for the output activations of the following affine layers. For the Llama 3.1 8B Instruct, Qwen3.5 9B, Ministral 8B Instruct 2410, Granite 4.1 8B, Gemma 2 9B IT, and Gemma 4 12B IT models: q_proj, k_proj, v_proj, o_proj, down_proj. For the Phi-4 model: qkv_proj, o_proj, down_proj.

Filtering of Model Refusal Cases and Prompt Formatting To exclude the effect of model-specific refusal behavior on the relation between representational typicality and performance, we filtered responses exhibiting basic lexical indicators of refusal. For the calibration set, prompts eliciting model refusal were excluded from contributing to layer-wise representation statistics. For the validation set, prompts eliciting refusal were excluded from the evaluation of model performance. All evaluations were done zero-shot and using standard instruction framing. See the Supplementary Materials for further details.

Formation of Calibration and Evaluation Sets With the exception of the ARC-Challenge and FOLIO datasets, unless otherwise specified we sampled 2,000 cases at random from each dataset's training set to form the calibration set. For the ARC-Challenge dataset we used the entire training set of 1,119 cases to form the calibration set, and for the FOLIO dataset we used the entire training set of 1,001 cases to form the calibration set. We use each dataset's entire validation set to evaluate performance.