Measuring LLM Generation Spaces with EigenScore

Anonymous Author(s)
Affiliation
Address
email

Abstract

An LLM's generation space for a given prompt — the range of semantically distinct outputs it could produce — provides a window into the model's implicit task representation. We currently lack a metric for characterizing this space. In this work, we argue that the EigenScore metric (originally developed for hallucination detection) captures the size of this generation space. To develop this understanding, we construct synthetic datasets of prompt pairs with known generation space relationships (complement, subset, etc.). We show that EigenScore reliably predicts a prompt's generation space size, outperforming other metrics like perplexity and entropy. We provide further evidence for this understanding of EigenScore by showing a strong connection between EigenScore and the length of reasoning tokens for the same prompt. Our work uses EigenScore to contribute a cognitive understanding of a model's generation space size and how it relates to reasoning abilities of LLMs.

1 Introduction

For humans, what "comes to mind" (Phillips et al., 2019) when faced with an open-ended question is shaped by factors like feature relevance (Mills and Phillips, 2023) and what we consider likely and good (Bear et al., 2020). However, we currently lack a way to quantify this notion of *generation space* for LLMs. Previous work has alluded to this notion, e.g., finding that post-training techniques like RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) often lead to a more collapsed space (Li et al., 2024) and less random generations (West and Potts, 2025).

Chen et al. (2024) introduce EigenScore, a metric for hallucination detection in factual question-answering (FactualQA) tasks. It is computed by constructing a covariance matrix of the sentence embeddings of K samples and computing the logarithm determinant of the covariance matrix. Conceptually, EigenScore captures the divergence and correlation relationship between embeddings of different sentences based on the LLM's internal representation. Thus, we posit that EigenScore approximates the generation space size for a given prompt. To evaluate this, we construct a dataset of prompt pairs, where each pair's generation space size has a clear set-theoretic relation.

Using our dataset, we show that a variant of EigenScore outperforms other metrics such as perplexity, energy, lexical similarity, etc., at mapping a prompt to the generation space size. Building on our results, we show a positive correlation between the length of reasoning tokens and EigenScore on various tasks. With this correlation, we argue that EigenScore captures the cognitive depth or task difficulty based on a model's internal representation of a prompt. Such a metric that quantifies generation space can enable a better understanding of model behavior (especially as it relates to prompt specificity (Murr et al., 2023; Lu et al., 2021; Santos et al., 2025)), controlling the diversity of LLM generations on open-ended tasks (Lanchantin et al., 2025; Padmakumar and He, 2023; Park et al., 2024), and the relationship between LLM generations and users' cognitive load in human-LLM interactions (Gerlich, 2025; Lee et al., 2025).

Dataset	Prompt A	Prompt B
Complement	Generate a poem about the moon	Generate anything that is not a poem about the moon
FactualQA	What is the fastest land animal?	Name a land animal.
Random Choice	Choose one from the following: cyan, pink	Choose one from the following: red, orange, pink, cyan, purple

Table 1: For each synthetic dataset, prompt A is an example of the prompt with a smaller generation size, and prompt B is a version of the prompt has a larger space size. Note that generation size and prompt length are not correlated.

2 Measuring Generation Space Size

We aim to find a mapping function f from a prompt and a model to the size of the model generation space, given the prompt. More formally, where $\mathcal{M}$ is a model and P is the set of prompts, we aim to find $f_{\mathcal{M}}$ (we drop the subscript below) such that

$$f_{\mathcal{M}} : P \mapsto \mathbb{R} \tag{1}$$

More specifically, given a prompt x, f outputs a real value. Further, we notate the logically possible generation space size for a prompt, abstracted away from $\mathcal{M}$, as $\mathcal{G}(x)$. As it is hard to quantify $\mathcal{G}(x)$, we use set-theoretic operations to create pairs of prompts, $\langle x, y \rangle$, where the set-theoretic relationship between x and y yields a clear comparison in terms of $\mathcal{G}$, such as $\mathcal{G}(x) > \mathcal{G}(y)$. With this set-up, we aim to find $f_{\mathcal{M}}$ that predicts this order on prompts in terms of their generation space size.

To evaluate different metrics (including EigenScore) as a possible f, we design a dataset of prompts with a ground truth of different generation space sizes. We rely on the following intuition: There are many different emails an LLM could generate given the prompt "Generate an email", each differing in length, subject, and style. When the prompt becomes more specified, with additional requirements like "Generate an email to my coworker about my birthday party invitation in three paragraphs", the size of all possible generations becomes more constrained. While tasks like FactualQA questions have a fixed generation space, open-ended questions – especially ones with irreducible randomness (Yadkori et al., 2024) like "Generate a random name" – have much larger generation spaces.

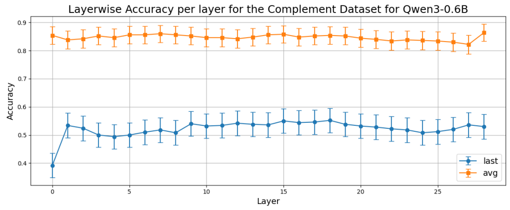
Datasets We construct the following three synthetic datasets, resulting in 3500 prompt pairs (x, x') (see Table 1 for examples) where $\mathcal{G}(x) > \mathcal{G}(y)$, as defined above. The prompt pairs include: (1) **Complement**: We take the complement of a prompt like "Generate a poem about the moon" to be "Generate *anything that is not* a poem about the moon". The latter has a much larger generation space. We generate 500 pairs of base prompts of open-ended generation tasks (e.g. email generation, persona generation, etc.) plus complement versions for each. (2) **FactualQA**: We randomly sampled 1000 questions from TriviaQA (Joshi et al., 2017) and 1000 questions from Natural Questions (Kwiatkowski et al., 2019) as the base prompts and created a synthetic dataset of 500 prompt pairs. (3) **Random Choice**: We can explicitly enumerate a set S in the prompt, along with instructing the model to pick an item from S. By varying the size of S across prompts, we can more directly control the possible generations to choose from. We use this method to create a dataset of 500 pairs of prompts (x, y) where $\mathcal{G}(x) < \mathcal{G}(y)$. More details can be found in Appendix A.1.

Evaluation criteria For each prompt pair, we evaluate a given function f's alignment between the predicted ordering of generation space sizes and the ground-truth ordering using pairwise accuracy, where a pair receives a score of 1 if $f(x) > f(y)$ (where $\mathcal{G}(x) > \mathcal{G}(y)$) and 0 otherwise.

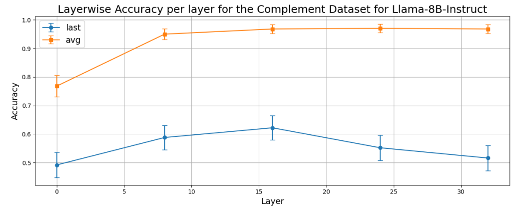
2.1 Mapping function candidates

Following the evaluations in Chen et al. (2024), we compare EigenScore with the following baselines: perplexity, energy (Liu et al., 2020), length-normalized entropy (Malinin and Gales, 2020), lexical similarity (Lin et al., 2023).

In addition to the original EigenScore E_{original}, we explore variants of EigenScore E_{output} and E_{average}. E_{output} (Output EigenScore) is the differential entropy in the sentence embedding space, where the sentence embeddings are obtained through an external sentence embedding model (`Roberta Large V1`). We also perform ablation studies on (1) which layer's embeddings to use and (2) whether to use the last token or average the tokens for the embeddings. We find that individual layers have



(a) Performance does not change with layer for Qwen-0.6B on the complement dataset. An EigenScore is calculated for each of the 29 layers.



(b) Performance does not change with layer for Llama-8B-Instruct on the complement dataset. An Eigen-Score is calculated for layer 0, 8, 16, 24, and 32.

Figure 1: Ablation studies on the layer to take the embeddings from and the token choice (last token versus averaging all tokens).

comparable performance. More critically, taking the mean of the tokens consistently lead to better performance than taking the last token (Figure 1). Thus we use the following variant of EigenScore:

$$E_{\text{average}} \;=\; \frac{1}{|S|\,K}\sum_{\ell\in S}\log\det\Big((JZ^{(\ell)})(JZ^{(\ell)})^{\top}+\alpha I_K\Big) \tag{2}$$

That is, let $H_{\ell,t}^{(n)} \in \mathbb{R}^d$ denote the hidden state for sequence $n \in \{1, \ldots, K\}$, layer $\ell \in \{1, \ldots, L\}$, and token t; let T_n be the sequence length; define $J = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^{\top}$ and a small regularizer $\alpha > 0$; and use the layer subset $S = \{20, \ldots, L-2\}$. Relative to E_{original} , E_{average} changes the representation and the aggregation in two ways: (1) for each layer ℓ and sequence n, replace the single (layer, token) embedding with $\bar{h}_{\ell}^{(n)} = \frac{1}{T_n-1}\sum_{t=1}^{T_n-1} H_{\ell,t}^{(n)}$; (2) for each ℓ, stack $\bar{h}_{\ell}^{(n)}$ across sequences to form $Z^{(\ell)}$ to compute the centered covariance, then average the layerwise scores over S. Thus, unlike E_{original} 's single-layer, single-token log-det, E_{average} aggregates over tokens (per layer) and layers.

2.2 Results

We evaluated these candidate metrics for Llama-8B-Instruct (Dubey et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Qwen3-0.6B (Yang et al., 2025), Qwen3-4B (Yang et al., 2025), and Qwen3-8B (Yang et al., 2025) (the 3 Qwen models allow us to examine the effect of model sizes). We find that E_{output} and E_{average} have the highest performance (Table 2).

Metric	Llama-8B-Instruct	Mistral-7B	Qwen-0.6B	Qwen-4B	Qwen-8B
Perplexity	0.634	0.355	0.599	0.539	0.525
Energy	0.599	0.567	0.619	0.576	0.505
Normalized Entropy	0.660	0.460	0.492	0.571	0.546
Lexical Similarity	0.613	0.558	0.616	0.521	0.544
E_{original}	0.607	0.488	0.568	0.527	0.505
E_{output}	**0.696**	0.543	**0.782**	0.607	0.586
E_{average}	0.688	**0.598**	0.668	**0.609**	**0.598**

Table 2: Average accuracy on the synthetic datasets for each metric for each model (with each dataset weighted equally). All reported values have a ± 0.02 margin of error, computed as $1.96 \times$ the standard error to represent 95% confidence interval. The best-performing metric for each model is in **bold**.

We also evaluated the role of model parameters such as top-k, sample size, and temperature. Consistent with Chen et al. (2024), varying the top-k parameter does not substantially affect performance, while increasing the sample size from 0 to 20 yields steady improvements (Fig 3 and 4). Unlike in hallucination detection, however, EigenScore achieves its best performance on our task at temperature 1.0 rather than 0.5. One possible explanation is that higher sampling randomness produces more diverse embeddings, which may better capture differential entropy when the output space is broader.

3 EigenScore Maps to Reasoning

Reasoning models — machines that "think" — can offer helpful insights into the cognitive taxes of reasoning. Based on human studies (Ericsson and Simon, 1980), we expect tasks with larger generation spaces to require more reasoning effort because the model must navigate a broader landscape of possible outputs, making more complex decisions about which path to pursue [1]. Using EigenScore, we systematize the connection between generation space size and cognitive effort by using the number of reasoning tokens in reasoning models to approximate the amount of deliberation required for a task (Levy et al., 2024; Lynsgøe Raaschou-jensen et al., 2025). Since reasoning models are loosely inspired by human reasoning, and human reasoning — in particular think-out-loud protocols (Van Someren et al., 1994; Wurgaft et al., 2025)— are closely linked with task difficulty and cognitive load, we use reasoning model lengths to establish a connection with the generation space size. In particular, we show that **tasks with larger generation spaces (measured by EigenScore) require more reasoning effort (measured by trace length)**. To test this, we use 6 datasets: Big Reasoning Traces (Allen Institute for AI (allenai), 2025), a modal and conditional reasoning dataset (Holliday et al., 2024), an epistemic reasoning dataset (Suzgun et al., 2024), our complement synthetic dataset, triviaQA (Joshi et al., 2017), and an everyday LLM use dataset (Wang et al., 2024).[2]

For each prompt, we obtain the reasoning traces from Qwen3-0.6B, Qwen3-4B, and Qwen3-8B and plot the correlation between different metrics and reasoning token length. We find that there is a moderate to strong positive correlation between E_{original} and the length of the reasoning tokens (see Figure 2).

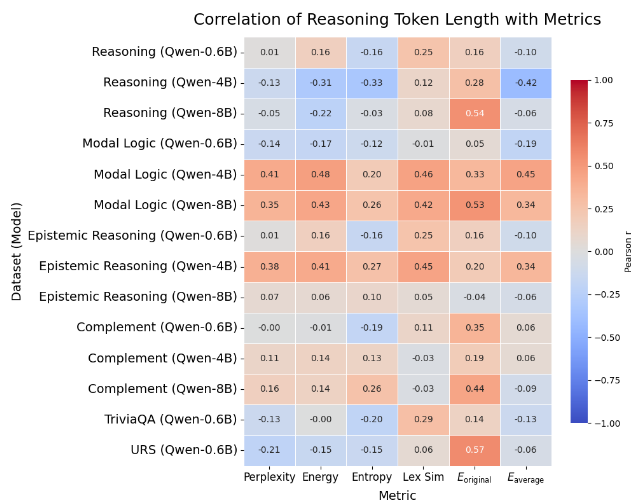


Figure 2: Pearson's r correlation between reasoning token length and various metrics across six datasets and three Qwen3 model sizes. The correlation between E_{original} and reasoning token length remains high across different datasets and models.

This further establishes EigenScore as a proxy for the space of "what comes to mind" for a language model, endowing EigenScore a cognitive interpretation, similar to existing uncertainty metrics (Hale, 2001; Smith and Levy, 2013; Frank, 2010). Beyond its previous applications of hallucination detection, it provides valuable cognitive insights into model representations, decoding, and reasoning. Future work can leverage the cognitive property of EigenScore to explore the connection to prompt specificty and grounding (Shaikh et al., 2023, 2025) and controllable generation by constraining or expanding the space size for different tasks.

[1]Note that longer traces can also reflect reasoning inefficiency (Sui et al., 2025), and high cognitive load could also lead to the absence of verbalization in humans. Despite these factors, we expect there to be a general correlation between reasoning token length and generation space, given the existing connection between reasoning and the nature of the tasks (Sprague et al., 2024; Liu et al., 2024; Aggarwal et al., 2025).

[2]Holliday et al. (2024) and Suzgun et al. (2024) are recent high-quality datasets that incorporate insights from contemporary semantic theory, modal logic, and epistemic logic, making them apt for evaluating reasoning abilities across tasks of varying difficulty.

References

Aggarwal, P., Kim, S., Lanchantin, J., Welleck, S., Weston, J., Kulikov, I., and Saha, S. (2025). Optimalthinkingbench: Evaluating over and underthinking in llms. *arXiv preprint arXiv:2508.13141*.

Allen Institute for AI (allenai) (2025). allenai/big-reasoning-traces: Large permissively licensed reasoning traces. `https://huggingface.co/datasets/allenai/big-reasoning-traces`. Accessed: 2025-08-09.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., and Cushman, F. (2020). What comes to mind? *Cognition*, 194:104057.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. (2024). Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Ericsson, K. A. and Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3):215.

Frank, S. L. (2010). Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, pages 81–89.

Gerlich, M. (2025). Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Holliday, W. H., Mandelkern, M., and Zhang, C. E. (2024). Conditional and modal reasoning in large language models. *arXiv preprint arXiv:2401.17169*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lanchantin, J., Chen, A., Dhuliawala, S., Yu, P., Weston, J., Sukhbaatar, S., and Kulikov, I. (2025). Diverse preference optimization. *arXiv preprint arXiv:2501.18101*.

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. (2025). The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–22.

Levy, M., Jacoby, A., and Goldberg, Y. (2024). Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Li, M., Shi, W., Pagnoni, A., West, P., and Holtzman, A. (2024). Predicting vs. acting: A trade-off between world modeling & agent modeling. *arXiv preprint arXiv:2407.02446*.

Lin, Z., Trivedi, S., and Sun, J. (2023). Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., and Griffiths, T. L. (2024). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Lynsgøe Raaschou-jensen, H. P., Fierro, C., and Søgaard, A. (2025). Predicting thinking time in reasoning models. *arXiv e-prints*, pages arXiv–2506.

Malinin, A. and Gales, M. (2020). Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Mills, T. and Phillips, J. (2023). Locating what comes to mind in empirically derived representational spaces. *Cognition*, 240:105549.

Murr, L., Grainger, M., and Gao, D. (2023). Testing llms on code generation with varying levels of prompt specificity. *arXiv preprint arXiv:2311.07599*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Padmakumar, V. and He, H. (2023). Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Park, P. S., Schoenegger, P., and Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770.

Phillips, J., Morris, A., and Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*, 23(12):1026–1040.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Santos, G. M., Julia, R. M. D. S., and Nascimento, M. Z. d. (2025). Diverse prompts: Illuminating the prompt space of large language models with map-elites. *arXiv preprint arXiv:2504.14367*.

Shaikh, O., Gligorić, K., Khetan, A., Gerstgrasser, M., Yang, D., and Jurafsky, D. (2023). Grounding gaps in language model generations. *arXiv preprint arXiv:2311.09144*.

Shaikh, O., Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. (2025). Navigating rifts in human-llm grounding: Study and benchmark. *arXiv preprint arXiv:2503.13975*.

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

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Sprague, Z., Yin, F., Rodriguez, J. D., Jiang, D., Wadhwa, M., Singhal, P., Zhao, X., Ye, X., Mahowald, K., and Durrett, G. (2024). To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Sui, Y., Chuang, Y.-N., Wang, G., Zhang, J., Zhang, T., Yuan, J., Liu, H., Wen, A., Zhong, S., Chen, H., et al. (2025). Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Suzgun, M., Gur, T., Bianchi, F., Ho, D. E., Icard, T., Jurafsky, D., and Zou, J. (2024). Belief in the machine: Investigating epistemological blind spots of language models. *arXiv preprint arXiv:2410.21195*.

Van Someren, M. W., Barnard, Y. F., Sandberg, J. A., et al. (1994). The think aloud method: a practical approach to modelling cognitive processes. *London: AcademicPress*, 11(6).

Wang, J., Mo, F., Ma, W., Sun, P., Zhang, M., and Nie, J.-Y. (2024). A user-centric multi-intent benchmark for evaluating large language models. *arXiv preprint arXiv:2404.13940*.

West, P. and Potts, C. (2025). Base models beat aligned models at randomness and creativity. *arXiv preprint arXiv:2505.00047*.

Wurgaft, D., Prystawski, B., Gandhi, K., Zhang, C. E., Tenenbaum, J. B., and Goodman, N. D. (2025). Scaling up the think-aloud method. *arXiv preprint arXiv:2505.23931*.

Yadkori, Y. A., Kuzborskij, I., György, A., and Szepesvári, C. (2024). To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Generation space size experiment details

A.1 Dataset construction details

Complement We generated the base prompts following templates about email, poem, Python program, short story, and persona generation. Each prompt is constructed following an existing template that adds modifiers to the item generation (full details below). Then, the complement version of the prompt is constructed by adding "anything that is not". Tab 4 shows some examples of the prompt pairs.

Table 3: The template used for the Complement dataset. Each base prompt is constructed by choosing a combination of a topic, context, qualifier, and outline

(a) **An email**

Field	Example values
Topics	job opportunities; an upcoming conference; a new product launch; a team milestone
Contexts	at a tech firm; for remote engineers; in the non-profit sector
Qualifiers	includes a discussion of my qualifications; asks about remote-work policies
Outlines	Greeting, Purpose, Qualifications, Next steps; Subject, Body, Closing

(b) **A poem**

Field	Example values
Topics	autumn leaves; lost love; a starry night; the ocean's whispers
Contexts	in a small town; during wartime; over the desert
Qualifiers	employs vivid imagery; uses iambic pentameter; is limited to 14 lines
Outlines	haiku (5-7-5); limerick; free verse

(c) **A Python program**

Field	Example values
Topics	sorting a list; scraping a website; converting CSV to JSON; analyzing text sentiment
Contexts	using merge sort; handling pagination; with nested objects
Qualifiers	includes docstrings; uses type hints; avoids external libraries
Outlines	main(), helper functions, guard block; CLI interface

(d) **A short story**

Field	Example values
Topics	a time-travel mishap; an unlikely friendship; a dystopian future; a family reunion
Contexts	in Victorian London; between a robot and a child; ruled by algorithms
Qualifiers	written in first person; contains a twist ending; under 500 words
Outlines	Freytag's pyramid; journal entries; letters format

(e) **A persona**

Field	Example values
Topics	a tech-savvy college student; a health-conscious parent; a budget traveler; a small business owner
Contexts	majoring in computer science; with two toddlers; backpacking in Southeast Asia
Qualifiers	includes demographic info; identifies pain points; lists preferred communication channels
Outlines	Background, Goals, Challenges; bullet points; short narrative example

Table 4: Original prompts and their complement versions.

Original Prompt	Complement Prompt
Generate a poem about the moon	Generate anything that is not a poem about the moon
Generate a story set in a dystopian future	Generate anything that is not a story set in a dystopian future
Generate a Python function to sort a list	Generate anything that is not a Python function to sort a list
Generate an email to request a recommendation letter	Generate anything that is not an email to request a recommendation letter
Generate a recipe using only 5 ingredients	Generate anything that is not a recipe using only 5 ingredients
Generate a haiku about the ocean	Generate anything that is not a haiku about the ocean
Generate a motivational quote	Generate anything that is not a motivational quote
Generate a summary of the French Revolution	Generate anything that is not a summary of the French Revolution

FactualQA rewrites Below we present the prompt used to label and rewrite the prompt pair for questions in TriviaQA and Natural Questions:

Instruction: First, determine if the following question has only one possible correct answer. For example, the question “What is the name of the first founding father?” has only one correct answer, while the question “What is the name of one founding father?” has multiple correct answers.
Output `1` if there is more than one correct answer (i.e., multiple possible generations), and output `0` if there is only one correct answer.
If there is only one correct answer, make minimal changes to the question so that the new question has more than one possible correct answer. For example, change “Name the largest river in Brazil” to “Name a river in Brazil,” where the former has only one correct answer while the latter has many.
If there is more than one correct answer, make minimal changes to the question so that the new question has only one correct answer. For example, change “Who is one founding father of the United States” to “Who is the first founding father of the United States.”
Output format: Output the number and the new question, separated by a comma.
Example:

Original prompt: `Name the largest river in Brazil.`
Output: `0, Name a river in Brazil`

Question: `{question}`

We used GPT-4o to classify whether a question from TriviaQA and Natural Questions only has one correct answer using the prompt above. Then, we instructed the model to generate a new version of the question, where it would constrain the space if the original question has more than one possible answer, and vice versa. Out of the sample of 1000 questions from Natural Questions, 579 were labeled as having only one correct answer, and out of the sample of 1000 questions from TriviaQA, 860 were labeled as having only one correct answer. Tab 5 shows more example prompt pairs.

Table 5: Original and rewritten questions.

Original Question	Rewritten Question
Who composed the Hungarian Dances in 1869?	Name a composer of Hungarian Dances.
What physical feature do all pinnipeds have?	What is the primary habitat of a specific pinniped species?
Who wrote the 1980 children’s book ‘The Twits’?	Name a children’s book written by Roald Dahl.
What were the first names of English author H G Wells?	What is one first name of English author H G Wells?

FactualQA Synthetic The synthetic dataset for question pairs where one question has one single correct answer and the other has multiple correct answers is constructed using a template with a superlative version of the question and a non-superlative one. To augment the dataset, we populated variables like country or continent with a randomly selected country or continent name from a pool of candidates. The full prompt template pairs and the country and continent candidates are in Tab 6. We used a total of 60 base prompts, 30 country names, and 6 continent names to populate 1000 unique prompt pairs for evaluation.

Table 6: Templates used to construct the factualQA Synthetic dataset.

(a) Example template pairs. Prompt A has a smaller generation space size than prompt B.

Prompt A	**Prompt B**
Who was the first president of {country}?	Name a president of {country}.
What is the capital of {country}?	Name a city in {country}.
What is the largest river in {country}?	Name a river in {country}.
What is the tallest mountain in {country}?	Name a mountain in {country}.
What is the longest river in {continent}?	Name a river in {continent}.
What is the most populated city in {country}?	Name a city in {country}.
What is the highest mountain in {continent}?	Name a mountain in {continent}.
What is the official language of {country}?	Name a language spoken in {country}.
What is the currency of {country}?	Name a currency used in {continent}.
Who was the 16th president of the United States?	Who was a president of the United States?

(b) Countries and continents to replace the placeholder.

Type	**List**
Countries	Argentina, Australia, Bangladesh, Belgium, Brazil, Canada, Chile, China, Colombia, Denmark, Egypt, Ethiopia, Finland, France, Germany, India, Indonesia, Iran, Iraq, Italy, Japan, Kenya, Mexico, Netherlands, Nigeria, Pakistan, Russia, South Africa, South Korea, United Kingdom
Continents	Asia, Africa, Europe, North America, South America, Australia

Table 7: Example categories and their items used to construct synthetic prompts for the random choice experiment.

Category	**Items**
Animals	cat, dog, sheep, horse, bird, whale, lion, tiger, bear, elephant, giraffe, zebra
Colors	red, blue, green, yellow, black, white, orange, purple, pink, gray, brown, cyan
Numbers	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
Fruits	apple, banana, cherry, grape, kiwi, lemon, mango, orange, pear, peach, plum, melon
Vehicles	car, truck, bus, motorcycle, bicycle, scooter, van, train, boat, plane, helicopter, submarine

Random Choice To construct the prompt pairs for the random choice experiment, we used a word bank of four categories: animals, colors, numbers, and vehicles. Each category contains 10 to 20 common words. The prompt pairs are constructed by first randomly choosing a category, then randomly choosing 2 (for prompt A) or 10 (for prompt B) items from the list to append to the sentence "Choose one from the following:". The full list of words are in Tab 7. To verify that each option has an equal probability of being chosen and that the space size is truly bigger for the bigger set, we calculate the logits distribution for each question and find that the logits distribution are uniform across the possible options available in the set.

A.2 Additional Datasets

Here we introduce three additional datasets, constructed based on three different set properties: subset, union, and intersection.

Subset The subset dataset is constructed by appending additional information (adding additional requirements) to each base generation task. The base generation tasks are the same as the complement dataset: email, poem, Python Program, short story, or persona generation. For each group, we create five prompts of increased specificity level by appending more and more requirements. We evaluate

Table 8: Evaluation results for three additional datasets: subset, union, and intersection for four models.

(a) **Subset**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-4B	Mistral-7B
perplexity	0.483 ± 0.02	0.374 ± 0.02	0.477 ± 0.02	0.450 ± 0.02
energy	0.501 ± 0.02	0.386 ± 0.02	0.472 ± 0.02	0.574 ± 0.02
normalized entropy	0.448 ± 0.02	0.416 ± 0.02	0.417 ± 0.02	0.471 ± 0.02
lexical similarity	0.706 ± 0.02	0.557 ± 0.02	0.547 ± 0.02	0.688 ± 0.02
E_{original}	0.464 ± 0.02	0.522 ± 0.02	0.456 ± 0.02	0.512 ± 0.02
E_{output}	0.718 ± 0.02	$\mathbf{0.684 \pm 0.02}$	0.571 ± 0.02	0.771 ± 0.02
E_{average}	$\mathbf{0.740 \pm 0.02}$	0.682 ± 0.02	$\mathbf{0.610 \pm 0.02}$	$\mathbf{0.779 \pm 0.02}$

(b) **Union**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-4B	Mistral-7B
perplexity	0.533 ± 0.04	0.540 ± 0.04	0.567 ± 0.04	0.549 ± 0.05
energy	0.524 ± 0.04	0.550 ± 0.04	0.563 ± 0.05	0.530 ± 0.05
normalized entropy	0.526 ± 0.04	0.480 ± 0.04	0.566 ± 0.05	0.505 ± 0.03
lexical similarity	0.585 ± 0.05	0.540 ± 0.04	0.616 ± 0.05	$\mathbf{0.556 \pm 0.04}$
E_{original}	0.554 ± 0.04	0.525 ± 0.04	0.568 ± 0.04	0.504 ± 0.04
E_{output}	$\mathbf{0.635 \pm 0.05}$	$\mathbf{0.616 \pm 0.04}$	$\mathbf{0.677 \pm 0.05}$	0.506 ± 0.04
E_{average}	0.569 ± 0.05	0.488 ± 0.04	0.610 ± 0.04	0.527 ± 0.03

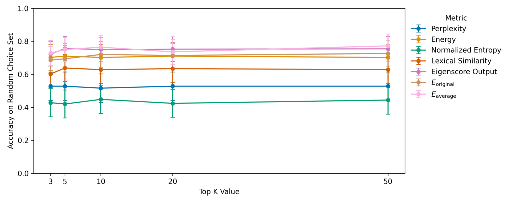
(c) **Intersection**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-4B	Mistral-7B
perplexity	0.574 ± 0.04	0.476 ± 0.04	0.477 ± 0.04	0.412 ± 0.04
energy	0.578 ± 0.04	0.422 ± 0.04	0.457 ± 0.04	0.564 ± 0.04
normalized entropy	0.615 ± 0.04	0.463 ± 0.04	0.439 ± 0.04	0.504 ± 0.04
lexical similarity	0.646 ± 0.04	0.450 ± 0.04	0.461 ± 0.04	0.683 ± 0.03
E_{original}	0.473 ± 0.04	0.558 ± 0.03	0.475 ± 0.04	0.439 ± 0.04
E_{output}	0.596 ± 0.05	0.495 ± 0.04	0.452 ± 0.04	0.655 ± 0.04
E_{average}	$\mathbf{0.687 \pm 0.04}$	$\mathbf{0.571 \pm 0.04}$	$\mathbf{0.505 \pm 0.04}$	$\mathbf{0.698 \pm 0.04}$

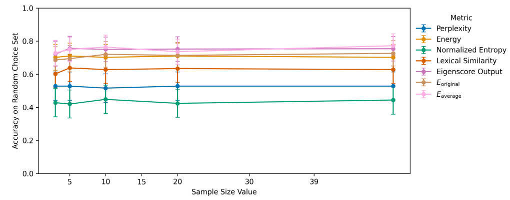
the pairwise accuracy (whether the more specific prompt has a lower score, given that its generation space should be a subset of its supersets). The dataset comprises of 180 sets of prompts and a total of 900 prompts.

Union The union dataset is constructed by taking the union (connecting generation tasks with the keyword "or"), which should theoretically increase the generation space. For each group, we create 4 base prompts (e.g. "come up with an idea for breakfast", "come up with an idea for lunch", "come up with an idea for afternoon snack", and "come up with an idea for dinner"), then we create a total of 15 prompts, including each possible combination of the base prompts, connected through "or". We evaluate whether the scores for the bigger sets (e.g. "come up with an idea for breakfast or lunch or dinner or afternoon snack") are bigger using pairwise comparisons. We created 60 distinct sets with 15 prompts in each set.

Intersection Each group in the intersection dataset comprises of 4 base prompts, which are overlapping requirements (e.g. "compose an email", "please write a piece that is 200 words long", "please write something that is three paragraphs in length", and "compose a piece using formal language"). Then, we can take the intersections by connecting each base prompt with the keyword "and", which effectively constrains the generation space by adding additional requirements. We created 60 unique sets (each with 15 prompts) and evaluate the pairwise comparison based on whether the score for each subset is smaller than the score of its supersets.

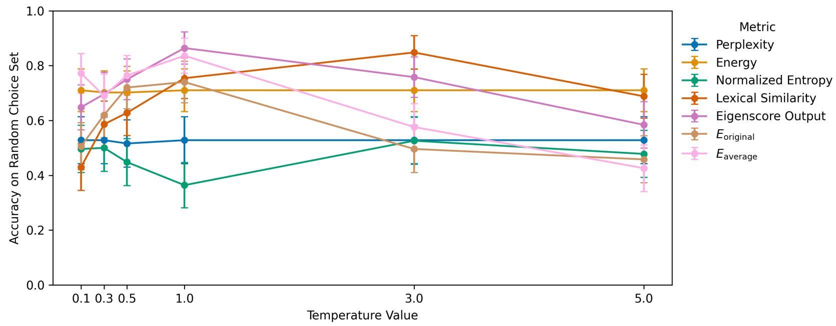


(a) Performance does not change with top-k.

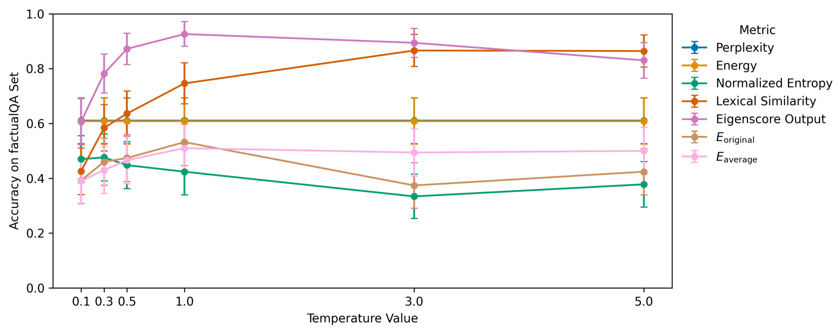


(b) Performance increases with sample size.

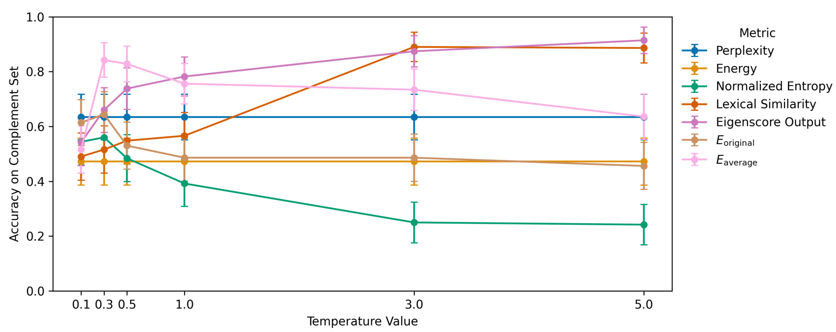
Figure 3: Ablation studies on top K and sample size.



(a) Performance is best when temperature is 1 for the random choice dataset.



(b) Performance is best when temperature is 1 for the factualQA synthetic dataset.



(c) Performance is best when temperature is 0.5 for the complement dataset.

Figure 4: Ablation for temperature

A.3 Prompt Length

In this section, we provide clarity on the connection between EigenScore and prompt length in our tasks. To address the concern that longer prompts contain more information and are correlated with various uncertainty measurements like entropy (Shannon, 1951), we intentionally construct prompt pairs in the Complement Set and Random Choice Set such that the longer prompt is the one with the bigger generation space. In the factualQA prompt pairs, the prompts have similar lengths, so prompt length is not a good predictor for the task of modeling generation space size. In Tab 9, we present the pairwise accuracy achieved by prompt length alone and correlation between E_{average} and prompt length, providing evidence that prompt length is not a confounding factor.

Table 9: Pairwise accuracy (mean $\pm$ 95% CI) and correlation of the prompt length. The accuracy is the score when using prompt length as the mapping function f, and correlation is between the prompt length and E_{average} .

Dataset	**Pairwise Accuracy**	**Correlation**
Complement	1.00	0.0066
Natural Questions	0.23 ± 0.03	-0.12
SyntheticQA	0.030 ± 0.03	0.058
TriviaQA	0.27 ± 0.03	-0.052
Random Choice	1.00	0.56

A.4 Full results

We present the full results on each dataset and include two additional models in Tab 10: Qwen3-0.6B (with reasoning) and Qwen3-4B (with reasoning). We observe that when reasoning mode is turned on, the EigenScore performance deteriorates.

B Length of Reasoning Tokens

Tab 11 shows the dataset used to calculate correlations and the size of each dataset, and Tab 12 shows some examples of prompts and their reasoning token lengths and EigenScores.

Table 11: The datasets used to examine the correlation with reasoning token lengths.

Dataset	**Source**	**Size**
Big Reasoning Traces	Allen Institute for AI (allenai) (2025)	1000
Modal Logic	Holliday et al. (2024)	3000
Epistemic Reasoning	Suzgun et al. (2024)	3000
Complement	synthetic	900
TriviaQA	Joshi et al. (2017)	1000
User-Intent	Wang et al. (2024)	1000

B.1 Exploratory analysis of token length and EigenScore

User Intent Dataset Wang et al. (2024) provides prompt and user-intent pairs, where user-intent are labels that each participant reported based on the given taxonomy. The possible labels are: Ask for Advice, FactualQA, Leisure, Seek Creativity, Solve Professional Problem, and Text Assistant. Below we calculate the average thinking token length and EigenScore for prompts in each category. Tab 13 shows that categories with longer reasoning token lens, such as `Solve Professional Problem` and `Seek Creativity` also have greater EigenScores. Similarly, tasks with shorter reasoning token length — including `Ask for Advice` and `FactualQA` — also have lower EigenScores. Tasks from `Solve Professional Problem` and `Seek Creativity` are more difficult tasks that often require more deliberation. The finding supports our hypothesis that there is a strong connection between EigenScore, reasoning token length, and the generation space size. The different EigenScores for each user-intent type suggests that EigenScore can be predictive of user intent behind a task.

Table 10: Accuracy breakdown for each dataset and for each model.

(a) **Complement**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-0.6B (R)	Qwen-4B	Qwen-4B (R)	Mistral-7B	Qwen-8B
perplexity	0.674 ± 0.04	0.594 ± 0.04	0.632 ± 0.04	0.530 ± 0.04	0.858 ± 0.03	0.412 ± 0.04	0.576 ± 0.04
energy	0.670 ± 0.04	0.516 ± 0.04	0.624 ± 0.04	0.530 ± 0.04	0.898 ± 0.03	0.540 ± 0.04	0.456 ± 0.04
normalized entropy	0.772 ± 0.04	0.354 ± 0.04	0.352 ± 0.04	0.690 ± 0.04	0.778 ± 0.04	0.314 ± 0.04	0.532 ± 0.04
lexical similarity	0.880 ± 0.03	0.668 ± 0.04	0.716 ± 0.04	0.736 ± 0.04	0.704 ± 0.04	0.560 ± 0.04	0.712 ± 0.04
E_{original}	0.566 ± 0.04	0.596 ± 0.04	0.452 ± 0.04	0.574 ± 0.04	0.434 ± 0.04	0.550 ± 0.04	0.500 ± 0.04
E_{output}	$\mathbf{0.954 \pm 0.02}$	$\mathbf{0.908 \pm 0.03}$	$\mathbf{0.958 \pm 0.02}$	0.860 ± 0.03	$\mathbf{0.930 \pm 0.02}$	0.758 ± 0.04	0.790 ± 0.04
E_{average}	0.940 ± 0.02	0.810 ± 0.03	0.754 ± 0.04	$\mathbf{0.880 \pm 0.03}$	0.876 ± 0.03	$\mathbf{0.762 \pm 0.04}$	$\mathbf{0.806 \pm 0.03}$

(b) **SyntheticQA**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-0.6B (R)	Qwen-4B	Qwen-4B (R)	Mistral-7B	Qwen-8B
perplexity	0.660 ± 0.04	0.610 ± 0.04	$\mathbf{0.610 \pm 0.04}$	0.318 ± 0.04	0.428 ± 0.04	0.086 ± 0.02	0.334 ± 0.04
energy	0.656 ± 0.04	0.608 ± 0.04	0.486 ± 0.04	$\mathbf{0.410 \pm 0.04}$	0.334 ± 0.04	0.484 ± 0.04	0.380 ± 0.04
normalized entropy	0.670 ± 0.04	0.434 ± 0.04	0.532 ± 0.04	0.290 ± 0.04	0.440 ± 0.04	0.362 ± 0.04	0.438 ± 0.04
lexical similarity	0.506 ± 0.04	0.738 ± 0.04	0.572 ± 0.04	0.290 ± 0.04	0.418 ± 0.04	$\mathbf{0.542 \pm 0.04}$	0.274 ± 0.04
E_{original}	0.472 ± 0.04	0.506 ± 0.04	0.518 ± 0.04	0.256 ± 0.04	0.508 ± 0.04	0.356 ± 0.04	0.412 ± 0.04
E_{output}	0.718 ± 0.04	$\mathbf{0.922 \pm 0.02}$	0.510 ± 0.04	0.358 ± 0.04	$\mathbf{0.796 \pm 0.04}$	0.280 ± 0.04	0.388 ± 0.04
E_{average}	$\mathbf{0.782 \pm 0.04}$	0.502 ± 0.04	0.556 ± 0.04	0.284 ± 0.04	0.606 ± 0.04	0.468 ± 0.04	$\mathbf{0.438 \pm 0.04}$

(c) **Random Choice**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-0.6B (R)	Qwen-4B	Qwen-4B (R)	Mistral-7B	Qwen-8B
perplexity	0.678 ± 0.04	0.516 ± 0.04	$\mathbf{0.546 \pm 0.04}$	0.696 ± 0.04	0.654 ± 0.04	0.464 ± 0.04	0.458 ± 0.04
energy	0.594 ± 0.04	0.702 ± 0.04	0.452 ± 0.04	$\mathbf{0.762 \pm 0.04}$	$\mathbf{0.712 \pm 0.04}$	$\mathbf{0.658 \pm 0.04}$	0.312 ± 0.04
normalized entropy	0.642 ± 0.04	0.378 ± 0.04	0.420 ± 0.04	0.690 ± 0.04	0.318 ± 0.04	0.628 ± 0.04	0.470 ± 0.04
lexical similarity	0.666 ± 0.04	0.738 ± 0.04	0.224 ± 0.04	0.680 ± 0.04	0.106 ± 0.03	0.622 ± 0.04	0.470 ± 0.04
E_{original}	$\mathbf{0.680 \pm 0.04}$	0.726 ± 0.04	0.510 ± 0.04	0.618 ± 0.04	0.656 ± 0.04	0.562 ± 0.04	0.542 ± 0.04
E_{output}	$\mathbf{0.680 \pm 0.04}$	$\mathbf{0.856 \pm 0.03}$	0.236 ± 0.04	0.704 ± 0.04	0.550 ± 0.04	0.600 ± 0.04	0.562 ± 0.04
E_{average}	0.628 ± 0.04	0.838 ± 0.03	0.234 ± 0.04	0.650 ± 0.04	0.378 ± 0.04	0.546 ± 0.04	$\mathbf{0.572 \pm 0.04}$

(d) **Natural Questions rewrite**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-0.6B (R)	Qwen-4B	Qwen-4B (R)	Mistral-7B	Qwen-8B
perplexity	0.521 ± 0.03	0.656 ± 0.03	0.524 ± 0.03	0.629 ± 0.03	0.472 ± 0.03	0.549 ± 0.03	0.704 ± 0.03
energy	0.507 ± 0.03	0.638 ± 0.03	0.558 ± 0.03	$\mathbf{0.671 \pm 0.03}$	0.522 ± 0.03	0.494 ± 0.03	$\mathbf{0.768 \pm 0.03}$
normalized entropy	0.568 ± 0.03	$\mathbf{0.721 \pm 0.03}$	0.615 ± 0.03	0.624 ± 0.03	0.465 ± 0.03	0.592 ± 0.03	0.725 ± 0.03
lexical similarity	0.379 ± 0.03	0.302 ± 0.03	0.425 ± 0.03	0.326 ± 0.03	$\mathbf{0.588 \pm 0.03}$	0.577 ± 0.03	0.704 ± 0.03
E_{original}	$\mathbf{0.675 \pm 0.03}$	0.528 ± 0.03	0.506 ± 0.03	0.595 ± 0.03	0.537 ± 0.03	0.509 ± 0.03	0.566 ± 0.03
E_{output}	0.461 ± 0.03	0.525 ± 0.03	0.549 ± 0.03	0.550 ± 0.03	0.426 ± 0.03	0.621 ± 0.03	0.616 ± 0.03
E_{average}	0.411 ± 0.03	0.635 ± 0.03	$\mathbf{0.658 \pm 0.03}$	0.648 ± 0.03	0.565 ± 0.03	$\mathbf{0.637 \pm 0.03}$	0.627 ± 0.03

(e) **TriviaQA rewrite**

Metric	Llama-8B-Instruct	Qwen-0.6B	Qwen-0.6B (R)	Qwen-4B	Qwen-4B (R)	Mistral-7B	Qwen-8B
perplexity	0.638 ± 0.03	0.621 ± 0.03	0.621 ± 0.03	0.520 ± 0.03	0.572 ± 0.03	0.263 ± 0.03	0.552 ± 0.03
energy	0.567 ± 0.03	0.629 ± 0.03	0.629 ± 0.03	0.506 ± 0.03	0.515 ± 0.03	$\mathbf{0.660 \pm 0.03}$	$\mathbf{0.607 \pm 0.03}$
normalized entropy	0.648 ± 0.03	0.572 ± 0.03	0.572 ± 0.03	0.562 ± 0.03	0.507 ± 0.03	0.405 ± 0.03	0.565 ± 0.03
lexical similarity	0.636 ± 0.03	0.632 ± 0.03	0.632 ± 0.03	0.573 ± 0.03	0.568 ± 0.03	0.490 ± 0.03	0.560 ± 0.03
E_{original}	0.640 ± 0.03	0.483 ± 0.03	0.483 ± 0.03	$\mathbf{0.590 \pm 0.03}$	0.456 ± 0.03	0.464 ± 0.03	0.504 ± 0.03
E_{output}	0.665 ± 0.03	$\mathbf{0.697 \pm 0.03}$	$\mathbf{0.697 \pm 0.03}$	0.562 ± 0.03	$\mathbf{0.580 \pm 0.03}$	0.456 ± 0.03	0.576 ± 0.03
E_{average}	$\mathbf{0.677 \pm 0.03}$	0.556 ± 0.03	0.556 ± 0.03	0.582 ± 0.03	0.539 ± 0.03	0.575 ± 0.03	0.546 ± 0.03

Table 13: Token length and EigenScore by user intent for data from Wang et al. (2024) (mean $\pm$ 95% CI). Both EigenScore and reasoning token lengths are calculated for Qwen3-8B. After filtering to only include English prompts, $N = 1000$

User Intent	**Token Len**	**EigenScore**
Ask for Advice	298.15 ± 31.1	-1.61 ± 0.02
FactualQA	295.42 ± 45.3	-1.63 ± 0.02
Leisure	359.19 ± 117.6	-1.59 ± 0.04
Seek Creativity	383.09 ± 132.8	-1.56 ± 0.05
Solve Professional Problem	656.10 ± 180.9	-1.50 ± 0.06
Text Assistant	328.38 ± 47.4	-1.64 ± 0.05

Modal and Conditional Reasoning Dataset Modal and conditional reasoning tasks differ in difficulty, with some tasks presumably requiring more deliberation than others. With this guiding

Table 12: Examples of token length and EigenScore for different prompts from the Complement Dataset and Modal Logic Dataset. All examples show cases where the prompt with bigger generation space correpond to longer reasoning token length and higher EigenScores. In the modal logic dataset, uDSmu tasks are significantly more difficult than DS tasks. (The model is Qwen3-8B). The prompt with longer reasoning length and EigenScore is in **bold** for each pair.

Task Type	Prompt	Token Len	EigenScore
Complement	Generate a story set in a dystopian future	342	-1.28
Complement	Generate anything that is not a story set in a dystopian future	**452**	**-1.22**
Complement	Generate a haiku about the ocean	359	-1.21
Complement	Generate anything that is not a haiku about the ocean	**433**	**-0.959**
Complement	Generate a to-do list for moving houses	450	-1.23
Complement	Generate anything that is not a to-do list for moving houses	**675**	**-1.01**
DS (Logic)	From “Either the pen is in my bag or it is on my desk” together with ”The pen isn’t on my desk”, can we infer “The pen is in my bag”?	704	-1.41
DS (Logic)	From “Either the umbrella is in the car or it tucked away in the closet” together with “The umbrella isn’t tucked away in the closet”, can we infer “The umbrella is in the car”?	532	-1.39
uDSmu (Logic)	Either the cat is napping on the couch or it must be playing in the bedroom. Also, it’s not the case that the cat must be playing in the bedroom. Can we infer that the cat is napping on the couch?	**1606**	**-1.21**
uDSmu (Logic)	Either the jacket is draped over the chair or it must be hanging in the closet. Also, it’s not the case that the jacket must be hanging in the closet. Can we infer that the jacket is draped over the chair?	**1262**	**-1.24**

thought, we categorized all inferences from Holliday et al. (2024) into two classes: Easy and Hard. For instance, we classified simple inference patterns, such as Modus Ponens and Modus Tollens, that students are introduced to in an introductory logic class, as Easy. Inferences that involve operations such as modal distribution over booleans were classified as Hard. Our classification was also guided by the accuracies reported in Holliday et al. (2024); we took it that models have difficulty solving harder tasks and thereby achieve lower accuracies on them.

Below we show the average reasoning token length and EigenScore for different tasks based on different difficulty levels, where we group different tasks into easy and hard. Tab 14 shows that the harder reasoning tasks have a longer token length and higher EigenScore.

Table 14: Comparison of Token Length and EigenScore for easy and hard modal and conditional reasoning tasks from the dataset used in Holliday et al. (2024)

Difficulty Level	Token Len	EigenScore
Easy	664.81 ± 15.39	-1.19 ± 0.01
Hard	**1254.93 ± 59.40**	**-0.96 ± 0.03**

Table 15: Token Length and EigenScore per task type.

Task Difficulty	Task Type	Token Len	EigenScore
Easy	AS	933.33 ± 118.50	-1.10 ± 0.06
	CONV	600.05 ± 37.78	-1.19 ± 0.03
	CT	795.42 ± 78.99	-1.19 ± 0.03
	DA	621.25 ± 29.49	-1.21 ± 0.03
	DS	549.66 ± 20.61	-1.16 ± 0.03
	INV	704.00 ± 40.22	-1.24 ± 0.03
	MP	441.77 ± 13.86	-1.09 ± 0.03
	MT	521.69 ± 21.72	-1.17 ± 0.03
	MiN	728.98 ± 27.71	-1.22 ± 0.02
	NMu	689.34 ± 41.07	-1.24 ± 0.03
Hard	CMP	**2643.60 ± 488.00**	**-0.40 ± 0.05**
	DSmi	1676.39 ± 108.32	-0.71 ± 0.05
	DSmu	709.02 ± 44.13	-1.25 ± 0.02
	MTmi	1869.09 ± 159.29	-0.50 ± 0.04
	MTmu	720.24 ± 56.45	-1.24 ± 0.02
	MuAg	891.98 ± 121.42	-1.25 ± 0.05
	MuDistOr	1170.68 ± 153.26	-1.12 ± 0.07
	NSFC	1018.05 ± 145.24	-1.21 ± 0.07
	WSFC	934.25 ± 190.73	-1.25 ± 0.05