Agreement in Representation Space for Open-Ended Self-Consistency

Anonymous ACL submission

Abstract

Self-consistency improves LLM reasoning by sampling multiple outputs and selecting the most consistent answer, but existing formulations largely rely on exact matching and therefore remain limited to tasks with categorical outputs. In this work, we study self-consistency in open-ended generation tasks such as code synthesis and text summarization. We hypothesize that consistency can be understood as a geometric property of the generation space, where semantically compatible generations concentrate in similar regions of representation space. To study this hypothesis, we introduce Embedding-Based Agreement (EBA), a simple training-free operationalization that estimates agreement by clustering sampled generations in embedding space. Through experiments on mathematical reasoning, code generation, and summarization, we show that agreement in representation space provides a robust and scalable signal of self-consistency for open-ended tasks. In particular, EBA consistently outperforms random selection and exhibits more stable scaling behavior than recent selection approaches based on LLM evaluation or uncertainty estimation. We further show that these agreement signals remain stable across model families and embedding spaces, even with native hidden representations. Finally, our analysis shows that the geometric location occupied by sampled generations is strongly correlated with generation quality: generations concentrated near central regions of representation space tend to correspond to more reliable outputs, whereas peripheral generations are substantially less accurate. Overall, our findings support viewing self-consistency as a property of the geometric organization of sampled generations rather than exact symbolic overlap.

1 Introduction

Self-consistency has emerged as one of the most effective inference-time strategies for improving the reasoning capabilities of large language models (LLMs). Classical self-consistency methods sample multiple reasoning paths and select the most frequent final answer through majority voting (Wang et al., 2022). This approach has proven highly effective in tasks with categorical outputs constrained to a limited set of possibilities, such as mathematical reasoning and multiple-choice question answering, where agreement can be naturally defined through exact answer matching.

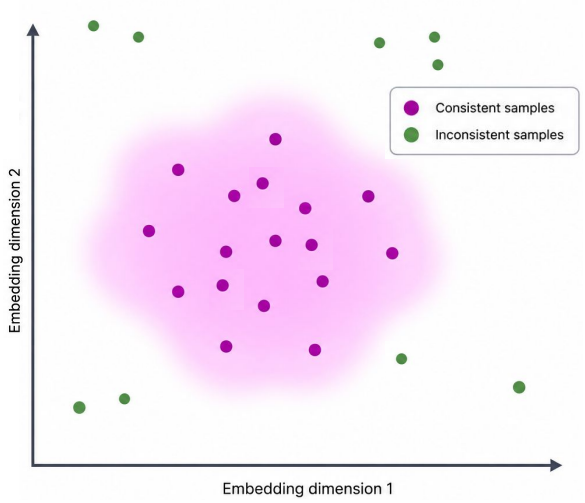

Figure 1: Illustration of our central hypothesis: self-consistency can be understood as a geometric property of the generation space. Semantically consistent generations (purple) are expected to accumulate in dense regions of representation space, forming dominant agreement structures. In contrast, inconsistent generations (green) tend to appear more isolated and sparsely distributed. Under this perspective, agreement can be estimated by identifying dominant concentration regions rather than relying on exact symbolic overlap.

However, extending self-consistency beyond closed-form tasks remains challenging. In open-ended generation settings such as code generation or summarization, responses that are semantically equivalent may differ substantially at the lexical or structural level, making exact matching ineffective. Recent approaches address this limitation

by relying on LLM-based evaluation (Chen et al., 2023) or uncertainty-based scoring signals (Kang et al., 2026). While effective in some settings, these methods raise broader questions about what agreement actually means in open-ended generation and how consistency signals should be estimated when outputs do not admit canonical symbolic forms.

In this work, we study self-consistency from the perspective of representation-space geometry. We hypothesize that semantically consistent generations occupy similar regions in embedding space and that agreement can therefore be understood as a geometric property of the distribution of sampled generations (see Figure 1). Building on this view, we introduce Embedding-Based Agreement (EBA), a simple training-free framework that estimates self-consistency by clustering sampled generations in representation space and selecting representative generations from dominant concentration regions.

Rather than presenting EBA primarily as a new decoding strategy, our goal is to investigate whether agreement in embedding space provides a reliable and scalable signal of self-consistency across both verifiable and open-ended tasks. Through experiments on mathematical reasoning (MATH500), code generation (HumanEval), and summarization (CNN/DM), we show that geometric agreement consistently improves generation selection quality over random sampling and existing selection approaches, and remains robust across model families, model scales, embedding spaces, and native hidden representations.

Our analysis further reveals several properties of agreement in generation space. First, agreement signals strengthen as the number of sampled generations increases, suggesting that consistency emerges from the collective organization of the sampled distribution. Second, embedding-based agreement exhibits more stable scaling behavior than recent approaches based on LLM evaluation or uncertainty estimation. Third, we find that the geometric location occupied by sampled generations strongly correlates with generation quality: generations concentrated near central regions of representation space consistently produce more reliable outputs, whereas peripheral generations are substantially less accurate. Finally, we show that embedding-based agreement closely connects to classical majority-voting self-consistency in closed-form reasoning tasks, suggesting a unified geometric interpretation of self-consistency across categorical and open-ended generation settings.

Overall, our findings suggest that self-consistency can be understood as a property of the geometric organization of sampled generations in representation space, opening a broader perspective on agreement estimation beyond exact symbolic matching.

2 Related Work

Self-consistency (SC) (Wang et al., 2022) improves reasoning in LLMs by sampling multiple reasoning paths and selecting the most frequent final answer through majority voting. This approach has proven highly effective for tasks with categorical and verifiable outputs, such as mathematical reasoning and multiple-choice question answering, where agreement can be determined through exact answer matching.

However, many generation tasks do not naturally admit such limited answer spaces. In open-ended settings such as summarization, code generation, or free-form reasoning, semantically equivalent outputs may differ substantially at the lexical or structural level, making exact-match voting insufficient for estimating consistency.

Several approaches have recently attempted to extend self-consistency beyond closed-form reasoning tasks. Universal Self-Consistency (USC) (Chen et al., 2023) prompts an LLM to identify the most consistent response among multiple candidates, replacing symbolic voting with model-based evaluation. More recently, lightweight evaluator-free approaches have been proposed. Self-Certainty (SCe) (Kang et al., 2026) estimates response quality from token-level uncertainty signals, while similarity-based voting methods (Jiang et al., 2025) leverage semantic similarity in embedding space through pairwise voting.

Our work is closely related to this latter direction, but differs in both motivation and formulation. Prior work primarily treats embeddings as a mechanism for soft answer matching or response ranking. In contrast, we interpret self-consistency itself as a geometric property of the generation space. Specifically, we hypothesize that semantically coherent generations form concentration regions in representation space, enabling agreement to emerge from the collective geometric organization of sampled outputs rather than from pairwise similarity aggregation alone.

More broadly, our work connects to a long line of research showing that neural representa-

tion spaces capture meaningful semantic and structural information, enabling applications such as semantic retrieval (Reimers and Gurevych, 2019; Karpukhin et al., 2020), clustering (Aggarwal and Zhai, 2012), and paraphrase identification (Lan and Xu, 2018). We build on these observations to study agreement in open-ended generation as a property of embedding-space structure rather than discrete symbolic overlap.

3 Self-Consistency as Agreement in Output Space

3.1 Embedding-space agreement

Classical self-consistency methods estimate agreement through exact matching or majority voting over categorical outputs. However, in open-ended generation tasks, semantically equivalent responses may differ substantially at the lexical or structural level, making symbolic agreement insufficient for capturing consistency.

In this work, we hypothesize that self-consistency in open-ended generation can instead be understood as a geometric property of the generation space. More specifically, independently sampled generations expressing similar semantic content are expected to occupy nearby regions in a shared representation space, as illustrated in Figure 1. Under this view, consistency emerges not from exact symbolic overlap, but from the formation of dense semantic regions induced collectively by multiple generations.

Formally, let x denote an input problem and let $\mathcal{G}(x) = \{g_1, g_2, \ldots, g_n\}$ be a set of generations sampled independently from a language model conditioned on x. Let $\phi(g_i) \in \mathbb{R}^d$ denote a vector representation of generation g_i, obtained through an embedding function ϕ. The sampled generations therefore induce a geometric structure in representation space: $\Phi(x) = \{\phi(g_1), \phi(g_2), \ldots, \phi(g_n)\}$.

Our central hypothesis is that the structure of $\Phi(x)$ contains information about the latent consistency of the sampled outputs. Semantically coherent generations are expected to form concentrated regions or modes in representation space, while inconsistent or low-quality generations should appear more isolated or fragmented. Under this perspective, self-consistency becomes a problem of identifying dominant semantic regions within the geometry induced by the sampled generations.

This interpretation generalizes the intuition underlying classical self-consistency. In closed-form reasoning tasks, majority voting operates over discrete symbolic answers, implicitly assuming that semantically equivalent outputs collapse to the same representation. Embedding-space agreement extends this idea to continuous and structurally diverse outputs, where semantic similarity must be inferred geometrically rather than symbolically.

3.2 Operationalizing Agreement in Representation Space

Building on this perspective, we operationalize self-consistency in open-ended generation through *Embedding-Based Agreement* (EBA). Given an input prompt x, we first sample N candidate generations $\{y_1, \ldots, y_N\}$ from a language model and map each generation into a continuous representation space through an embedding function $f(\cdot)$, producing embeddings $\{e_1, \ldots, e_N\}$.

Under our hypothesis, agreement among generations is reflected in the geometric organization of this representation space. Rather than requiring exact lexical overlap, semantically compatible generations are expected to accumulate in nearby regions, forming concentration structures that reflect dominant modes of agreement.

To estimate this structure, we cluster the embeddings associated with the sampled outputs. Clusters containing larger numbers of generations are interpreted as regions exhibiting stronger agreement among the sampled responses. In this sense, clustering plays a role analogous to majority voting in classical self-consistency: instead of counting identical symbolic answers, agreement is estimated through the accumulation of semantically similar generations in representation space. We then identify the dominant cluster and select as final prediction the sample located closest to its centroid, favoring generations that are most representative of the dominant agreement region.

In our implementation, clustering is performed using agglomerative clustering (Müllner, 2011), with the number of clusters determined automatically through silhouette analysis (Rousseeuw, 1987). Unless otherwise specified, embeddings are obtained using pretrained embedding models independently from the generation model itself (implementation details are provided in Appendix A).

4 Experimental Setup

Datasets. We evaluate Embedding-Based Agreement (EBA) on three generation tasks covering

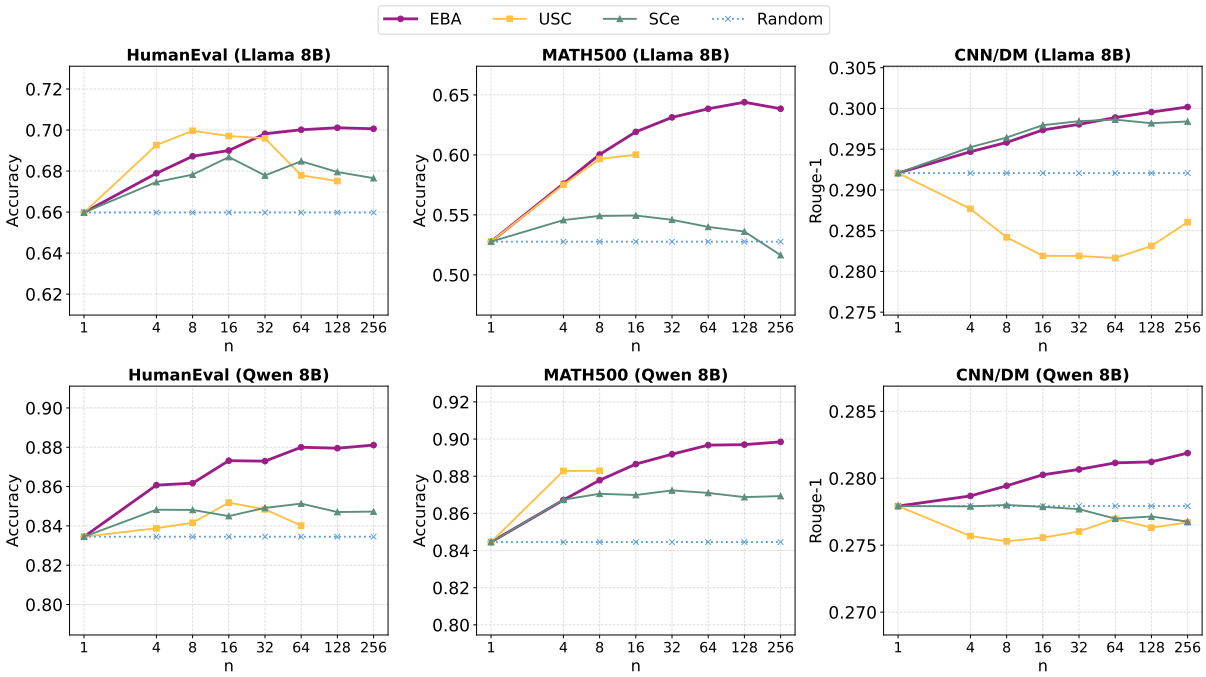

Figure 2: Performance of EBA, random selection, Universal Self-Consistency (USC), and self-certainty (SCe) across increasing numbers of sampled generations on HumanEval, MATH500 and CNN/DM using Llama 3-8B (top) and Qwen 3-8B (bottom). EBA consistently improves with increasing numbers of samples and tends to plateau at higher values, whereas USC and SCe generally plateau earlier or even degrade. This suggests more stable scaling behaviour for embedding-based agreement.

both verifiable and open-ended settings: HumanEval (Chen et al., 2021) for code generation, CNN/DailyMail (CNN/DM) (Hermann et al., 2015; See et al., 2017) for summarization, and MATH500 (Hendrycks et al., 2021; Lightman et al., 2023) for mathematical reasoning. We report accuracy for HumanEval and MATH500, and Rouge-1 (Lin, 2004) for CNN/DM. Unless otherwise specified, agreement on MATH500 is computed directly over full generations.

Models. Our main experiments use Llama 3.1-8B (Grattafiori et al., 2024) and Qwen 3-8B (Team, 2025) as generation models, together with Qwen-Embedding-8B (Zhang et al., 2025) as embedding function ϕ for representation extraction.

Reference methods. We compare EBA against Universal Self-Consistency (USC) (Chen et al., 2023), Self- Certainty (SCe) (Kang et al., 2026), and random selection.

Implementation details. Agreement regions are estimated through agglomerative clustering with cosine distance, where the number of clusters is selected automatically using the Silhouette score. We use a sampling temperature of $T = 0.7$ with $top_p = 1.0$ and report results for n between 1 and 256.[1] To reduce noise in the results, we use 50 bootstrapping iterations for every reported result, drawing sets of generations from a fixed pool of 512 sampled generations for each input. For USC, we use as the selector the same LLM that was used to generate the samples. Additional implementation details, prompts, and hardware specifications are provided in Appendix A and Appendix B.

5 Empirical Evidence of Embedding-Space Agreement

This section evaluates agreement in representation space as a signal of self-consistency in open-ended generation. We study the effectiveness, scaling behavior, and robustness of Embedding-Based Agreement (EBA) across models and representation spaces (full results in Appendix D).

5.1 Agreement in embedding space improves selection quality

Figure 2 shows the performance of Embedding-Based Agreement (EBA) across code generation (HumanEval), mathematical reasoning

[1]For USC, the evaluator model can run out of context for large n and tasks with long generations, in which case we report results up to the context limit.

(MATH500), and summarization (CNN/DM), together with several reference selection methods. Across both Llama and Qwen models, EBA consistently outperforms random selection, indicating that agreement in representation space provides a substantially stronger signal than selecting arbitrary generations from the sampled set.

The improvements are particularly pronounced in code generation and mathematical reasoning, where EBA improves over random selection by up to 4 and 17 points respectively, depending on the model. The effect is smaller but still consistent for summarization (approximately $+1$ Rouge-1 point), where generations tend to occupy more concentrated regions of the output space and evaluation metrics may under-reward semantically equivalent paraphrastic outputs. Overall, these results suggest that agreement can emerge from the geometric organization of sampled generations even when responses differ lexically or structurally, supporting the view of self-consistency as a property of the generation distribution rather than exact symbolic matching.

5.2 Agreement strengthens with increased sampling

As shown in Figure 2, EBA exhibits consistent and near-monotonic improvements as the number of sampled generations increases across datasets and model families. The effect is particularly pronounced in HumanEval and MATH500, where performance tends to improve over a broad range of sampling budgets before gradually leveling off at larger values of n. In contrast, summarization exhibits smaller but still positive gains as the number of generations increases.

These results suggest that agreement signals become increasingly informative as the generation distribution is explored more thoroughly. As additional generations are sampled, dominant concentration regions in representation space may become more clearly defined, making agreement estimates more reliable. The observed scaling behavior supports the view of self-consistency as a geometric property emerging from the collective organization of sampled generations rather than exact lexical overlap between outputs.

5.3 Geometric agreement provides a robust signal of self-consistency

Figure 2 additionally compares EBA against two recent approaches for extending self-consistency beyond exact matching: Universal Self-Consistency (USC) and Self-Certainty (SCe). While both methods generally improve over random selection, EBA exhibits more stable behavior across tasks and sampling regimes.

USC frequently degrades as the number of candidate generations increases, for example it loses around two points on HumanEval for the Llama model. This behavior suggests increasing difficulty in reliably identifying agreement through direct LLM-based comparison over larger candidate sets. Moreover, USC is inherently constrained by the maximum context length of the underlying model, limiting its scalability to large numbers of generations.

SCe demonstrates substantially stronger scaling behavior than USC and remains competitive across several settings. However, its performance is less stable as sampling increases; for example, on MATH500 with Llama 8B, performance peaks early before dropping by roughly four points between $n = 16$ and $n = 256$. In contrast, EBA consistently matches or outperforms SCe on MATH500 while remaining competitive on open-ended tasks such as HumanEval and CNN/DM. Overall, these results suggest that geometric agreement provides a robust and scalable signal of self-consistency across diverse generation settings.

5.4 Agreement is stable across model scales and embedding spaces

Figure 3 analyzes the robustness of EBA across different language model scales and embedding spaces. We evaluate both larger generation models (Llama 70B and Qwen 32B) and alternative embedding models (Qwen and Gemma embeddings (Vera et al., 2025)). Across all configurations, the same qualitative behavior observed in Figure 2 remains consistent: agreement-based selection continues to outperform random selection and benefits from increased sampling.

While absolute performance and improvement magnitudes vary across models and embedding spaces, several qualitative patterns remain consistent. In particular, EBA exhibits strong and largely monotonic gains on MATH500 and generally positive trends on CNN/DM, whereas HumanEval displays more variable behavior, particularly for larger models. Larger models also tend to exhibit smaller improvements overall, which may be explained by their stronger baseline performance and consequently smaller potential gains from selection.

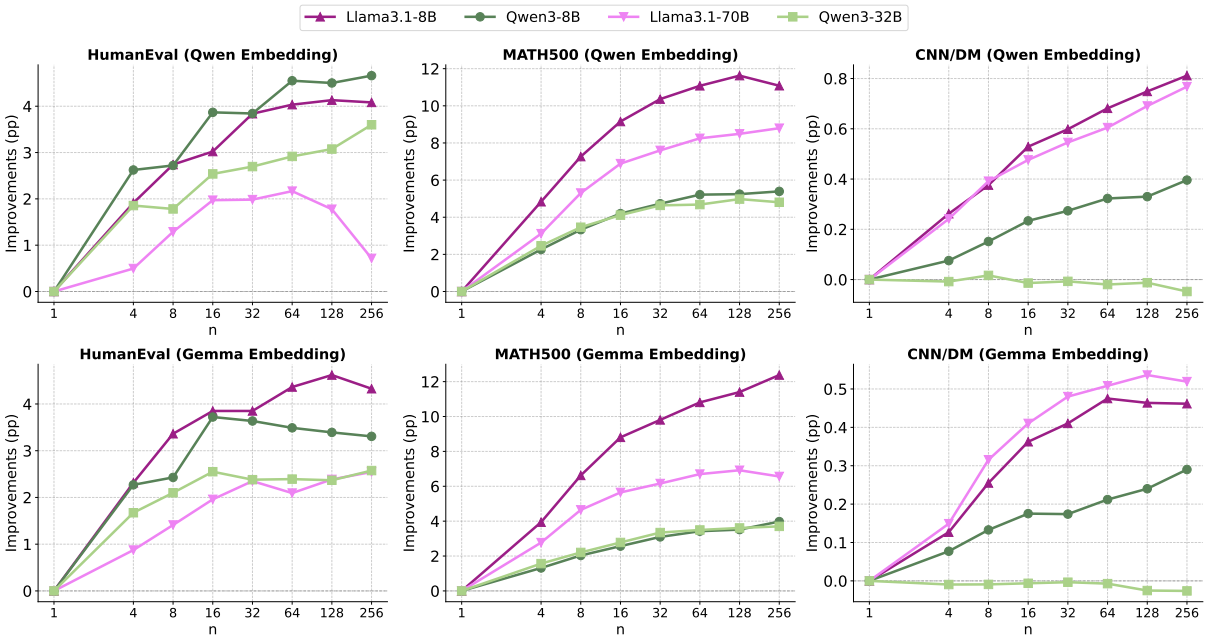

Figure 3: Improvement of embedding-based agreement (EBA) as the number of sampled generations increases across code generation (HumanEval), summarization (CNN/DM), and mathematical reasoning (MATH500). The y-axis reports gain in **percentage points (pp)** over random performance. Results are shown for multiple language models and embedding representations, including Qwen and Gemma embeddings. Across model scales and embedding spaces, EBA exhibits consistent scaling behavior, indicating that agreement in embedding space reflects a robust and model-agnostic property of the generation distribution.

These results indicate that the proposed framework captures a general property of the geometry of sampled generations rather than an artifact of a particular representation space.

5.5 Agreement emerges in native hidden representations

Figure 4 evaluates EBA using representations extracted directly from the generation model itself, replacing external embedding models with the final sampled token's hidden representation. Across all datasets and model families, EBA continues to exhibit the same qualitative behavior observed with pretrained embedding models: agreement-based selection consistently improves with increased sampling and remains substantially stronger than random selection.

Although performance is typically slightly lower than with specialized embedding models, with differences often within ±1–2 points across datasets, several qualitative patterns remain similar. Improvements are generally stronger and more stable for MATH500 and CNN/DM, while HumanEval exhibits more variable behavior, particularly at larger sampling budgets. These results suggest that useful agreement signals may already be present in native hidden representations and do not rely exclusively on external semantic embedding spaces. More broadly, they are consistent with the view that aspects of self-consistency may be reflected in the internal organization of sampled generations in representation space.

6 Analysis of Self-Consistency in Open-Ended Generation

Results from Section 5 show that agreement in representation space is a strong signal of self-consistency across tasks. We further study the underlying geometry and its relationship to self-consistency, as well as the effects of local and global structure on generation quality (full results in Appendix D).

6.1 Connection to closed-form self-consistency

Classical self-consistency operates in discrete answer spaces, where sampled generations are reduced to final answers and aggregated through majority voting. In contrast, EBA targets open-ended generation settings in which outputs cannot be naturally mapped to a small set of categorical responses. A natural question, therefore, is whether embedding-based agreement remains connected to

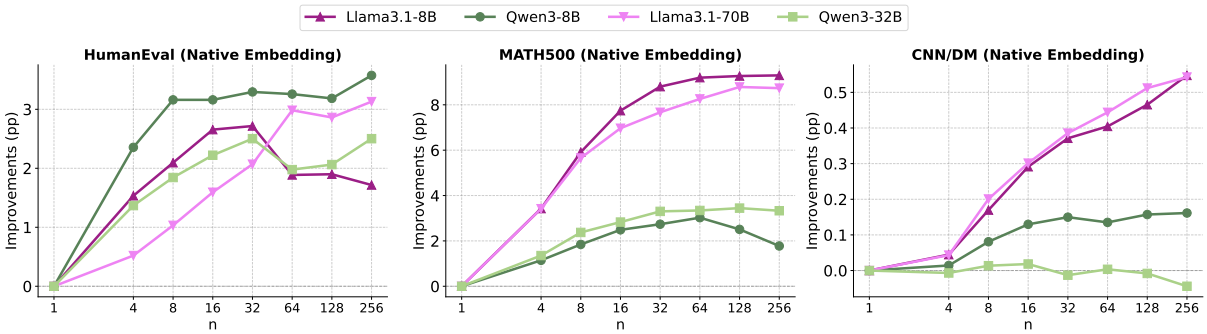

Figure 4: Improvement of embedding-based agreement (EBA) using native hidden representations as the number of sampled generations increases across all datasets. The y-axis reports gain in **percentage points (pp)** over random performance. Results are shown for multiple language models.

the original intuition underlying closed-form self-consistency.

To study this connection, we evaluate EBA on MATH500 under two representation settings: (i) using the complete generation, including both reasoning trace and final answer, and (ii) using only the extracted final answer. We compare both variants against standard majority-voting self-consistency.

Figure 5 shows that embedding-based agreement exhibits scaling behavior remarkably similar to classical self-consistency. Across both Llama 3-8B and Qwen 3-8B, performance improves consistently as additional generations are sampled, closely following the monotonic trend of majority voting. This similarity is particularly strong when agreement is computed over isolated final answers, where EBA approaches the performance of exact-match self-consistency across all sample sizes. Similar trends are observed for larger models (see Appendix C).

The comparison between full-generation and answer-only agreement further highlights an important property of open-ended generation spaces. While both variants exhibit similar scaling dynamics, agreement over complete generations consistently underperforms answer-level agreement, suggesting that embedding representations remain sensitive to differences in reasoning structure, lexical choice, and chain-of-thought organization even when generations lead to the same final answer.

Overall, these findings suggest that EBA can be viewed as a continuous extension of classical self-consistency. When outputs become effectively discrete, agreement in representation space behaves similarly to majority voting, while naturally extending the same idea to more diverse generation settings.

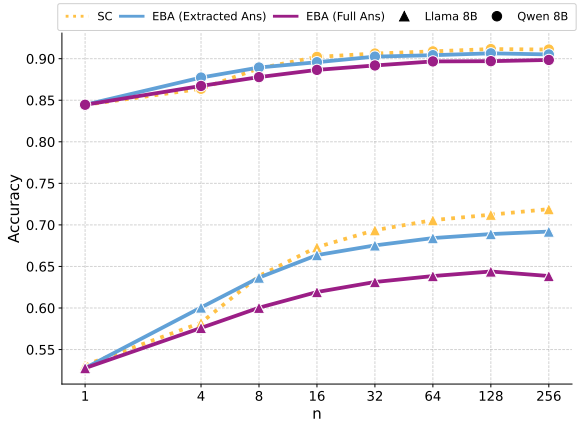

Figure 5: Comparison of self-consistency (SC) and embedding-based aggregation method using the complete generation (EBA (Full Ans)) and using only the extracted final answer (EBA (Extracted Ans)) on MATH500. Results are shown for Llama 8B and Qwen 8B models.

6.2 Generation quality correlates with geometric structure

While EBA operationalizes agreement through clustering, this discretization is not the only possible way to estimate agreement in continuous representation spaces. An alternative view is to treat agreement as a global geometric property of the sampled generations without explicitly partitioning the space into clusters. Under this perspective, generations located near the global centroid may correspond to semantically central or highly representative outputs, whereas generations located farther away may reflect more isolated or inconsistent behaviors. Figure 6 compares EBA against two centroid-based selection strategies on MATH500: selecting the generation closest to the global centroid and selecting the generation farthest from it.

The results reveal a striking geometric pattern.

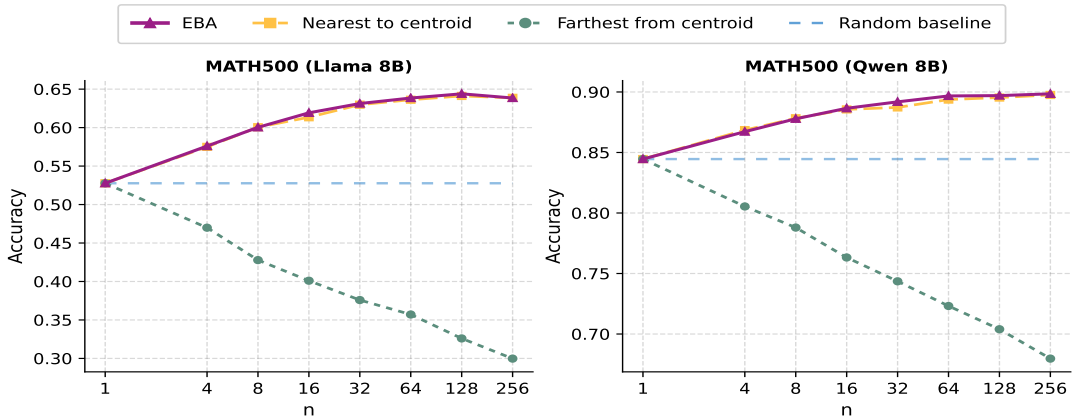

Figure 6: Comparison of generation selection strategies on MATH500 for Llama 8B and Qwen 8B as the number of sampled generations (n) increases. Selecting generations closest to the global centroid achieves performance nearly similar to EBA, whereas selecting generations farthest from the centroid causes substantial performance degradation. The dashed line denotes the random baseline corresponding to performance at $n = 1$.

Across both Llama and Qwen models, selecting generations closest to the global centroid achieves performance remarkably similar to EBA across all sampling regimes. In contrast, selecting generations farthest from the centroid produces dramatic degradations in accuracy, reaching drops of nearly 20 points for large sample sizes. These results indicate that the geometric location occupied by a generation in representation space is strongly correlated with its quality: generations concentrated near central regions tend to correspond to higher-quality outputs, whereas peripheral generations are substantially less reliable.

At the same time, the near-equivalence between EBA and global-centroid selection suggests that the explicit discretization introduced by clustering contributes only marginally under the current experimental setting. Several factors may explain this behavior. Prior work (Liang et al., 2022) has shown that embedding representations often populate narrow high-dimensional conic regions, which may reduce the separability of distinct semantic modes. Additionally, the relatively simple clustering strategy used in this work may fail to capture finer-grained structure in the generation distribution. More broadly, open-ended generation spaces may not exhibit sharply separated semantic regions, but rather smoother concentration gradients organized around dominant semantic directions.

7 Conclusion

In this work, we studied self-consistency beyond categorical answer spaces. We hypothesized that semantically compatible generations concentrate in nearby regions of embedding space, and introduced Embedding-Based Agreement (EBA), a simple training-free framework that estimates agreement by clustering sampled generations and selecting representative outputs from dominant concentration regions.

Across mathematical reasoning, code generation, and summarization, we showed that agreement in embedding space provides a reliable and scalable selection signal. EBA consistently improves over random selection and exhibits more stable scaling behavior than recent approaches based on LLM evaluation or uncertainty estimation. Moreover, the observed agreement patterns remain robust across model families, model scales, embedding spaces, and native hidden representations.

We further connected EBA to classical self-consistency on closed-form reasoning tasks, showing that EBA closely follows the scaling behavior of exact-match majority voting while naturally extending to open-ended generation. Our analysis also reveals that the geometric location occupied by sampled generations is strongly correlated with generation quality, with central regions in representation space consistently associated with more reliable outputs than peripheral regions.

Overall, our findings support viewing self-consistency as a geometric property of sampled generations rather than a phenomenon tied exclusively to exact symbolic overlap. This perspective provides a unified interpretation of agreement across discrete and open-ended tasks, and opens new directions for inference-time generation selection based on representation-space structure.

Limitations

While our results suggest that embedding-based agreement captures a meaningful signal of self-consistency in open-ended generation, several limitations remain.

Representation geometry and semantic alignment. Our approach assumes that semantically compatible generations occupy nearby regions in representation space. Although this assumption holds sufficiently well to produce consistent improvements across tasks, the results of Section 6 suggest that current embedding spaces exhibit relatively low-separation geometric structure, with generations often concentrating in a small number of broad regions. This may reflect known properties of neural embedding spaces, such as anisotropy and concentration in narrow high-dimensional cones. As a result, embedding proximity does not always perfectly correspond to semantic equivalence, potentially limiting the precision of agreement estimation.

Limitations of clustering-based agreement estimation. EBA operationalizes agreement through agglomerative clustering and heuristic model-selection criteria based on silhouette analysis. While this provides a simple and fully training-free mechanism for estimating dominant concentration regions, the relatively small gap observed between EBA and global centroid selection suggests that current clustering strategies may not fully capture the local structure of the generation space. More expressive geometric modeling or representation-learning approaches may therefore yield improved estimates of semantic agreement.

Computational considerations. Although EBA avoids additional LLM evaluation calls and remains substantially cheaper than evaluator-based approaches such as USC, it still requires generating multiple samples and computing their representations. For large sampling budgets or long generations, embedding extraction and storage may introduce additional computational overhead. At the same time, our experiments suggest that simpler alternatives based on native embeddings may provide attractive practical trade-offs.

Scope of evaluation. Our experiments focus on summarization, code generation, and mathematical reasoning as representative open-ended generation settings. While these tasks cover both verifiable and non-verifiable outputs, additional evaluation on domains such as dialogue, long-form generation, or agentic reasoning would be necessary to fully assess the generality of the proposed perspective.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. volume 28.

Chunyang Jiang, Yonggang Zhang, Yiyang Cai, Chi-Min Chan, Yulong Liu, Mingming Chen, Wei Xue, and Yike Guo. 2025. Semantic voting: A self-evaluation-free approach for efficient llm self-improvement on unverifiable open-ended tasks. *arXiv preprint arXiv:2509.23067*.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2026. Scalable best-of-n selection for large language models via self-certainty. *Advances in neural information processing systems*, 38:19720–19745.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th international*

conference on computational linguistics, pages 3890–3902.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Implementation details

Clustering. We use the same temperature and sampling strategy for all model sizes. Specifically, we generate 512 samples per input using temperature $T = 0.7$ and $top_p = 1.0$ sampling in a zero-shot setting. We then create subsets of size 4, 8, 16, 32, 64, 128, and 256. We do not report experiments for $n = 512$ because we perform 50 bootstrap resamples, which would not be feasible when using all 512 samples. We use the vLLM framework for generation, whereas native embeddings are extracted using Hugging Face, as it provides access to hidden-dimensional representations.

For all datasets except the HumanEval dataset, clustering is applied directly to the generated responses, which are also the embedded representations. For HumanEval, clustering is instead performed on the extracted final answers, which are likewise used as the embedded representations. Regarding the embedding model, we use the generations produced by each model independently as input. Hierarchical clustering (Müllner, 2011)[2] is performed by constructing the complete clustering tree and selecting the partition that maximizes the Silhouette score. We use average linkage and cosine distance, as embeddings are computed using cosine similarity. All the code is available at the projects' repository[3].

Reference methods For USC, we use the temperature values recommended for each model to obtain the best performance, i.e. for Llama models ($T = 0.7$, $top_p = 0.8$, $top_k = 20$) and for Qwen models ($T = 0.6$, $top_p = 0.9$). We use the prompts from the original paper (Chen et al., 2023) with minor modifications depending on the dataset (see Appendix B). When USC fails to choose one answer we randomly select one.

For SCe, we use the authors' released implementation[4] while adapting it to our experimental setup, i.e., retaining their self-certainty scoring and selection procedure while applying it to our generated samples. We also use a bootstrapping of 50 for these experiments.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[3] We will made public upon acceptance.

[4] https://github.com/backprop07/Self-Certainty

Datasets. All datasets are publicly available[5][6][7]. Due to computational constraints, we use a subset of 1,000 samples for CNN/DM from the 3.0.0 version. Specifically, we select the first 1,000 samples of the dataset.

For evaluation in the case of MATH dataset, we adopted an LLM-based evaluation, in line with a widely-used library[8]. Since both predictions and gold answers often contain LaTeX expressions and certain expressions can be written in different but mathematically equivalent ways (e.g., \frac{2}{3} and 2/3), string matching is not always a reliable metric. To address this, we used the Qwen 8B model for evaluation. This choice was supported by an empirical study in which we measured the agreement between Qwen 8B and human-annotated labels on 200 instances, achieving a 95.5% match. For HumanEval dataset, we first, extract the final answer with a combination of filters and regular expressions and then, we evaluate the function with the official evaluation harness[9] which runs the function against a set of unit tests.

Computational Infrastructure. All experiments were conducted using NVIDIA A100 GPUs. We use 1 GPU for small model inference (Llama 8B and Qwen 8B), 2 GPUS for Qwen 32B and 4 GPUs for Llama70B model inference. For both embeddings models we use 1 GPU and clustering was made in CPU.

Licences. To the best of our knowledge, all the data sources we use are available for non-commercial use.

- **MATH** (Hendrycks et al., 2021; Lightman et al., 2023): It is under a MIT License[10].
- **CNN/DailyMail** (Hermann et al., 2015; See et al., 2017): It is under Apache-2.0 license[11].
- **HumanEval** (Chen et al., 2021): It is under a MIT License[12].

Regarding the models we used:

- **Llama models** (Grattafiori et al., 2024): They are licensed under the Llama 3.1 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved[13].
- **Qwen models** (Team, 2025; Zhang et al., 2025): They are publicly accessible under Apache 2.0[14][15].
- **Gemma Embeddings model** (Vera et al., 2025): It is under Gemma License[16].

B Prompts

We use the same prompts for all models, while using different prompts for dataset. For USC We use the prompts from the original paper (Chen et al., 2023) with minor modifications depending on the dataset.

C Analysis of Self-Consistency in Open-Ended Generation for Larger Models

We replicate the MATH500 evaluation using larger backbone models to test the scalability of embedding-based agreement (EBA). Results show (see Figure 7) the same qualitative behavior: EBA improves monotonically with additional samples and closely follows classical self-consistency trends across all model sizes.

As in smaller models, agreement computed over final answers is more closely aligned with majority voting than full-generation embeddings, although both exhibit consistent scaling patterns. These findings confirm that the relationship between embedding-space agreement and self-consistency remains stable as model capacity increases.

D Full results

This appendix includes full tabular versions of the figures. Tables 1, 2, and 3 present the results of Figure 2. Tables 4, 5, and 6 present the results of Figures 3 and 4. Table 7 and Table 8 present the results for Figure 5 and Figure 7 respectively. Finally, Table 9 presents the result of the Figure 6.

[5] https://huggingface.co/datasets/HuggingFaceH4/MATH-500
[6] https://huggingface.co/datasets/openai/openai_humaneval
[7] https://huggingface.co/datasets/abisee/cnn_dailymail
[8] https://github.com/openai/simple-evals
[9] https://github.com/openai/human-eval
[10] https://github.com/openai/prm800k/tree/main?tab=MIT-1-ov-file
[11] https://huggingface.co/datasets/abisee/cnn_dailymail#licensing-information
[12] https://github.com/openai/human-eval?tab=MIT-1-ov-file
[13] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct/blob/main/LICENSE
[14] https://github.com/QwenLM/Qwen3#license-agreement
[15] https://huggingface.co/Qwen/Qwen3-Embedding-8B/blob/main/LICENSE
[16] https://ai.google.dev/gemma/terms

Prompt for CNN/DailyMail

Summarize the following text in less than 4 sentences. Generate a concise and coherent summary that captures the main points of the text:
<INPUT>

Summary:

Prompt for USC for CNN/DailyMail

I have the following text: <INPUT>
I have generated the following summaries of the text: <RESPONSES>

Evaluate the responses.
Select the most consistent response based on majority consensus.
Start your answer with "The most consistent response is Response X" (without quotes).

Prompt for HumanEval

<INPUT>

Prompt for USC for HumanEval

I have generated the following code-completion outputs for this function:
<INPUT>
<RESPONSES>

Evaluate the responses.
Select the most consistent response based on majority consensus.
Start your answer with "The most consistent response is Response X" (without quotes).

Prompt for MATH500

Solve the following math problem step by step. The last line of your response should be of the form Answer: $ANSWER (without quotes) where $ANSWER is the answer to the problem.

<INPUT>

Remember to put your answer on its own line after "Answer:", and you do not need to use a \\boxed command.

Prompt for USC for MATH500

I have generated the following responses to the mathematical problem:
<INPUT>
<RESPONSES>

Evaluate the responses.
Select the most consistent response based on majority consensus.
Start your answer with "The most consistent response is Response X" (without quotes).

Model	Method	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	EBA	65.98 ± 2.38	67.89 ± 2.23	68.72 ± 2.12	69.00 ± 1.97	69.82 ± 1.54	70.01 ± 1.96	70.11 ± 1.52	70.06 ± 1.58
	SCe	65.98 ± 2.38	67.46 ± 0.78	67.83 ± 0.47	68.68 ± 0.57	67.79 ± 0.43	68.48 ± 0.47	67.95 ± 0.36	67.65 ± 0.51
	USC	65.98 ± 2.38	69.27± 2.05	69.96±1.99	69.71±2.02	69.60 ±1.79	67.79± 2.17	67.51 ±2.10	–
	Random	65.98	65.98	65.98	65.98	65.98	65.98	65.98	65.98
Qwen 8B	EBA	83.45 ± 2.10	86.07 ± 1.85	86.17 ± 1.46	87.32 ± 1.17	87.29 ± 1.26	88.00 ± 1.38	87.95 ± 1.21	88.11 ± 1.31
	SCe	83.45 ± 2.10	84.82 ± 0.37	84.81 ± 0.32	84.50 ± 0.27	84.91 ± 0.22	85.13 ± 0.40	84.70 ± 0.31	84.73 ± 0.18
	USC	83.45 ± 2.10	83.88 ± 1.87	84.16 ± 1.75	85.18 ± 1.59	84.84 ± 1.90	84.01 ± 2.13	–	–
	Random	83.45	83.45	83.45	83.45	83.45	83.45	83.45	83.45

Table 1: HumanEval results across sampling budgets (n), comparing Embedding-Based Agreement (EBA), Universal Self-Consistency (USC), Self-Certainty (SCe), and random selection for Llama 8B and Qwen 8B. Results show accuracy scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Method	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	EBA	52.77 ± 1.56	57.60 ± 1.37	60.03 ± 1.23	61.92 ± 1.11	63.13 ± 1.22	63.84 ± 1.25	64.39 ± 1.20	63.85 ± 1.05
	SCe	52.77 ± 1.56	54.57 ± 0.52	54.91 ± 0.41	54.95 ± 0.36	54.59 ± 0.43	54.00 ± 0.36	53.61 ± 0.42	51.65 ± 0.16
	USC	52.77 ± 1.56	57.50 ± 1.33	59.66 ± 1.35	60.02 ± 1.44	–	–	–	–
	Random	52.77	52.77	52.77	52.77	52.77	52.77	52.77	52.77
Qwen 8B	EBA	84.46 ± 1.09	86.72 ± 1.15	87.78 ± 0.90	88.65 ± 0.73	89.18 ± 0.60	89.67 ± 0.76	89.70 ± 0.63	89.85 ± 0.51
	SCe	84.46 ± 1.09	86.74 ± 0.13	87.05 ± 0.20	86.99 ± 0.21	87.24 ± 0.16	87.10 ± 0.10	86.88 ± 0.11	86.93 ± 0.15
	USC	84.46 ± 1.09	88.28 ± 0.94	88.29 ± 0.81	–	–	–	–	–
	Random	84.46	84.46	84.46	84.46	84.46	84.46	84.46	84.46

Table 2: MATH500 results across sampling budgets (n), comparing Embedding-Based Agreement (EBA), Universal Self-Consistency (USC), Self-Certainty (SCe), and random selection for Llama 8B and Qwen 8B. Results show accuracy scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Method	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	EBA	29.21 ± 0.41	29.47 ± 0.10	29.58 ± 0.11	29.74 ± 0.12	29.81 ± 0.11	29.89 ± 0.10	29.96 ± 0.11	30.02 ± 0.10
	SCe	29.20 ± 0.41	29.50 ± 0.04	29.60 ± 0.03	29.80 ± 0.03	29.80 ± 0.04	29.90 ± 0.02	29.80 ± 0.03	29.80 ± 0.02
	USC	29.20 ± 0.41	28.80 ± 0.14	28.40 ± 0.11	28.20 ± 0.12	28.20 ± 0.13	28.20 ± 0.13	28.30 ± 0.11	28.60 ± 0.11
	Random	29.20	29.20	29.20	29.20	29.20	29.20	29.20	29.20
Qwen 8B	EBA	27.79 ± 0.10	27.87 ± 0.08	27.94 ± 0.09	28.03 ± 0.07	28.07 ± 0.07	28.11 ± 0.06	28.12 ± 0.07	28.19 ± 0.08
	SCe	27.80 ± 0.10	27.80 ± 0.01	27.80 ± 0.02	27.80 ± 0.01	27.80 ± 0.01	27.70 ± 0.01	27.70 ± 0.01	27.70 ± 0.02
	USC	27.80 ± 0.10	27.60 ± 0.09	27.50 ± 0.09	27.60 ± 0.09	27.60 ± 0.09	27.70 ± 0.10	27.60 ± 0.07	27.70 ± 0.09
	Random	27.80	27.80	27.80	27.80	27.80	27.80	27.80	27.80

Table 3: CNN/DM results across sampling budgets (n), comparing Embedding-Based Agreement (EBA), Universal Self-Consistency (USC), Self-Certainty (SCe), and random selection for Llama 8B and Qwen 8B. Results show ROUGE-1 scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Embedding	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	Qwen Emb.	65.98 ± 2.38	67.89 ± 2.23	68.72 ± 2.12	69.00 ± 1.97	69.82 ± 1.54	70.01 ± 1.96	70.11 ± 1.52	70.06 ± 1.58
	Gemma Emb.	65.98 ± 2.38	68.29 ± 2.21	69.34 ± 2.15	69.83 ± 1.64	69.83 ± 2.07	70.34 ± 1.84	70.60 ± 1.24	70.30 ± 1.50
	Native Emb.	65.98 ± 2.38	67.51 ± 2.34	68.07 ± 2.06	68.63 ± 1.92	68.70 ± 1.56	67.87 ± 1.79	67.88 ± 1.50	67.70 ± 1.56
Llama 70B	Qwen Emb.	79.89 ± 1.88	80.39 ± 1.63	81.18 ± 1.76	81.87 ± 1.50	81.88 ± 1.40	82.06 ± 1.44	81.67 ± 1.32	80.61 ± 1.34
	Gemma Emb.	79.89 ± 1.88	80.77 ± 1.52	81.30 ± 1.81	81.85 ± 1.49	82.24 ± 1.47	81.99 ± 1.30	82.28 ± 1.19	82.44 ± 1.08
	Native Emb.	79.89 ± 1.88	80.41 ± 1.88	80.93 ± 1.62	81.49 ± 1.50	81.96 ± 1.64	82.88 ± 1.59	82.76 ± 1.29	83.02 ± 1.21
Qwen 8B	Qwen Emb.	83.45 ± 2.10	86.07 ± 1.85	86.17 ± 1.46	87.32 ± 1.17	87.29 ± 1.26	88.00 ± 1.38	87.95 ± 1.21	88.11 ± 1.31
	Gemma Emb.	83.45 ± 2.10	85.72 ± 1.69	85.88 ± 1.39	87.17 ± 1.23	87.09 ± 1.36	86.94 ± 1.05	86.84 ± 1.03	86.76 ± 0.90
	Native Emb.	83.45 ± 2.10	85.80 ± 1.81	86.61 ± 1.45	86.61 ± 1.19	86.74 ± 1.20	86.71 ± 1.13	86.63 ± 1.19	87.02 ± 1.14
Qwen 32B	Qwen Emb.	87.16 ± 1.87	89.01 ± 1.52	88.94 ± 1.30	89.70 ± 1.39	89.85 ± 1.08	90.07 ± 0.98	90.23 ± 0.99	90.76 ± 0.99
	Gemma Emb.	87.16 ± 1.87	88.83 ± 1.26	89.26 ± 1.28	89.71 ± 1.17	89.54 ± 1.05	89.55 ± 1.08	89.52 ± 0.88	89.73 ± 0.82
	Native Emb.	87.16 ± 1.87	88.52 ± 1.27	89.00 ± 1.41	89.38 ± 1.04	89.66 ± 1.28	89.13 ± 1.03	89.22 ± 1.26	89.66 ± 1.17

Table 4: EBA performance on HumanEval across model scales (Llama 8B, Llama 70B, Qwen 8B, Qwen 32B) and embedding spaces, including Qwen embeddings, Gemma embeddings, and native hidden representations. Results show accuracy scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Embedding	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	Qwen Emb.	52.77 ± 1.56	57.60 ± 1.37	60.03 ± 1.23	61.92 ± 1.11	63.13 ± 1.22	63.84 ± 1.25	64.39 ± 1.20	63.85 ± 1.05
	Gemma Emb.	52.77 ± 1.56	56.71 ± 1.30	59.39 ± 1.16	61.57 ± 1.60	62.57 ± 1.30	63.58 ± 1.42	64.17 ± 1.17	65.14 ± 1.06
	Native Emb.	52.77 ± 1.56	56.19 ± 1.20	58.68 ± 1.52	60.51 ± 1.17	61.57 ± 1.26	61.96 ± 1.04	62.04 ± 1.13	62.06 ± 0.75
Llama 70B	Qwen Emb.	70.04 ± 1.44	73.15 ± 1.29	75.34 ± 1.16	76.92 ± 1.05	77.64 ± 1.05	78.29 ± 0.98	78.53 ± 0.84	78.83 ± 0.85
	Gemma Emb.	70.04 ± 1.44	72.81 ± 1.31	74.68 ± 1.34	75.68 ± 1.06	76.20 ± 1.11	76.74 ± 1.04	76.96 ± 0.92	76.60 ± 1.04
	Native Emb.	70.04 ± 1.44	73.46 ± 1.33	75.69 ± 0.96	77.00 ± 1.03	77.71 ± 0.96	78.30 ± 1.03	78.82 ± 0.74	78.77 ± 0.71
Qwen 8B	Qwen Emb.	84.46 ± 1.09	86.72 ± 1.15	87.78 ± 0.90	88.65 ± 0.73	89.18 ± 0.60	89.67 ± 0.76	89.70 ± 0.63	89.85 ± 0.51
	Gemma Emb.	84.46 ± 1.09	85.77 ± 0.91	86.49 ± 0.96	87.02 ± 0.82	87.55 ± 0.69	87.88 ± 0.78	87.97 ± 0.73	88.43 ± 0.68
	Native Emb.	84.46 ± 1.09	85.60 ± 0.96	86.30 ± 1.02	86.94 ± 0.75	87.19 ± 0.82	87.48 ± 0.74	86.96 ± 0.77	86.23 ± 0.64
Qwen 32B	Qwen Emb.	86.57 ± 1.08	89.03 ± 0.97	90.03 ± 0.82	90.68 ± 0.55	91.21 ± 0.62	91.25 ± 0.50	91.55 ± 0.47	91.38 ± 0.49
	Gemma Emb.	86.57 ± 1.08	88.14 ± 0.82	88.78 ± 0.81	89.35 ± 0.90	89.92 ± 0.74	90.06 ± 0.71	90.18 ± 0.67	90.27 ± 0.70
	Native Emb.	86.57 ± 1.08	87.92 ± 1.01	88.94 ± 0.87	89.40 ± 1.00	89.87 ± 0.73	89.91 ± 0.70	90.02 ± 0.64	89.90 ± 0.54

Table 5: EBA performance on MATH500 across model scales (Llama 8B, Llama 70B, Qwen 8B, Qwen 32B) and embedding spaces, including Qwen embeddings, Gemma embeddings, and native hidden representations. Results show accuracy scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Embedding	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	Qwen Emb.	29.21 ± 0.41	29.47 ± 0.10	29.58 ± 0.11	29.74 ± 0.12	29.81 ± 0.11	29.89 ± 0.10	29.96 ± 0.11	30.02 ± 0.10
	Gemma Emb.	29.21 ± 0.41	29.33 ± 0.12	29.46 ± 0.11	29.57 ± 0.09	29.62 ± 0.09	29.68 ± 0.09	29.67 ± 0.09	29.67 ± 0.08
	Native Emb.	29.21 ± 0.41	29.25 ± 0.09	29.38 ± 0.12	29.50 ± 0.12	29.58 ± 0.11	29.61 ± 0.08	29.67 ± 0.08	29.75 ± 0.08
Llama 70B	Qwen Emb.	29.48 ± 0.44	29.72 ± 0.11	29.87 ± 0.12	29.95 ± 0.10	30.02 ± 0.10	30.08 ± 0.10	30.17 ± 0.08	30.25 ± 0.09
	Gemma Emb.	29.48 ± 0.44	29.63 ± 0.11	29.79 ± 0.09	29.89 ± 0.11	29.96 ± 0.10	29.99 ± 0.09	30.01 ± 0.09	30.00 ± 0.08
	Native Emb.	29.48 ± 0.44	29.52 ± 0.14	29.68 ± 0.10	29.78 ± 0.11	29.86 ± 0.10	29.92 ± 0.10	29.99 ± 0.07	30.02 ± 0.07
Qwen 8B	Qwen Emb.	27.79 ± 0.10	27.87 ± 0.08	27.94 ± 0.09	28.03 ± 0.07	28.07 ± 0.07	28.11 ± 0.06	28.12 ± 0.07	28.19 ± 0.08
	Gemma Emb.	27.79 ± 0.10	27.87 ± 0.08	27.92 ± 0.09	27.97 ± 0.07	27.97 ± 0.08	28.00 ± 0.07	28.03 ± 0.06	28.08 ± 0.07
	Native Emb.	27.79 ± 0.10	27.81 ± 0.09	27.87 ± 0.09	27.92 ± 0.07	27.94 ± 0.07	27.93 ± 0.09	27.95 ± 0.06	27.95 ± 0.09
Qwen 32B	Qwen Emb.	25.49 ± 0.12	25.49 ± 0.10	25.51 ± 0.10	25.48 ± 0.11	25.49 ± 0.08	25.47 ± 0.12	25.48 ± 0.11	25.45 ± 0.10
	Gemma Emb.	25.49 ± 0.12	25.48 ± 0.10	25.48 ± 0.11	25.49 ± 0.08	25.49 ± 0.11	25.49 ± 0.09	25.47 ± 0.09	25.47 ± 0.11
	Native Emb.	25.49 ± 0.12	25.49 ± 0.09	25.51 ± 0.08	25.51 ± 0.11	25.48 ± 0.08	25.50 ± 0.10	25.49 ± 0.08	25.45 ± 0.09

Table 6: EBA performance on CNN/DM across model scales (Llama 8B, Llama 70B, Qwen 8B, Qwen 32B) and embedding spaces, including Qwen embeddings, Gemma embeddings, and native hidden representations. Results show ROUGE-1 scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Method	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	EBA (Full Ans)	52.77 ± 1.56	57.60 ± 1.37	60.03 ± 1.23	61.92 ± 1.11	63.13 ± 1.22	63.84 ± 1.25	64.39 ± 1.20	63.85 ± 1.05
	EBA (Extracted Ans)	52.77 ± 1.56	60.06 ± 1.43	63.66 ± 1.24	66.36 ± 1.12	67.53 ± 1.02	68.41 ± 1.04	68.90 ± 1.01	69.20 ± 0.80
	SC	52.77 ± 1.56	58.22 ± 0.98	63.85 ± 1.29	67.32 ± 1.14	69.36 ± 1.02	70.57 ± 0.98	71.23 ± 0.78	71.90 ± 0.70
Qwen 8B	EBA (Full Ans)	84.46 ± 1.09	86.72 ± 1.15	87.78 ± 0.90	88.65 ± 0.73	89.18 ± 0.60	89.67 ± 0.76	89.70 ± 0.63	89.85 ± 0.51
	EBA (Extracted Ans)	84.46 ± 1.09	87.73 ± 0.85	88.94 ± 0.88	89.58 ± 0.51	90.25 ± 0.64	90.42 ± 0.44	90.64 ± 0.46	90.51 ± 0.34
	SC	84.46 ± 1.09	86.38 ± 0.94	88.84 ± 0.88	90.22 ± 0.78	90.62 ± 0.57	90.88 ± 0.48	91.16 ± 0.43	91.10 ± 0.34

Table 7: Comparison of self-consistency (SC) and embedding-based agreement (EBA) using the complete generated response (EBA Full Ans) and only the extracted final answer (EBA Extracted Ans) on MATH500. Results are reported across different sample sizes (n) for Llama 8B and Qwen 8B models. Results show accuracy scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Method	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 70B	EBA (Full Ans)	70.04 ± 1.44	73.15 ± 1.29	75.34 ± 1.16	76.92 ± 1.05	77.64 ± 1.05	78.29 ± 0.98	78.53 ± 0.84	78.83 ± 0.85
	EBA (Extracted Ans)	70.04 ± 1.44	75.57 ± 1.24	77.74 ± 1.09	79.42 ± 0.83	80.51 ± 0.70	81.08 ± 0.66	81.30 ± 0.64	81.50 ± 0.46
	SC	70.04 ± 1.44	74.86 ± 1.17	78.34 ± 0.87	79.99 ± 0.75	80.69 ± 0.77	81.10 ± 0.54	81.50 ± 0.51	81.72 ± 0.38
Qwen 32B	EBA (Full Ans)	86.57 ± 1.08	89.03 ± 0.97	90.03 ± 0.82	90.68 ± 0.55	91.21 ± 0.62	91.25 ± 0.50	91.55 ± 0.47	91.38 ± 0.49
	EBA (Extracted Ans)	86.57 ± 1.08	89.62 ± 0.90	90.68 ± 0.71	91.60 ± 0.41	92.41 ± 0.46	92.72 ± 0.31	92.99 ± 0.36	93.30 ± 0.27
	SC	86.57 ± 1.08	88.86 ± 1.05	91.00 ± 0.75	92.03 ± 0.60	92.57 ± 0.45	92.90 ± 0.49	93.16 ± 0.31	93.28 ± 0.26

Table 8: Comparison of self-consistency (SC) and embedding-based agreement (EBA) using the complete generated response (EBA Full Ans) and only the extracted final answer (EBA Extracted Ans) on MATH500. Results are reported across different sample sizes (n) for Llama 70B and Qwen 32B models. Results show accuracy scores multiplied by 100 with corresponding standard deviation as a function of the number of sampled generations (n).

Model	Method	$n = 1$	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
Llama 8B	EBA	52.77 ± 1.56	57.60 ± 1.37	60.03 ± 1.23	61.92 ± 1.11	63.13 ± 1.22	63.84 ± 1.25	64.39 ± 1.20	63.85 ± 1.05
	Nearest	52.77	57.50	60.05	61.35	62.98	63.62	64.12	63.87
	Farthest	52.77	46.99	42.78	40.10	37.59	35.71	32.60	29.98
	Random	52.77	52.77	52.77	52.77	52.77	52.77	52.77	52.77
Qwen 8B	EBA	84.46 ± 1.09	86.72 ± 1.15	87.78 ± 0.90	88.65 ± 0.73	89.18 ± 0.60	89.67 ± 0.76	89.70 ± 0.63	89.85 ± 0.51
	Nearest	84.46	85.50	86.21	86.06	86.24	86.55	85.92	85.31
	Farthest	84.46	82.47	81.22	79.96	78.52	77.80	76.16	74.73
	Random	84.46	84.46	84.46	84.46	84.46	84.46	84.46	84.46

Table 9: Comparison of generation selection strategies on MATH500 for Llama 8B and Qwen 8B across increasing numbers of sampled generations (n). EBA selects generations using agreement-based clustering, while *Nearest* and *Farthest* choose generations according to their distance from the global centroid in embedding space. *Random* corresponds to the baseline performance at $n = 1$. Results show accuracy scores multiplied by 100 with corresponding standard deviation (when possible) as a function of the number of sampled generations (n).

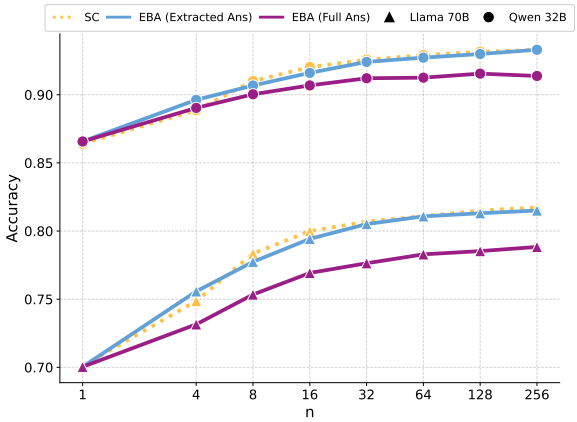

Figure 7: Comparison of self-consistency (SC) and embedding-based aggregation method using the complete generation (EBA (Full Ans)) and using only the extracted final answer (EBA (Extracted Ans)) on MATH500. Results are shown for Llama 70B and Qwen 32B models.