Granuscore: A Reference-Free Measure of Granularity for Text Analysis and Question Answering

Anonymous ACL submission

Abstract

Natural language conveys information at varying levels of abstraction, from fine-grained references to broad descriptions. While granularity is fundamental to human communication, existing measures mostly capture surface detail or sentence specificity rather than abstraction within semantic hierarchies.

We introduce **Granuscore**, a reference-free measure of granularity that leverages structural properties of a hierarchical embedding space. We validate Granuscore both empirically and conceptually. Granuscore reliably recovers granularity orderings on the GRANOLA-EQ dataset and aligns with expected differences in abstraction across sections of scientific papers. Across multiple domains, we further show that granularity captures systematic, non-linear variation in sentence specificity beyond sentence length. Finally, we apply Granuscore to several question answering datasets and demonstrate that it provides a useful lens for characterizing dataset difficulty and understanding systematic differences in model performance and outputs. Together, these results position Granuscore as a scalable and broadly applicable tool for analyzing granularity in text.

1 Introduction

Natural language varies not only in *what* information is conveyed, but also in *how abstract or fine-grained* that information is expressed. A speaker may refer to a person as *a skateboarder* or as *Tony Hawk*, or locate an event in *America* rather than *San Diego*. Such choices preserve truth conditions while changing the level of abstraction at which referential units are realized. We refer to this dimension as *granularity*.

Granularity is not incidental: humans actively adapt it to conversational context, shared knowledge, and epistemic uncertainty. When facts are uncertain or only partially known, speakers often prefer coarser descriptions that remain informative without overcommitting (e.g., stating that someone was born in America rather than in San Diego). Conversely, when common ground is established, more fine-grained references become appropriate (Yona et al., 2024). Granularity should therefore be understood as a deliberate communicative strategy that balances informativeness, reliability, and audience expectations

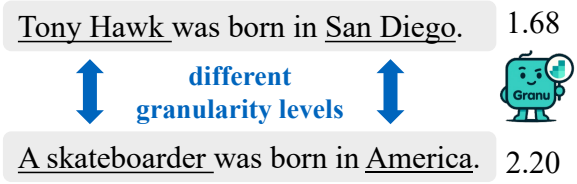

Figure 1: Two sentences with underlined referential units that differ in their level of granularity. Replacing a fine-grained referent with a more abstract one changes the granularity of the sentence, reflected in Granuscore.

A closely related concept is *sentence specificity*, which describes how much concrete information a sentence provides and how well it stands on its own, such that different readers arrive at a similar interpretation (Li et al., 2016; Ko et al., 2019). While increasing granularity often leads to higher specificity, the two are not equivalent. Sentences can become more specific by adding descriptive details without changing the abstraction level of their underlying referents. Granularity instead captures where referential choices lie within a structured hierarchy of abstraction, whereas specificity reflects the overall amount of detail in a sentence.

While granularity has been implicitly discussed in work on specificity, informativeness, and semantic hierarchies, prior approaches do not provide a reference-free, scalable measure of granularity at the sentence level that captures hierarchical abstraction (Thoppilan et al., 2022; Adiwardana et al., 2020; Ko et al., 2019; Li et al., 2016).

We introduce **Granuscore**, a reference-free mea-

sure of granularity in text. It exploits structural properties of a hierarchical embedding space to estimate the level of abstraction. Lower Granuscore values correspond to more fine-grained (i.e., more granular) expressions, while higher values indicate more abstract references. Figure 1 illustrates this behavior using two sentences with similar structure but different levels of granularity. Granuscore identifies the referential units (e.g., *Tony Hawk* vs. *a skateboarder*, *San Diego* vs. *America*) and assigns a scalar score reflecting their granularity. The sentence referring to a named individual and a city is more fine-grained and therefore receives a lower Granuscore (1.68) than the sentence using a profession and a country (2.20).

Beyond providing a numerical characterization of granularity at the sentence level, Granuscore enables large-scale analysis of text collections and model behavior. Prior work has shown that the abstraction level influences user satisfaction in dialogue systems (Adiwardana et al., 2020; Thoppilan et al., 2022) and accessibility in simplified language settings (OECD, 2013; Anschütz et al., 2025). Granuscore can serve both as an analytical lens and as a potential training signal for adapting model responses to appropriate levels of abstraction without requiring additional human annotation.

In this work, we focus on the analytical use of Granuscore. We apply it to question answering datasets to study how granularity relates to dataset difficulty, model performance, and systematic differences between questions, gold answers, and model outputs.

Our contributions are as follows:

- We introduce **Granuscore**, a reference-free measure of informational granularity in text.
- We validate Granuscore both empirically and conceptually. Granuscore reliably recovers human-annotated granularity orderings on the GRANOLA-EQ (Yona et al., 2024) dataset and aligns with expected differences in granularity across sections of scientific papers.
- We show that Granuscore captures systematic variation in sentence specificity beyond sentence length.
- We demonstrate the practical relevance of Granuscore in question answering by showing that it serves as a useful lens for characterizing QA datasets' difficulty and understanding systematic differences in model performance and outputs.

2 Background and Related Work

Term Specificity Prior work has studied *term specificity* as a property to identifying index terms to distinguish one class of documents from other classes. Kim (2006) identifies several types of term specificity, among which *hierarchical specificity* and *posting specificity* are most closely related to our notion of granularity. Hierarchical specificity captures semantic narrowness within a hierarchy (e.g., *cat* being more specific than *domestic animal*), while posting specificity reflects how selectively a term is used within a given collection (e.g., *thesaurus* having fewer postings than *indexing* in a bibliographic database).

We translate these notions and operationalize them using structural properties of a hierarchical embedding space. Hierarchical specificity is captured through the relative position of a term within the embedding hierarchy. Posting specificity is reflected in the local neighborhood structure of a term, indicating how selectively it is embedded among related terms.

Sentence Specificity Sentence specificity refers to the extent to which a sentence conveys concrete information and supports consistent interpretation across readers (Li et al., 2016; Ko et al., 2019). Prior work has shown its relevance for reading comprehension (Dixon, 1987) and establishing common ground in dialogue (Djalali et al., 2011).

Existing approaches typically rely on human annotation or supervised prediction of sentence specificity (Gao et al., 2019; Ko et al., 2019; Li et al., 2016), with sentence length serving as a strong baseline feature. In contrast, our work focuses on granularity as a reference-free measure of abstraction within semantic hierarchies.

Training Signals for Informativeness and Interestingness The informativeness of model responses plays a central role in user engagement and response quality (Adiwardana et al., 2020; Thoppilan et al., 2022). While early work relies on human annotation to supervise informativeness (Adiwardana et al., 2020; Thoppilan et al., 2022), more recent approaches use LLM-based judges to obtain relative preference signals by comparing response pairs (Wu et al., 2025). Relatedly, Onozeki and Inaba (2025) introduce interestingness as a training

signal and assign scores using an LLM judge.

In contrast to these approaches, which depend on human supervision or pairwise or model-based judgments, Granuscore provides a reference-free and scalable signal that measures granularity.

Hierarchy Transformers Chen et al. (2024) introduce HIERARCHY TRANSFORMERS, which retrain transformer encoders to explicitly encode hierarchical structure in the embedding space using a hyperbolic geometry modeled as a Poincaré ball.

We build on this by exploiting the resulting hierarchical embedding space. In particular, we use the distance to the center of the embedding space as a primary signal for granularity.

Granularity Evaluation Recent work has begun to evaluate the granularity of language model outputs. Huang et al. (2023) propose an automatic benchmark for measuring specificity using transitive relations derived from Wikidata, a notion that closely aligns with our understanding of granularity. However, the induced orderings can yield unintuitive comparisons, for example, ranking *Mexico* as more granular than *Colombia*, or ranking *historian* as more granular than *writer*. We therefore acknowledge this dataset but refrain from using it in our experiments.

Yona et al. (2024) introduce GRANOLA-EQ, a question answering dataset with explicitly controlled answer granularity levels. They show that standard decoding methods tend to produce overly granular and often incorrect answers. We build on this work by using GRANOLA-EQ to train Granuscore. In addition, we extend their analysis by applying granularity estimation to a broader set of QA datasets, studying how granularity relates to model outputs, model correctness, and dataset difficulty.

3 Granuscore

Granuscore measures semantic granularity by exploiting structural properties of a hierarchical embedding space, where lower scores correspond to finer-grained expressions. We build on the HIERARCHY TRANSFORMER MODEL[1] introduced by Chen et al. (2024). We use the variant trained on the WordNet hierarchy, as WordNet provides broad-coverage, general-purpose commonsense structure. The model embeds lexical entities into a hyperbolic space modeled as a Poincaré ball. In this geometry, hierarchical relations are encoded by radial distance from the origin: more specific concepts are embedded further away, while more general concepts lie closer to the center.

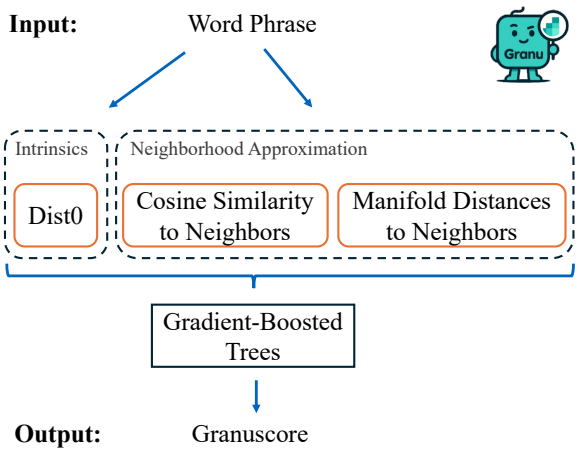

Figure 2: Granuscore pipeline: extraction of hierarchical depth (*Dist0*) and local neighborhood, followed by gradient-boosted trees to a scalar granularity score.

We denote the radial distance of an embedding from the origin as *Dist0*. This feature directly captures hierarchical depth and serves as a primary signal for granularity.

Beyond hierarchical position, we hypothesize that the *local structure* of the embedding space also reflects granularity. Fine-grained concepts tend to exhibit a steeper decay in similarity to neighboring entities, while broader and more homogeneous similarity distributions characterize coarse-grained concepts. Granuscore, therefore, combines global hierarchical information with features derived from the local similarity neighborhood.

Figure 2 provides an overview of the Granuscore pipeline. Given an input word phrase, the model first obtains a hierarchical embedding and extracts *Dist0*. It then approximates the local neighborhood structure via similarity-based retrieval over a Wikidata-derived embedding index. The resulting features are finally mapped to a scalar granularity score using gradient-boosted trees.

3.1 Neighborhood Construction

As the embedding space is unbounded, we approximate its local structure using a finite subset of entities. We construct this proxy space from 50,000 randomly sampled Wikidata entities[2]. For each entity, we use its title as the textual representation.

[1] https://huggingface.co/Hierarchy-Transformers/HiT-MiniLM-L12-WordNetNoun

[2] https://huggingface.co/datasets/philippesaade/wikidata

Wikidata offers broad topical coverage and a relatively clean entity structure, making it a suitable general-purpose semantic reference. We choose an index size of 50,000 entities as a trade-off between computational efficiency and neighborhood fidelity. In preliminary experiments, this size yielded stable neighborhood structures, while larger indices substantially increased runtime.

All entity embeddings are indexed using FAISS (Douze et al., 2025). For a given input embedding, we retrieve the top 1,000 nearest neighbors using cosine similarity.

3.2 Granularity Prediction

We map the extracted feature vectors to scalar granularity scores using gradient-boosted decision trees, implemented with LIGHTGBM (Ke et al., 2017). We use the raw similarity and distance values as input features, allowing the model to capture fine-grained patterns in the local neighborhood geometry that would be lost if pre-aggregation were used. Details on the training procedure and model hyperparameters are provided in Appendix A

3.3 Dataset

We train the LIGHTGBM model on GRANOLA-EQ (Yona et al., 2024), an extension of the ENTITYQUESTIONS dataset (Sciavolino et al., 2021). GRANOLA-EQ provides multiple answers per question, ordered by increasing granularity level.

During preprocessing, we remove samples with more than five answers (fewer than 0.3% of the data) to reduce noise from excessively fine-grained distinctions. The resulting dataset contains, on average, approximately three answers per question (2% with one answer, 23% with two, 60% with three, 14% with four, and 1% with five answers).

Due to the construction of GRANOLA-EQ, the same answer entity may receive different granularity levels across samples depending on question context (e.g., *England* appears 87 times with a mean granularity of 3.15 and variance 0.20). This variability constitutes a limitation of the dataset and should be taken into account when interpreting inter-question comparisons.

To prevent data leakage, we enforce that no answer entity appears in more than one dataset split. The final dataset consists of 6,314 training samples and 1,230 samples each for development and test. Additional dataset statistics are provided in Appendix B.

3.4 Extension to Multi-Word Inputs

We extend Granuscore to sentences and longer text spans by decomposing the text into referential units. We use a SPACY-based splitter (Honnibal et al., 2020). Noun phrases are kept intact, while stopwords and non-informative symbols are removed. This preserves referential expressions that convey granularity while avoiding fragmentation of multi-word concepts.

We avoid decomposing inputs into atomic facts, as a single fact may contain multiple entities with different granularity levels, making fine-grained attribution difficult. For example, in Figure 1, both *Tony Hawk* and *San Diego* independently contribute to the sentence's granularity, such that modifying either referent changes the perceived granularity. Moreover, atomic fact decomposition can introduce or duplicate lexical material not explicitly present in the original sentence, which may bias the resulting scores Wanner et al. (2024).

Finally, we compute the granularity score for a multi-word input by averaging the granularity scores of its referential units. If no units are extracted (e.g., the input consists only of stopwords), we assign a Granuscore of 6.

4 Results

We evaluate Granuscore on GRANOLA-EQ (subsection 4.1), compare scores across paper sections (subsection 4.2), and analyze correlations with human-annotated sentence specificity (subsection 4.3).

4.1 GRANOLA-EQ

We evaluate Granuscore on GRANOLA-EQ using three metrics: Kendall's τ, Intra-sample Kendall's τ, and Intra-sample Exact ordering accuracy.

Kendall's τ is computed across all answer pairs from all samples. It reflects how well the model assigns granularity scores across unrelated entities, capturing global ordering behavior. In contrast, Intra-sample Kendall's τ is computed per sample and averaged, measuring within-question ordering quality. Exact accuracy reports the proportion of samples in which all answers are ordered perfectly by Granuscore (i.e., Kendall's $\tau = 1$).

Table 1 reports the performance of Granuscore across dataset splits. Incorporating neighborhood-based features improves performance over the *Dist0*-only baseline across all metrics and splits.

Split	τ	$\Delta\tau$	Intra-τ	ΔIntra-τ	Exact	ΔExact
Train	0.681	+0.225	0.902	+0.189	0.836	+0.220
Dev	0.476	+0.052	0.714	+0.047	0.655	+0.045
Test	0.483	+0.036	0.734	+0.006	0.678	+0.002

Table 1: Performance of Granuscore on GRANOLA-EQ. Kendall's τ measures global ordering consistency across all answers, Intra-τ measures average within-question ordering, and Exact reports the proportion of samples with perfectly ordered answers. Δ values denote improvements over a baseline that orders answers solely by *Dist0*.

Level	1	2	3	4	5
Granuscore	1.47	1.98	2.23	2.27	2.06

Table 2: Average Granuscore values per granularity level on the GRANOLA-EQ test split.

Section	Mean	Std
Introduction	2.176	0.082
Related Work	2.126	0.085

Table 3: Comparison of Introduction and Related Work section Granuscore.

On the test split, Granuscore achieves a global Kendall's τ of 0.483, an Intra-sample Kendall's τ of 0.734, and an Exact accuracy of 0.678.

Bootstrap resampling on the test split shows a statistically significant improvement in global Kendall's τ over the *Dist0*-only baseline ($p < 0.05$). In contrast, improvements in Intra-sample Kendall's τ and Exact accuracy are not statistically significant. On the development split, bootstrap resampling shows statistically significant improvements for both global and Intra-sample Kendall's τ ($p < 0.05$).

To contextualize absolute values of Granuscore, Table 2 reports average scores per granularity level learned from GRANOLA-EQ on the test split. Differences between adjacent granularity levels 1 to 4 are 0.51 (levels 1–2), 0.25 (levels 2–3), and 0.04 (levels 3–4), indicating progressively smaller score increments at higher granularity levels. Granularity level 5 is represented by only 6 samples in the test split. Consequently, its lower average score should be interpreted with caution.

Pairwise comparisons between all granularity levels are statistically significant (one-sided Wilcoxon signed-rank test; lower granularity $<$ higher granularity; $p \leq 3.1 \times 10^{-2}$).

4.2 Granuscore Across Paper Sections

We evaluate whether Granuscore reflects systematic differences in granularity across paper sections by applying it to scientific articles from the S2ORC corpus (Lo et al., 2020). We sample the first 1,000 papers and retain only those containing standard *Introduction*, *Related Work*, and *Conclusion* sections to ensure comparability and exclude atypical document structures.

For each paper and section, we select a paragraph containing at least ten referential units. For the Introduction sections, we select the first paragraph that matches this criterion. For the Related Work sections, we skip the opening paragraph, as it often serves as a brief transition, and select the first subsequent paragraph that meets the criterion. If no such paragraph exists, we fall back to the opening paragraph if it also meets the criterion. This procedure yields 978 papers for comparison.

Table 3 shows the Granuscore values for the Introduction and Related Work sections. Granuscore differs significantly between the two sections (two-sided t-test: $p \leq 4.2 \times 10^{-47}$; Wilcoxon signed-rank test: $p \leq 8.4 \times 10^{-46}$), with 70.04% of paired comparisons exhibiting a higher score for Introduction than for Related Work

Relative to the Granuscore reference scale in Table 2, both sections fall between granularity levels 2 and 3. The observed difference between the two sections (0.05) is smaller than the transition between adjacent granularity levels (e.g., 2–3, 0.25), placing it within the range of within-level variation.

4.3 Correlation to Sentence Specificity

We analyze the relationship between Granuscore and sentence specificity using datasets with human-annotated specificity from Ko et al. (2019) and Li et al. (2016). The former covers movie reviews, tweets, and Yelp reviews, while the latter contains annotated sentences from news articles.

Sentence length (word count) alone is a strong baseline predictor of specificity, with Spearman cor-

Domain	Expl. Dev. (Length)	Expl. Dev. (Len+Gran)	Δ
movie	0.38	0.47	+0.08
twitter	0.24	0.37	+0.12
yelp	0.52	0.56	+0.04
news	0.45	0.55	+0.10

Table 4: Explained deviance of generalized additive models (GAMs) predicting sentence specificity. All granularity smooth terms are statistically significant ($p < 10^{-12}$).

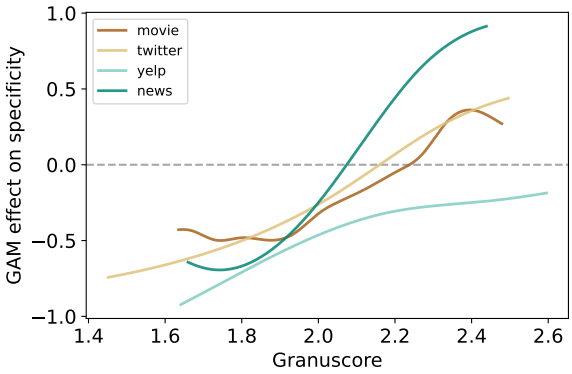

Figure 3: Effect of Granuscore on sentence specificity across domains. Lower specificity scores correspond to more specific sentences. The plotted range is restricted to the 1st–99th percentiles of Granuscore to avoid sparse-support regions.

relations of 0.445 (Twitter), 0.581 (movie reviews), 0.676 (Yelp), and 0.665 (news).

To quantify the effect of Granuscore on sentence specificity beyond sentence length, we fit Generalized Additive Models (GAMs). This allows us to capture non-linear effects of sentence length (word count) and Granuscore. Table 4 shows that adding Granuscore consistently improves explained deviance over a length-only baseline across all domains. Absolute gains range from +0.04 (Yelp) to +0.12 (Twitter), corresponding to relative improvements of 8–50%.

Figure 3 shows the estimated effect of Granuscore on sentence specificity across domains, restricted to the data-supported Granuscore range. In all domains, the relationship is non-linear. At low levels of granularity, Granuscore exhibits a negative effect on the specificity score, indicating that lower Granuscore values are associated with more specific sentences. As Granuscore increases, the magnitude of this negative effect decreases.

For the movie, Twitter, and news domains, the effect crosses zero at mid-range Granuscore values, after which higher Granuscore is associated with less specific sentences. In contrast, the Yelp domain does not exhibit a sign change. The effect remains negative throughout the observed range, although its magnitude steadily decreases, indicating diminishing influence at higher Granuscore.

For completeness, we show the effect of sentence length on specificity in Appendix C.

5 Applying Granuscore to QA Datasets

We apply Granuscore to several widely used QA datasets to investigate how granularity affects dataset properties and model performance. We use the public splits of FACTS PARAMETRIC (Cheng et al., 2025) (1,047 samples), SIMPLEQA (Wei et al., 2024) (4,255), SQUAD (Rajpurkar et al., 2016) (10,570), and TRUTHFULQA (Lin et al., 2022) (817), resulting in a total of 16,689 samples.

For each dataset, we compute Granuscore for both questions and gold answers. When multiple gold answers are provided, we average their Granuscore values.

To relate granularity to model behavior, we evaluate model correctness using QWEN3-8B (Yang et al., 2025). We select this model as a well-established mid-sized language model that supports both standard and reasoning-based generation, allowing us to additionally compare performance with and without explicit reasoning.

For answer generation, we use a maximum output length of 512 tokens for standard and 2,048 tokens for reasoning-based generation. We instruct the model to produce answers of at most five sentences. We retain only responses that terminate before reaching the respective token limits, ensuring that all evaluated outputs are complete and not truncated. Further details on prompting, correctness evaluation, and model performance are given in Appendix D.

In the following, we report results for QWEN3-8B without explicit reasoning. We find that reasoning outputs exhibit slightly lower Granuscore but do not substantially alter trends. Detailed comparisons are reported in Appendix D.

Granuscore of Questions and Gold Answers In Figure 4, we relate model correctness to the Granuscore of questions and gold answers across QA datasets. Question and gold answer granularity differ substantially between datasets: with the exception of gold answer Granuscore of SIMPLEQA and FACTS PARAMETRIC, all pairwise comparisons are statistically significant (Mann-Whitney

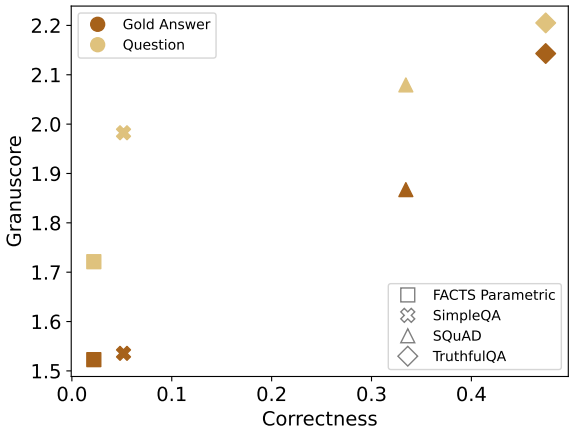

Figure 4: Granuscore versus answer correctness across QA datasets.

Type	Correct	Wrong	Not Att.
Question	2.09	2.02	2.07
Gold Answer	1.95	1.71	1.82
Answer	2.15	2.11	2.19
Gap	0.19	0.40	0.37

Table 5: Mean Granuscore values by response outcome (Correct, Wrong, and Not Attempted). Gap denotes the difference between output and gold-answer granularity.

U, $p \leq 8.8 \times 10^{-65}$). Model correctness also varies strongly across datasets, with accuracies of 2.20% on FACTS PARAMETRIC, 5.17% on SIMPLEQA, 33.44% on SQUAD, and 47.44% on TRUTHFULQA.

Across datasets, question and gold answer Granuscore are systematically related to model performance. Datasets characterized by lower Granuscore questions and answers exhibit substantially lower correctness, while higher Granuscore datasets are associated with improved performance.

Granuscore Across Response Outcomes In Table 5, we report mean Granuscore values for questions, gold answers, and model outputs, stratified by response outcome (correct, incorrect, and not-attempted). Granuscore distributions differ significantly across outcomes (Mann-Whitney U, $p \leq 2.32 \times 10^{-8}$), except for question granularity between correct and not attempted responses.

Model outputs associated with wrong answers exhibit the lowest average Granuscore (2.11), followed by correct answers (2.15), while not-attempted responses show the highest output granularity (2.19). The latter is expected, as abstentions typically consist of explicit statements indicating inability to provide an answer.

On the input side, incorrect responses are associated with lower-granularity questions and gold answers, followed by not-attempted cases and then correct responses.

Finally, we analyze the *granularity gap*, defined as the difference between model output and gold-answer Granuscore. The gap is substantially larger for incorrect and not-attempted responses than for correct ones. Using five-fold cross-validation, a logistic regression with granularity gap as the sole predictor achieves an average AUC of 0.66 ($\pm$ 0.004), indicating a moderate, stable association between granularity mismatch and response failure.

6 Discussion

Validation of Granuscore Granuscore's performance on the test split of GRANOLA-EQ highlights that granularity ranking, particularly for global ordering, is a challenging task. One key reason is that granularity in GRANOLA-EQ is explicitly context-dependent: the same entity may correspond to different granularity levels depending on the surrounding question.

This challenge is further reflected across the granularity scale (cf. Table 2). As granularity becomes coarser, the score differences between adjacent levels decrease. This suggests that identifying fine-grained entities is comparatively straightforward, whereas distinguishing between already coarse-grained entities is substantially more challenging, as the differences between them (e.g., an occupation versus a generic object) become increasingly ambiguous.

Consistent with this observation, Intra-sample Kendall's τ values show that ordering answers within the same question is substantially easier than establishing a global ordering across samples. Compared to a baseline that relies solely on hierarchical depth (*Dist0*), Granuscore achieves a statistically significant improvement of 3.6 percentage points in global Kendall's τ, reaching 0.483. Given the context dependence of granularity annotations and the intrinsic ambiguity of global comparisons, this result indicates strong overall ordering performance. In contrast, improvements for Intra-sample ordering are minor and not statistically significant, suggesting that neighborhood-based features primarily contribute to resolving non-trivial global ordering relationships across entities. Taken together, these findings indicate that Granuscore captures an

aggregated, context-averaged notion of granularity.

As an additional sanity check, we compared Granuscore values across the Introduction and Related Work sections of scientific papers. Based on discourse conventions, we expect Related Work sections to contain more fine-grained, detail-rich descriptions of prior studies, whereas Introduction sections typically provide higher-level overviews. Consistent with this expectation, Granuscore assigns lower scores to Related Work sections than to Introductions, further supporting the validity of the measure.

Correlation to Sentence Specificity We show that Granuscore captures a systematic aspect of sentence specificity beyond sentence length, which serves as a strong baseline indicator (Gao et al., 2019; Ko et al., 2019). The consistency of this relationship across heterogeneous domains further supports the robustness of Granuscore as a general granularity measure.

Granularity and QA Dataset Difficulty Our analysis of QA datasets shows that granularity provides a useful lens for characterizing dataset difficulty and explaining systematic differences in model performance.

We observe that incorrect responses are associated with the lowest question granularity, followed by not-attempted and then correct responses. This pattern suggests that question Granuscore serves as an indicator of uncertainty: lower-granularity questions are more likely to elicit incorrect or not-attempted answers. Granuscore may therefore be incorporated as a signal in deciding when to rely on a model's internal knowledge versus when to invoke external tools.

A similar, and more pronounced, pattern emerges for gold-answer granularity. Model correctness is strongly positively correlated with gold-answer Granuscore, with datasets containing higher-granularity answers consistently yielding better performance. Together, these results indicate that granularity captures dataset-level difficulty and helps explain variation in QA performance beyond surface-level properties.

Finally, we find that incorrect answers also exhibit lower output granularity than correct ones. In these cases, models tend to remain at the granularity level of the question rather than adapting their responses for complex inputs. This behavior suggests limited granularity adaptation under challenging conditions, where producing a coarser, more general answer might be preferable. This observation is consistent with the findings of Yona et al. (2024), who show that language models struggle to adjust their granularity appropriately.

Future Directions A natural next step is to use Granuscore as a training signal for language models. Analogous to prior work that incorporates objectives such as informativeness or interestingness (Adiwardana et al., 2020; Thoppilan et al., 2022; Onozeki and Inaba, 2025), Granuscore could guide models to produce responses with controlled levels of granularity.

In particular, learning to adapt output granularity based on contextual factors such as question difficulty represents a promising direction for future research. Beyond general response generation, Granuscore may also support controlled language adaptation, such as simplification, including summarization (Stoll et al., 2022) and definition generation (Ellinger et al., 2025), where appropriate granularity is crucial for producing accessible yet informative text.

7 Conclusion

We introduced Granuscore, a reference-free measure of granularity that quantifies the level of abstraction expressed in text using a hierarchical embedding space. Granuscore reliably recovers granularity orderings on the controlled GRANOLA-EQ benchmark, aligns with expected differences in abstraction across scientific paper sections, and captures systematic variation in sentence specificity beyond sentence length.

Applied to question answering, Granuscore provides a useful lens for characterizing dataset difficulty and understanding systematic differences in model performance and outputs.

Limitations

Embedding Backbone Granuscore relies on a single hierarchical embedding model fine-tuned on the WordNet hierarchy. We choose this variant because WordNet provides broad-coverage, general-purpose commonsense structure. This choice might limit granularity estimation for domains that are poorly represented in WordNet. Future work could explore domain-specific hierarchical models and evaluate their impact, whereas we intentionally focus on general applicability in this work.

Limited Model Evaluation on QA Datasets We demonstrate the practical relevance of Granuscore for question answering by analyzing its relationship to dataset difficulty and model performance. While our QA analysis spans multiple datasets and covers a total of 16,689 samples, we evaluate granularity effects using a single language model. Consequently, our findings primarily illustrate how it relates to QA in this specific setting, rather than establishing model-agnostic conclusions. A more extensive evaluation is beyond the scope of this paper and left for future work. We view our QA analysis as an application-driven case study demonstrating the usefulness of Granuscore rather than a comprehensive benchmark.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Miriam Anschütz, Thanh Mai Pham, Eslam Nasrallah, Maximilian Müller, Cristian-George Craciun, and Georg Groh. 2025. German4All – A Dataset and Model for Readability-Controlled Paraphrasing in German. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 390–407, Hanoi, Vietnam. Association for Computational Linguistics.

Jiaoyan Chen, Yuan He, Ian Horrocks, and Zhangdie Yuan. 2024. Language Models as Hierarchy Encoders. In *Advances in Neural Information Processing Systems 37*, pages 14690–14711, Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, Adam Bloniarz, Aijun Bai, Andrew Wang, Anfal Siddiqui, Arturo Bajuelos Castillo, Aviel Atias, Chang Liu, Corey Fry, and 46 others. 2025. The FACTS Leaderboard: A Comprehensive Benchmark for Large Language Model Factuality. *arXiv preprint*. ArXiv:2512.10791 [cs].

Peter Dixon. 1987. The processing of organizational and component step information in written directions. *Journal of Memory and Language*, 26(1):24–35.

Alex Djalali, David Clausen, Sven Lauer, Katrin Schultz, and Christopher Potts. 2011. Modeling Expert Effects and Common Ground Using Questions under Discussion. In *AAAI Fall Symposium: Building Representations of Common Ground with Intelligent Agents*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. *arXiv preprint*. ArXiv:2401.08281 [cs].

Lukas Ellinger, Miriam Anschütz, and Georg Groh. 2025. Simplifications Are Absolutists: How Simplified Language Reduces Word Sense Awareness in LLM-Generated Definitions. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 342–351, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yifan Gao, Yang Zhong, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2019. Predicting and Analyzing Language Specificity in Social Media Posts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6415–6422.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2023. Can Language Models Be Specific? How? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 716–727, Toronto, Canada. Association for Computational Linguistics.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Giyeong Kim. 2006. Relationship between index term specificity and relevance judgment. *Information Processing & Management*, 42(5):1218–1229.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain Agnostic Real-Valued Specificity Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6610–6617.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the Annotation of Sentence Specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3921–3927, Portorož, Slovenia. European Language Resources Association (ELRA).

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human

Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

OECD. 2013. OECD Skills Outlook 2013: First Results from the Survey of Adult Skills. *OECD Skills Outlook*, 2013. Publisher: OECD Publishing.

Hiroki Onozeki and Michimasa Inaba. 2025. Enhancing Coherence and Interestingness in Knowledge-Grounded Dialogue Generation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 1–19, Hanoi, Vietnam. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple Entity-Centric Questions Challenge Dense Retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marlene Stoll, Martin Kerwer, Klaus Lieb, and Anita Chasiotis. 2022. Plain language summaries: A systematic review of theory, guidelines and empirical research. *PLOS ONE*, 17(6):e0268789. Publisher: Public Library of Science.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, and 41 others. 2022. LaMDA: Language Models for Dialog Applications. *arXiv preprint*. ArXiv:2201.08239 [cs].

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024. DnDScore: Decontextualization and Decomposition for Factuality Verification in Long-Form Text Generation. *arXiv preprint*. ArXiv:2412.13175 [cs].

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint*. ArXiv:2411.04368 [cs].

Tianyi Wu, Jingwei Ni, Bryan Hooi, Jiaheng Zhang, Elliott Ash, See-Kiong Ng, Mrinmaya Sachan, and Markus Leippold. 2025. Balancing Truthfulness and Informativeness with Uncertainty-Aware Instruction Fine-Tuning. *arXiv preprint*. ArXiv:2502.11962 [cs].

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Narrowing the Knowledge Evaluation Gap: Open-Domain Question Answering with Multi-Granularity Answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6737–6751, Bangkok, Thailand. Association for Computational Linguistics.

Parameter	Value
Boosting type	GBDT
Objective	Regression
Evaluation metric	RMSE
Early stopping	200
Number of iterations	10,000
Learning rate	0.0107
Number of leaves	20
Maximum depth	Unlimited
Minimum data in leaf	35
Feature fraction	0.634
Bagging fraction	0.989
Bagging frequency	10
Dropout rate (DART)	0.1
Maximum bins	255
Random seed	0
Device	CPU

Table 6: LightGBM hyperparameters used for training Granuscore.

Type	Template
location	Where is [MASK] located?
	Where did [MASK] die?
founder	Who founded [MASK]?
company	Which company is [MASK] produced by?
owner	Who owns [MASK]?
performer	Who performed [MASK]?
person	Who is [MASK]'s child?
	Who is [MASK] married to?

Table 7: Relation categories and corresponding question templates used in GRANOLA-EQ.

A LightGBM Model Training

We train Granuscore using a LightGBM regression model on GRANOLA-EQ. Hyperparameters are selected with Optuna (Akiba et al., 2019), using 50 optimization trials on the development split. The final hyperparameter configuration is shown in Table 6.

B Granola-EQ

GRANOLA-EQ covers multiple relation types, represented through masked question templates. Table 7 lists the relation categories included in the dataset together with their corresponding question templates. Table 8 reports the distribution of relation types across the training, development, and test splits. The dataset covers a diverse range of relations, with location- and person-based questions forming the largest categories, while founder- and company-related questions occur less frequently.

Question Type	Train	Dev	Test
location	1452	278	132
person	707	144	320
performer	502	92	31
owner	411	104	89
founder	295	40	56
company	216	117	192
total	6314	1230	1230

Table 8: Distribution of relation types across the training, development, and test splits of GRANOLA-EQ.

Gran. Level	Count
1	216
2	2027
3	5217
4	1253
5	61
Mean	2.88
Variance	0.48

Table 9: Distribution of granularity levels in GRANOLA-EQ.

As shown in the table, the relative proportions differ across splits. This variation arises from enforcing a strict split by answer entity, ensuring that the same answer does not appear in multiple splits and preventing data leakage across training and evaluation.

Distribution of Granularity Levels

The distribution of answers across granularity levels for the full dataset is given in Table 9. The mean and variance indicate that most answers are located at coarse levels.

Context-dependent granularity variation

The granularity level of an entity is context-dependent: the same entity can be associated with different levels of abstraction depending on the question being asked. Table 10 illustrates this effect for the entity *England*. The distribution shows that identical entities may appear at multiple granularity levels across contexts, reflecting variation in the semantic resolution required by different questions. Consequently, Granuscore reflects an aggregated, context-averaged view of granularity, which is well-suited for global analysis and comparison.

Gran. Level	Count
2	2
3	71
4	13
5	1
Mean	3.15
Variance	0.20

Table 10: Distribution of annotated granularity levels for the entity *England* in GRANOLA-EQ.

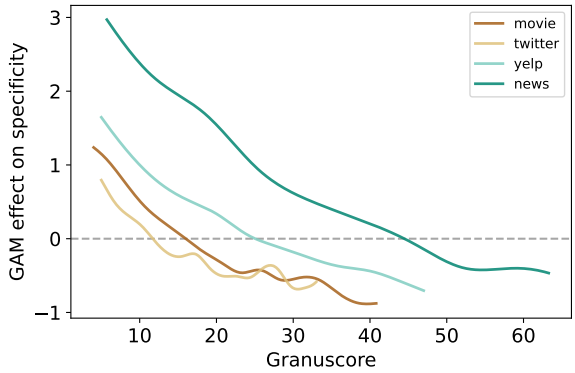

Figure 5: Effect of Length on sentence specificity across domains. Longer sentences correspond to more specific sentences. The plotted range is restricted to the 1st–99th percentiles of Granuscore to avoid sparse-support regions.

C Correlation to Sentence Specificity

For completeness, Figure 3 shows the estimated effect of sentence length (measured as word count) on sentence specificity across domains, restricted to the data-supported Granuscore range. As expected, we observe a negative relationship: as sentence length increases, specificity scores decrease, indicating more specific sentences.

The strength of this effect varies by domain. For Twitter, specificity decreases most rapidly with increasing length, followed by movie reviews, Yelp reviews, and news articles. This pattern reflects domain-specific length distributions, as Twitter texts are typically much shorter than the other domains, while news articles tend to be longer and more descriptive.

D QA Datasets

Response Generation and Evaluation For answer generation, we use a maximum output length of 512 tokens for standard and 2,048 tokens for reasoning-based generation. We retain only responses that terminate before reaching the respective token limits, ensuring that all evaluated outputs are complete and not truncated.

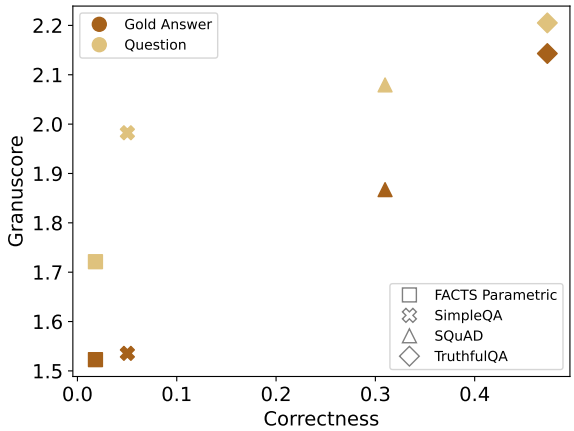

Figure 6: Granuscore versus answer correctness across QA datasets for QWEN-8B with reasoning.

Models are instructed to produce answers of at most five sentences using the following prompt:

User Prompt: Relation Satisfaction

```
Answer the following query in at most five
complete sentences: <query>
```

Model responses are evaluated using `gpt-4.1-nano-2025-04-14` as an LLM-based judge, following the prompt template introduced in SIMPLEQA (Wei et al., 2024).

Granuscore of Question and Gold Answers In Figure 6, we relate model correctness of QWEN-8B with reasoning to the Granuscore of questions and gold answers across QA datasets. Model correctness varies strongly across datasets, with accuracies of 1.81% on FACTS PARAMETRIC, 5.03% on SIMPLEQA, 30.98% on SQUAD, and 47.32% on TRUTHFULQA.

Overall, the observed patterns mirror those of the non-reasoning setting. Datasets characterized by lower-granularity questions and answers are associated with substantially lower model correctness, whereas datasets with higher granularity tend to yield improved performance.

Granuscore Across Response Outcomes Table 12 reports mean Granuscore values for questions, gold answers, reasoning traces, and model outputs for QWEN-8B with reasoning, stratified by response outcome. As in the non-reasoning setting, Granuscore differs significantly across out-

	Qwen-8B				Qwen-8B Reasoning			
Dataset	Correct	Wrong	Not Att.	Corr. (Att.)	Correct	Wrong	Not Att.	Corr. (Att.)
FACTS Parametric	2.20	92.55	5.25	2.32	1.81	78.61	19.58	2.26
SimpleQA	5.17	93.84	0.99	5.22	5.03	91.09	3.88	5.23
SQuAD	33.44	63.37	3.19	34.54	30.98	58.37	10.65	34.67
TruthfulQA	47.44	46.32	6.24	50.60	47.32	38.95	13.73	54.85

Table 11: Model outcomes (%) across QA datasets for QWEN-8B with and without reasoning. *Correct*, *Incorrect*, and *Not Att.* are proportions over all samples. *Acc. (Att.)* denotes accuracy computed over attempted answers only.

Type	Correct	Wrong	Not Att.
Question	2.09	2.02	2.07
Gold Answer	1.95	1.71	1.82
Reasoning	2.12	2.08	2.15
Answer	2.12	2.09	2.17
Gap	0.19	0.40	0.37

Table 12: Mean Granuscore values by response outcome for QWEN-8B with reasoning. Gap denotes the difference between output and gold-answer granularity.

comes (Mann-Whitney U, $p \leq 1.76 \times 10^{-7}$), except for question granularity between correct and not-attempted responses.

Wrong answers are associated with the lowest output granularity (2.09), followed by correct answers (2.12), while not-attempted responses exhibit the highest granularity (2.17), reflecting explicit abstentions. Granuscore for intermediate reasoning traces follows a similar trend and thus does not yield additional insights beyond those observed for final answers.

Finally, the granularity gap remains substantially larger for incorrect and not-attempted responses. A logistic regression using granularity gap as the sole predictor yields an average AUC of 0.66 ($\pm$ 0.004), closely matching the non-reasoning results.

Comparison between Non-Reasoning and Reasoning Overall differences in correctness and granularity between non-reasoning and reasoning variants are negligible. The primary change concerns the proportion of non-attempted responses. As shown in Table 11, the percentage of non-attempted answers increases consistently when reasoning is enabled, ranging from +2.89 percentage points on SIMPLEQA to +14.33 percentage points on FACTS PARAMETRIC.

This shift indicates that the reasoning variant adopts a more conservative answering strategy, abstaining more frequently from producing an answer. As a consequence, accuracy over attempted answers generally increases, reflecting higher correctness among the remaining responses. An exception is FACTS PARAMETRIC, where accuracy over attempted answers slightly decreases by 0.06 percentage points.