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 granularity, from fine-grained references to broad descriptions. While granularity is fundamental to human communication, existing measures mostly capture surface detail or sentence specificity. We introduce **Granuscore**, a reference-free measure of granularity that leverages structural properties of a hierarchical embedding space. Granuscore reliably recovers hierarchical orderings on the GRANOLA-EQ dataset and captures expected differences in granularity across discourse contexts. Across domains, we further show that Granuscore explains non-linear variation in sentence specificity beyond sentence length. Finally, we apply Granuscore to four question-answering benchmarks and analyze how granularity differs for questions, gold answers, and model outputs across response outcomes. The analysis reveals consistent differences in model behavior and provides a principled lens for characterizing the difficulty of QA datasets. Together, the results position Granuscore as a scalable, broadly applicable tool for analyzing granularity in text.

	Example
29.32	Tony Hawk was born in San Diego.
40.34	Tony Hawk was born in California.
54.77	Tony Hawk was born in the United States.
74.53	A skateboarder was born in the United States.
81.60	A sportsman was born in the United States.

Figure 1: Sentences with referential units varying in granularity. Units that differ across sentences are underlined. Replacing fine-grained terms with coarser alternatives increases sentence granularity: lower Granuscores indicate finer expressions.

1 Introduction

Natural language varies not only in *what* information is conveyed, but also in *how coarsely or finely* that information is expressed. Consider the sentences in Figure 1. A speaker may refer to a person as *Tony Hawk*, *a skateboarder*, or *a sportsman*, and may locate an event in *San Diego*, *California*, or *the United States*. These alternatives preserve the underlying fact while referring to it at different levels. We refer to this dimension as *granularity*: the level of abstraction at which entities or events are represented in language (Mulkar-Mehta et al., 2011; Rosch et al., 1976; Hobbs, 1985).

Granularity is not incidental: speakers adapt the level of abstraction of their descriptions depending on conversational context and task requirements (Mulkar-Mehta et al., 2011; Hobbs, 1985). When uncertain, speakers often prefer coarser fact descriptions that remain informative without overcommitting. Conversely, when common ground is established, more fine-grained references become appropriate (Yona et al., 2024). Granularity should therefore be understood as a deliberate strategy that balances reliability and audience expectations.

Prior work suggests that linguistic granularity affects how information is perceived and used. In dialogue systems, too fine-grained or coarse responses can reduce user satisfaction (Adiwardana et al., 2020; Thoppilan et al., 2022). Similarly, in simplified language settings, controlling granularity is important for accessibility and comprehension as it reduces the cognitive load (OECD, 2024; Anschütz et al., 2025). However, studying such effects systematically is difficult because existing approaches do not provide a scalable, reference-free measure of granularity at the sentence level.

Our contributions are as follows:

- We introduce **Granuscore**, a reference-free measure of granularity that exploits structural properties of a hierarchical embedding space.
- We validate Granuscore both empirically and conceptually. It reliably recovers human-annotated orderings on GRANOLA-EQ (Yona

et al., 2024) and captures expected granularity differences across discourse contexts.

- We show that across domains Granuscore explains non-linear variation in sentence specificity beyond sentence length.
- We demonstrate the practical relevance of Granuscore for question answering. Evaluating six language models on four QA benchmarks, we identify consistent differences in granularity between questions, gold answers, and model outputs across response outcomes. These patterns provide a principled lens for characterizing QA dataset difficulty and analyzing model behavior.
- We release Granuscore as a pip package to ensure reproducibility and enable its usage for further research or production.

2 Background and Related Work

Granularity Mulkar-Mehta et al. (2011) describe granularity in natural language as shifts between coarse and fine descriptions, where higher-level representations abstract from more detailed components. Related perspectives appear in cognitive science, where concepts are organized at different levels of abstraction within taxonomies (Rosch et al., 1976). Further, foundational work by Hobbs (1985) argues that intelligent reasoning requires representing the world at multiple levels of granularity and switching between them as needed, allowing complex phenomena to be modeled through simpler abstractions.

A related property is *term specificity*, which refers to identifying index terms distinguishing one class of documents from others. In particular, Kim (2006) describe *hierarchical specificity* as a term's position within a generic–specific hierarchy, where narrower terms correspond to more specific concepts, matching the notion of granularity.

We capture these ideas using structural properties of a hierarchical embedding space. Unlike approaches relying on manually constructed hierarchies, this enables estimating granularity without being restricted to predefined vocabularies.

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).

Although finer-grained references often increase sentence specificity, granularity and sentence specificity capture different properties. Sentence specificity reflects the amount of descriptive information conveyed by a sentence, whereas granularity describes the level at which referential expressions occur within a semantic hierarchy. A sentence can therefore become more specific by adding descriptive details without changing the granularity of its referents. For example, *"The skateboarder won the competition"* becomes more specific in *"The skateboarder won the competition and set a new record."*. The referents remain at the same granularity level, but the sentence conveys more information.

Granularity Evaluation While granularity has been implicitly discussed in work on specificity, informativeness, and semantic hierarchies (Thoppilan et al., 2022; Adiwardana et al., 2020; Ko et al., 2019; Li et al., 2016), existing automatic evaluations typically rely on taxonomy depth (e.g., WordNet hypernym levels (Miller, 1994) or hierarchical relations in knowledge graphs such as Wikidata (Vrandečić and Krötzsch, 2014; Huang et al., 2023)). However, these approaches require entities to exist in the underlying taxonomy and therefore provide limited coverage for free-form text. In contrast, embedding-based approaches can operate directly on arbitrary text.

Huang et al. (2023) propose an automatic benchmark for measuring specificity using transitive relations derived from Wikidata. However, the induced orderings can yield unintuitive comparisons, for example, ranking *Mexico* as more granular than *Colombia*, or *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 dataset to train Granuscore and extend their analysis by applying granularity estimation to a broader set of QA datasets, studying how granularity relates to model outputs, correctness, and dataset difficulty.

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

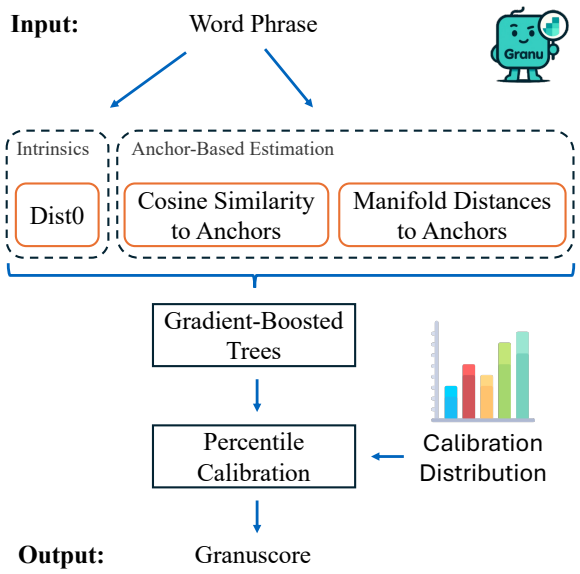

Figure 2: **Granuscore pipeline**: extraction of hierarchical depth (*Dist0*) and comparison to anchor entities, followed by gradient-boosted trees and percentile calibration to produce a scalar granularity score.

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, scalable signal that measures granularity on an absolute and interpretable scale.

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 (HiT) proposed by Chen et al. (2024), who train transformer encoders to represent hierarchical structure in a hyperbolic embedding space modeled as a Poincaré ball. In this geometry, hierarchical relations are represented by radial distance from the origin: more specific concepts lie farther from the center, while more general concepts lie closer. We denote this radial distance as *Dist0*, which captures hierarchical depth and serves as a primary signal for granularity. We use the variant trained on the WordNet hierarchy, as WordNet (Miller, 1994) provides a broad-coverage commonsense structure.

While *Dist0* captures global hierarchical position, additional signals can be obtained by relating to other entities in the space. Therefore, we compare the input embedding against a set of anchor entities and derive features from the resulting pairwise relations. In our default configuration, we use 999 randomly sampled fixed anchors, which performed best in our ablation (Appendix G). Alternative strategies are described in Section 3.3.

Figure 2 illustrates the resulting pipeline. Given an input word or phrase, the model first obtains a hierarchical embedding and extracts *Dist0*. It then computes pairwise similarity and distance features to the anchor entities using a Wikidata-derived embedding index. To map these features to a scalar granularity score, we train gradient-boosted decision trees using LIGHTGBM (Ke et al., 2017). The model operates directly on the raw similarity and distance values, allowing it to capture fine-grained interaction patterns that would be lost under pre-aggregation. Details on the training procedure and model hyperparameters are provided in Appendix E. Because the resulting raw scores depend on the annotations of GRANOLA-EQ, we convert them to percentile scores using a fixed calibration distribution. We choose the WordNet noun set (approximately 119k concepts), which was also used to train the HiT model, providing an annotator-independent alignment. Appendix F.3 shows how annotation levels map to raw and percentile scores.

3.1 Dataset

To train the LIGHTGBM model, we use GRANOLA-EQ (Yona et al., 2024), an extension of the ENTITYQUESTIONS dataset (Sciavolino et al., 2021). Each dataset entry consists of a question and a set of answers referring to the same underlying *reference entity* at different levels of granularity. We refer to the ordered list of such answers as an *answer hierarchy*, and to the individual answers as *granularity realizations*.

During preprocessing, we remove entries with more than four granularity realizations (fewer than 1.2% of the data) as these typically reflect inconsistencies introduced during generation. The resulting dataset contains, on average, approximately three realizations per question (2% with one, 22% with two, 62% with three, and 14% with four).

Since GRANOLA-EQ was generated by prompting an LLM to list increasingly coarse answers, the number of realizations per question

varies, and no fixed hierarchical structure is enforced (e.g., *city* → *state* → *country*). The LLM implicitly determines the resolution of the answer hierarchy it considers appropriate for a given question. To obtain comparable training targets, we normalize the answer levels to a continuous scale from 1 (most fine-grained) to 4 (most coarse-grained); for example, a hierarchy with three answers is mapped to levels $1, 2.5, 4$.

Due to the construction of GRANOLA-EQ, the same entity may appear at different granularity levels across dataset entries depending on the question context (e.g., *England* appears 487 times with a mean granularity of 3.25 and variance 0.44). We retain these variations, allowing the model to learn from multiple granularity annotations of the same realization and encouraging generalization.

Finally, to prevent data leakage, we enforce that no granularity realization appears in more than one split. The final dataset consists of 6,702 training samples and 1,220 samples each for development and test. For dataset details, see Appendix F.

3.2 Extension to Multi-Word Inputs

Because Granuscore is defined for individual referential units, we extend it to sentences and longer text spans by decomposing the text. We use a SPACY-based splitter (Honnibal et al., 2020). Noun phrases are kept intact, while stop words and non-informative symbols are removed. This preserves referential expressions that convey granularity while avoiding fragmentation of multi-word concepts. If no referential units can be identified (e.g., the input consists solely of stop words), we assign a Granuscore of 100, corresponding to the coarsest granularity score.

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 the person and the location 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 using a two-step aggregation. First, we compute a sentence-level Granuscore by averaging the scores of the extracted referential units within each sentence. We then aggregate across sentences by taking the mean of the bottom 80% of sentence Granuscores, which reduces the influence of unusually high scores. We evaluate a range of alternative aggregation strategies and compare them in Appendix H. Based on this ablation, we adopt this aggregation as the default, as it shows the strongest performance.

3.3 Methods

To evaluate the effectiveness of Granuscore, we compare it against several baselines and variants that estimate granularity using lexical, hierarchical, or embedding-based signals:

- **Word Count:** Number of words in the text. Negative word count so that higher scores correspond to coarser concepts.
- **WordNet Hierarchy:** Average depth of mapped WordNet synsets; deeper nodes correspond to finer concepts.
- **GPT-4.1 mini:** Few-shot prompting to estimate granularity (Appendix D).
- **HiT Dist0:** radial distance *dist0* only.
- **Nearest Neighbors (NN):** top-k cosine-similar entities as anchors.
- **Random:** k dynamically sampled anchors.
- **Random Anchors:** fixed k random anchors.
- **Radial Anchors:** fixed set of k anchors sampled across HiT *dist0* distance bins.

For Nearest Neighbors, Random, and Random Anchors, we evaluate both HiT and MiniLM embeddings. Radial Anchors are only defined for HiT, as they rely on the HiT *Dist0* radial structure. Additional details on the methods are provided in Appendix C. Exact versions of all LLMs used throughout the paper are listed in Appendix B.

3.4 Evaluation Approaches

We evaluate Granuscore across three complementary settings that test different aspects. First, we measure how well methods recover controlled granularity orderings. Second, we examine whether granularity scores capture differences across discourse contexts. Finally, we analyze how Granuscore relates to sentence specificity.

GRANOLA-EQ We first evaluate all methods on the test set of GRANOLA-EQ. We use *Pairwise Accuracy*, defined as the percentage of correctly ordered granularity realizations. For realization R_i and R_j with gold ordering $R_{i,\text{gold}} < R_{j,\text{gold}}$, a prediction is considered correct if $R_{i,\text{pred}} < R_{j,\text{pred}}$.

Pairs with identical gold granularity levels are excluded because the resolution of the GRANOLA-EQ annotations does not define a unique ordering.

We compute this metric in two settings. In the *global* setting, pairwise accuracy is computed across all dataset entries, measuring the ability of a method to assign consistent granularity scores to unrelated entities. This task is particularly challenging because entities may belong to different semantic dimensions and must be placed on a shared granularity scale. For example, a model must compare realizations such as *skateboarder* and *California*, originating from different hierarchies (e.g., *Tony Hawk* → *American skateboarder* → *skateboarder* → *sportsman* and *San Diego* → *California* → *United States* → *America*).

In the *intra-entry* setting, pairwise accuracy is computed within the hierarchy of each entry. This evaluates how well methods recover the local ordering of semantically related realizations.

Discourse Contexts Large-scale annotations of granularity for longer text are difficult to obtain, so we instead rely on naturally occurring discourse differences as an unsupervised proxy. Scientific papers provide a suitable testbed, as their standardized section structure reflects distinct discourse functions: Introduction sections typically describe the broader research context using more coarse-grained references, whereas Related Work sections contain more fine-grained references to specific prior methods, datasets, and papers.

We apply Granuscore to scientific articles from the S2ORC corpus (Lo et al., 2020). We sample 1,000 papers and compare the granularity of Introduction and Related Work sections. Further details on the sampling and filtering procedure are provided in Appendix J.

Sentence Specificity Finally, we examine how Granuscore relates to sentence specificity. We analyze human-annotated datasets from Ko et al. (2019) and Li et al. (2016). The former covers movie reviews, tweets, and Yelp reviews, while the latter contains sentences from news articles.

Sentence length (word count) is a strong baseline predictor with Spearman correlations of 0.45 (Twitter), 0.58 (movie reviews), 0.68 (Yelp), and 0.67 (news). We therefore quantify Granuscore's contribution to sentence specificity beyond sentence length by fitting Generalized Additive Models (GAMs). This allows us to isolate the contribution of each predictor via explained deviance.

Method	Global PW Acc.	Intra PW Acc.	Exact
Word Count	50.49	51.54	28.80
WordNet†	58.12	67.32	60.39
GPT-4.1 mini	76.76	81.58	58.21
HiT Dist0	80.82	87.86	73.50
MiniLM NN	67.40	69.50	48.97
MiniLM Random	67.55	70.57	48.80
MiniLM RandomAnch	67.45	71.15	49.15
HiT NN	81.79	86.47	70.17
HiT Random	80.80	84.74	69.32
HiT RadialAnch	82.31	88.35	71.71
HiT RandomAnch	83.00	88.15	72.48
HiT Dist0 + NN	82.86	87.02	71.54
HiT Dist0 + Random	82.01	86.82	70.85
HiT Dist0 + RadialAnch	*83.22*	*88.83*	*73.85*
HiT Dist0 + RandomAnch	**83.76**	**89.03**	**74.36**

Table 1: Comparison of methods on the GRANOLA-EQ test set. We report Global Pairwise Accuracy (PW Acc.), Intra-entry Pairwise Accuracy (Intra PW Acc.), and Exact Ordering Accuracy (Exact). Bold indicates the best result and italic indicates the second-best result across methods. †: WordNet could only derive a granularity level for 17.61% of the realizations.

4 Results

4.1 GRANOLA-EQ

Table 1 reports the performance of all methods on GRANOLA-EQ using Global and Intra-entry Pairwise Accuracy and Exact Ordering Accuracy. Additional metrics are provided in Appendix F.4.

Across all methods, intra-entry pairwise accuracy consistently exceeds global pairwise accuracy, indicating that ranking realizations within the same semantic hierarchy is easier than assigning consistent scores across unrelated entities. Exact ordering accuracy is consistently lower, reflecting the greater difficulty of recovering the full hierarchy rather than individual pairwise relations.

The radial depth signal already provides a strong baseline: HiT Dist0 achieves 80.82% global pairwise accuracy and 87.86% intra-entry accuracy.

In comparison, anchor-based HiT methods generally outperform it in global pairwise accuracy, with HiT Random Anchors outperforming HiT Dist0 by 2.18 percentage points. Notably, anchors sampled across the embedding space outperform nearest neighbors, suggesting that global structure is more informative for estimating granularity than local similarity.

Combining HiT Dist0 with Random Anchors achieves the highest scores across all metrics. Compared to HiT Dist0, it improves by +2.94 (global)

and +1.17 (intra-entry) points, and over HiT Random Anchors by +0.76 and +0.88 (bootstrap resampling, $N = 20{,}000$, $p < 0.002$ global; $p < 0.05$ intra-entry).

In contrast, all MiniLM-based variants perform substantially worse (best global: 67.55%) and yield nearly identical scores regardless of the anchor selection strategy. This suggests that the underlying embedding geometry provides a weaker signal for granularity than HiT.

Among additional baselines, the WordNet hierarchy achieves 58.12% global pairwise accuracy despite covering only 17.61% of realizations. This shows that lexical taxonomies contain meaningful signals for granularity when applicable, but their use is limited by coverage. GPT-4.1 mini performs competitively in pairwise ordering but shows lower exact ordering accuracy. Here, manual inspection indicates that the model frequently assigns identical granularity levels to multiple realizations, thereby reducing its ability to recover the full hierarchy. Finally, the Word Count baseline performs close to random (50%), confirming that granularity is not reflected in sentence length.

Overall, these results show that granularity is best captured by combining hierarchical depth and anchor comparisons in the embedding space.

4.2 Granuscore Across Paper Sections

Beyond the gold-labeled setting, we evaluate whether Granuscore captures differences across discourse contexts. Across the sampled papers, 68.71% of paired comparisons exhibit a higher Granuscore for the Introduction than for the Related Work section, indicating that Introduction sections tend to use more coarse-grained language. This difference is highly significant (paired t-test: $p \leq 5.49 \times 10^{-37}$; Wilcoxon signed-rank test: $p \leq 5.73 \times 10^{-39}$) with a moderate paired effect size ($d_z = 0.42$; Cohen, 2013). Consistent with this ordering, Introduction sections also have a higher average Granuscore (75.29 ± 4.94) than Related Work sections (72.43 ± 5.54).

This pattern aligns with the rhetorical roles of these sections.

4.3 Correlation to Sentence Specificity

Finally, we analyze how Granuscore relates to sentence specificity. Table 2 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),

Domain	Expl. Dev. (Length)	Expl. Dev. (Len+Gran)	Δ
movie	0.38	0.46	+0.08
twitter	0.24	0.36	+0.12
yelp	0.52	0.56	+0.04
news	0.45	0.55	+0.10

Table 2: Explained deviance of generalized additive models (GAMs) predicting sentence specificity. All smooth terms are significant ($p < 2.41 \times 10^{-10}$).

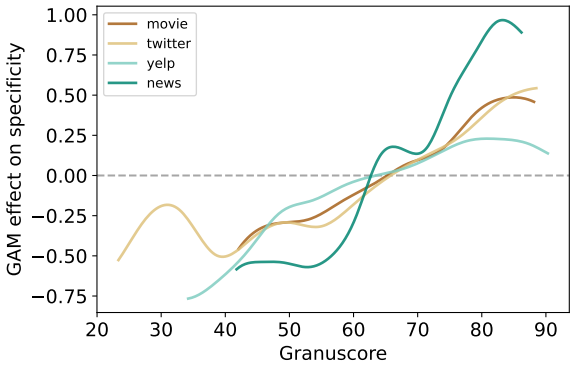

Figure 3: Effect of Granuscore on sentence specificity across domains. Lower specificity scores correspond to more specific sentences. The plotted range is restricted to the 1st–99th percentiles of Granuscore to avoid sparse-support regions.

corresponding to relative improvements of 7–50%. In Figure 3, we show the estimated effect of Granuscore on sentence specificity. Across all domains, the relationship is non-linear. Lower Granuscore negatively affects the specificity score, indicating an association with more specific sentences. As Granuscore increases, the magnitude of this negative effect decreases. The effect crosses zero between values of roughly 63–66, after which higher scores are associated with less specific sentences. For completeness, the effect of sentence length is shown in Appendix I.

Overall, these results show that Granuscore captures a systematic component of sentence specificity while remaining distinct from it. Although granularity alone does not determine specificity, incorporating Granuscore consistently improves specificity prediction across domains. This pattern aligns with the intuition that references to fine-grained entities tend to appear in more specific sentences, whereas coarse-grained concepts are more common in less specific ones.

5 Applying Granuscore to QA Datasets

We apply Granuscore to several widely used QA datasets to investigate how granularity affects

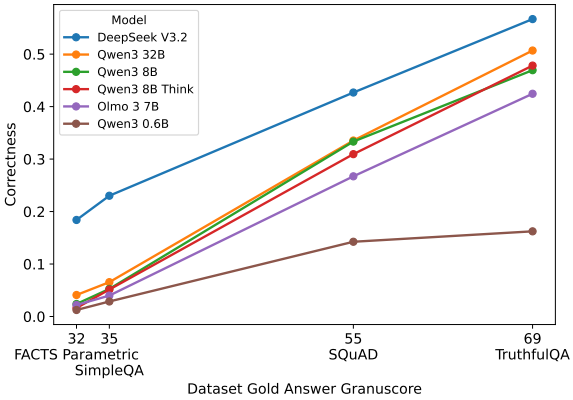

Figure 4: Relationship between dataset-level gold answer Granuscore and model correctness across QA benchmarks. Higher Granuscore datasets are associated with higher correctness across models. All pairwise differences in Granuscore between datasets are statistically significant (Mann–Whitney U, $p \leq 1.1 \times 10^{-3}$).

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.

To relate granularity to model behavior, we evaluate model correctness using QWEN3 0.6B, QWEN3-8B, and QWEN3-32B (Yang et al., 2025), OLMO 3 7B (Olmo et al., 2025), and DEEPSEEK V3.2 (DeepSeek-AI et al., 2025). These models cover a broad range of model sizes and represent well-established open-weight language models. For QWEN3-8B, we additionally evaluate both standard generation and reasoning-enabled generation ("think" mode) to compare performance with and without explicit reasoning.

Model responses are evaluated using GPT-4.1 nano as an LLM-based judge, following the prompt template introduced in SIMPLEQA (Wei et al., 2024). Further details are given in Appendix L.

Granuscore Gold Answers In Figure 4, we relate model correctness to the Granuscore of gold answers across QA datasets. For each of the models, correctness varies strongly across datasets, with mean accuracies of 4.99% on FACTS PARAMETRIC, 7.80% on SIMPLEQA, 30.24% on SQUAD, and 43.47% on TRUTHFULQA.

Datasets with lower Granuscores exhibit substantially lower accuracy, while higher Granuscore datasets are associated with improved performance. This pattern is consistent across all evaluated models, suggesting that Granuscore captures a model-independent aspect of question difficulty. Larger models, such as DEEPSEEK V3.2, achieve consistently higher correctness across datasets, indicating greater knowledge coverage, but follow the same overall trend. In contrast, the smallest model, QWEN3 0.6B, exhibits a weaker slope, likely reflecting general capacity limitations. We observe a similar trend when analyzing the Granuscore of questions (Appendix K). In contrast, replacing Granuscore with answer or question length does not yield a comparable relationship (Appendix K), indicating that the observed effect is not explained by superficial length differences.

Type	Correct	Wrong	Not Att.
Question	70.1 ± 1.5	65.4 ± 0.3	67.2 ± 0.6
Gold Answer	59.4 ± 4.1	45.8 ± 0.7	48.7 ± 2.7
Answer	69.6 ± 1.7	66.0 ± 1.4	72.5 ± 1.3

Table 3: Granuscore (mean $\pm$ std. across models) by response outcome (Correct, Wrong, and Not Attempted). Granuscore distributions differ significantly across outcomes (Mann–Whitney U, $p \leq 2.42 \times 10^{-13}$).

Granuscore Across Response Outcomes In Table 3, we report mean Granuscore values for questions, gold answers, and model outputs, stratified by response outcome (correct, incorrect, and not-attempted). Model outputs associated with wrong answers exhibit the lowest average Granuscore (66.0), followed by correct answers (69.6), while not-attempted responses show the highest output granularity (72.5). The latter is expected, as abstentions typically consist of general statements indicating an 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 0.62 ($\pm$ 0.005), indicating a moderate, stable association between granularity mismatch and response failure.

6 Discussion

Comparison of Methods Our results on GRANOLA-EQ highlight the importance of hierarchical structure for estimating granularity.

Methods based on HiT consistently outperform approaches using standard sentence embeddings, indicating that granularity is closely tied to hierarchical relations rather than surface similarity. The radial depth signal (Dist0) already provides a strong baseline as it directly reflects hierarchical depth in the embedding space. However, anchor-based comparisons further improve performance, particularly for global ordering. Comparing entities to anchors across the embedding space provides additional relational context, enabling a more reliable and stable estimation of granularity across unrelated hierarchies. Hence, the best results are achieved when combining both signals.

This challenge of comparing unrelated hierarchies is also reflected in the evaluation metrics. We observe a consistent gap between intra-entry and global accuracy: intra-entry comparisons operate within a shared semantic hierarchy (e.g., city → state → country), whereas global comparisons require ordering entities from unrelated hierarchies on a common scale. Despite this difficulty, Granuscore provides a strong signal for estimating general granularity across heterogeneous hierarchies.

Correlation to Sentence Specificity Further, we show that Granuscore explains non-linear variation in 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 supports the robustness of Granuscore as a general granularity measure.

Granularity and QA Performance Our QA analysis case study reveals consistent patterns linking granularity and response outcomes.

First, across all models and 16,689 QA-samples, we observe clear differences in the Granuscore across response outcomes. Questions and gold answers associated with incorrect or not-attempted responses exhibit significantly lower Granuscore values than those associated with correct responses. The effect is particularly pronounced for gold answers, while the difference in question granularity is present but weaker. These findings suggest that Granuscore may serve as a proxy for the difficulty of question–answer pairs and could be incorporated as a signal for deciding when to rely on a model's internal knowledge versus external tools.

At the dataset level, we observe a complementary trend: datasets with lower granularity, both for gold answers and questions, are substantially harder for models. This suggests that granularity provides a useful lens for characterizing differences in QA-difficulty that are not explained by superficial properties such as answer or question length.

Finally, we observe that incorrect responses tend to exhibit lower output granularity than correct ones. In these cases, models often remain at the level of detail implied by the question rather than adapt their responses to a more appropriate granularity based on their confidence in the answer. This aligns with findings that models struggle to adjust answer granularity (Yona et al., 2024) and Kalai et al. (2025) arguing that benchmark evaluations incentivize models to guess overly specific answers.

Future Directions A natural next step is to use Granuscore as a training signal for language models. Prior work has shown that optimizing for properties such as informativeness and interestingness can improve response quality and user engagement (Adiwardana et al., 2020; Thoppilan et al., 2022; Onozeki and Inaba, 2025). Similarly, Granuscore could encourage models to generate responses at appropriate levels of granularity. In particular, models could learn to align output granularity with their confidence: when uncertain about fine-grained details, they may respond at a coarser but reliable level (e.g., a broader category or time period). Such behavior mirrors human communication and could help reduce overly fine-grained incorrect answers while preserving informative responses.

Beyond 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 that quantifies the granularity expressed in text using a hierarchical embedding space. Granuscore reliably recovers granularity orderings on the controlled GRANOLA-EQ benchmark, aligns with expected differences across scientific paper sections, and captures non-linear variation in sentence specificity beyond sentence length.

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

Limitations

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 in 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.

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].

Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Routledge, New York.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [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.

Jerry R. Hobbs. 1985. Granularity. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'85, pages 432–435, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. *arXiv preprint*. ArXiv:2509.04664 [cs].

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.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Rutu Mulkar-Mehta, Jerry Hobbs, and Eduard Hovy. 2011. Granularity in Natural Language Discourse. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

OECD. 2024. Do Adults Have the Skills They Need to Thrive in a Changing World?: Survey of Adult Skills 2023. *OECD Skills Studies*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *arXiv preprint*. ArXiv:2512.13961 [cs].

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.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

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.

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].

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

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.

	Bike Example
51.60	He fixed his CUBE road bike using a rusty wrench.
57.52	He fixed his road bike using a rusty wrench.
67.24	He fixed his bike using a rusty wrench.
72.70	He fixed his bike using a wrench.
78.15	He fixed his bike using a tool.

Figure 5: Illustration of semantic abstraction. Starting from the specific statement *"He fixed his CUBE road bike using a rusty wrench"*, it can be generalized by abstracting the vehicle and the instrument.

	Café Example
47.20	I bought a cappuccino at the small Italian café.
55.68	I bought a warm milk drink at the small Italian café.
61.81	I bought a drink at the small Italian café.
69.94	I bought a drink at the café.
75.01	I bought a drink at the restaurant.

Figure 6: Illustration of semantic abstraction. Starting from the specific statement *"I bought a cappuccino at the small Italian café"*, it can be generalized by abstracting the drink type and the venue.

	Chair Example
45.55	He sits on his old wooden chair.
56.67	He sits on his wooden chair.
69.30	He sits on a chair.
75.16	He sits on furniture.

Figure 7: Illustration of semantic abstraction. Starting from the specific statement *"He sits on his old wooden chair"*, it can be generalized by abstracting the seating option.

A Example Sentences

We present example sentences with controlled granularity levels together with their assigned Granuscore values in Figure 5, Figure 6, and Figure 7. They illustrate how the same underlying fact can be expressed at different levels of granularity.

B Model Access

To support reproducibility, Table 4 lists all models used in this paper, including their names, exact versions, and access providers.

C Reference Construction

As the embedding space is unbounded, we approximate its structure using a finite subset of entities. We construct this proxy space from 50,000 randomly sampled Wikidata entities[1]. For each entity, we use its title as the textual representation. 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 the Random Anchor and Radial Anchor methods, anchors are sampled once in advance from the index and reused for all queries. For the Nearest Neighbor method, we retrieve the top nearest neighbors using cosine similarity for each query individually. For the Random Neighbor baseline, neighbors are sampled at random from the index for each query.

D LLM Prompt

We use the following few-shot prompt to annotate granularity levels with an LLM. The prompt includes three example semantic hierarchies comprising a total of 14 realizations together with their expected granularity levels.

User Prompt: LLM Granularity

```
You are an expert annotator for
granularity.

Your task is to assign a granularity
score to an answer using a 4-point Likert
scale. Granularity refers to how fine- /
coarse-grained an answer is.

Always assign exactly one score: 1, 2, 3,
or 4.

—
{examples}
—

Now assign a granularity score. Output
only the score.

Answer: "{answer}"

Granularity:
```

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

Name	Version	Access Provider
HiT	HiT-MiniLM-L12-WordNetNoun	Local
MiniLM	all-MiniLM-L6-v2	Local
GPT-4.1 nano	gpt-4.1-nano-2025-04-14	OpenAI Batch API
GPT-4.1 mini	gpt-4.1-mini-2025-04-14	OpenAI Batch API
Qwen3 0.6B	N/A	Local
Olmo 3	Olmo-3-7B-Instruct	Local
Qwen3 8B	N/A	Local
Qwen3 8B Think	N/A	Local
Qwen3 32B	N/A	OpenRouter
DeepSeek V3.2	N/A	OpenRouter

Table 4: Specific model versions used in our experiments. For each model we provide the exact version and the access provider.

Parameter	**Value**
Boosting type	GBDT
Objective	Regression
Evaluation metric	RMSE
Early stopping	200
Number of iterations	10,000
Learning rate	0.0257596
Number of leaves	138
Maximum depth	Unlimited
Minimum data in leaf	57
Feature fraction	0.751449
Bagging fraction	0.638041
Bagging frequency	7
Dropout rate (DART)	0.1
Maximum bins	255
Device	CPU

Table 5: LightGBM hyperparameters used for training Granuscore.

Rel.	Question Template
P112	Who founded [X]?
P127	Who owns [X]?
P131	Where is [X] located?
P159	Where is the headquarter of [X]?
P170	Who created [X]?
P175	Who performed [X]?
P176	Which company is [X] produced by?
P19	Where was [X] born?
P20	Where did [X] die?
P26	Who is [X] married to?
P264	What music label represents [X]?
P276	Where is [X] located?
P40	Who is [X]'s child?
P50	Who is the author of [X]?
P69	Where was [X] educated?
P740	Where was [X] founded?

Table 6: Question template for each relation type in the dataset.

E 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 5.

F Granola-EQ

GRANOLA-EQ covers multiple relation types, represented through masked question templates. Table 6 lists the relation categories included in the dataset together with their corresponding question templates. Table 7 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. As shown in the table, the relative proportions differ across splits. This variation arises from enforcing a strict split-by-granularity-realization rule, ensuring that the same realization does not appear in multiple splits and preventing data leakage between training and evaluation.

F.1 Distribution of Granularity Resolution

Table 8 shows the distribution of granularity resolution across the full dataset. Granularity resolution denotes the number of levels available for a given entity. The mean and variance indicate that most entities exhibit a moderate number of granularity levels, with the majority containing three levels.

F.2 Context-dependent granularity variation

Granularity is context-dependent: the same entity may occur at different levels of abstraction depending on the question. As described in Section 3.1, GRANOLA-EQ preserves these contextual variations. Table 9 illustrates this effect for the entity *England*, showing that identical entities appear at multiple normalized granularity levels across dataset entries. Granuscore therefore reflects an

Rel.	Train	Dev	Test
P20	650	36	49
P19	635	45	64
P69	600	48	30
P276	585	28	55
P159	496	76	67
P26	482	58	142
P131	452	96	61
P176	444	73	124
P50	443	64	104
P170	439	105	56
P264	398	68	95
P127	361	83	147
P40	337	83	127
P112	235	38	57
P175	79	315	35
P740	66	4	7
Total	6702	1220	1220

Table 7: Distribution of relation types across the training, development, and test splits of GRANOLA-EQ.

Gran. Resolution	Count
1	206
2	1966
3	5650
4	1320
Mean	2.88
Variance	0.44

Table 8: Distribution of granularity resolution in GRANOLA-EQ. The granularity resolution indicates the number of distinct granularity levels available for an entity.

aggregated, context-averaged view of granularity suitable for global analysis.

F.3 Granuscore Across Annotated Granularity Levels

To examine whether Granuscore meaningfully distinguishes between granularity levels, Table 10 reports the average raw and percentile scores for each normalized level in the GRANOLA-EQ test set.

Both the raw Granuscore values and their percentile equivalents increase consistently with the annotated granularity levels. The distances between adjacent levels are relatively uniform, indicating that Granuscore reflects the intended ordering of abstraction levels. This regular spacing suggests that Granuscore provides an interpretable scale of semantic granularity rather than producing small, arbitrary numerical differences. Pairwise comparisons between all levels are statistically significant (one-sided Wilcoxon signed-rank test; lower granularity $<$ higher granularity; $p \leq 6.4 \times 10^{-6}$).

Gran. Level	Count
1	2
2	13
2.5	114
3	160
4	198
Mean	3.25
Variance	0.44

Table 9: Distribution of normalized granularity levels for the realization *England* in GRANOLA-EQ.

Level	1	2	2.5	3	4
Percentile	28.54	47.33	57.27	64.12	77.29
Raw	1.55	1.99	2.19	2.41	2.79

Table 10: Average Granuscore per normalized granularity level on the GRANOLA-EQ test set using the HiT Random Anchor method. We report both the percentile score and the corresponding raw value.

F.4 Additional Metrics

Table 11 reports additional evaluation metrics on GRANOLA-EQ, including Kendall's τ, Pearson's r, and Intra-sample Kendall's τ. Kendall's τ and Pearson's r are computed across realizations from all samples, measuring how well predicted granularity scores follow the global ordering of annotated levels across unrelated entities. In contrast, Intra-sample Kendall's τ is computed within each question and then averaged, reflecting how well a method preserves the ordering of realizations within individual hierarchies.

Under these metrics, the LLM baseline (GPT-4.1 MINI) achieves the highest scores, followed by HiT Random Anchor. However, this result should be interpreted with caution. The LLM frequently assigns identical granularity levels to multiple answers, resulting in many tied comparisons. Since Kendall's τ excludes tied pairs from the computation, these ties effectively remove more difficult comparisons and leave only easier ordering decisions, artificially inflating the score.

For this reason, Kendall-based metrics can overestimate the performance of models that produce many tied predictions. To provide a more faithful evaluation of hierarchy recovery, we therefore report Pairwise Accuracy and Exact Ordering Accuracy in Section 4.1, which evaluate the ordering of all answer pairs and penalize tied predictions.

Method	Kendall τ	Pearson r	Intra Kendall τ
Word Count	4.15	1.07	7.91
WordNet†	25.54	28.15	61.15
GPT-4.1 mini	**69.84**	**74.04**	**88.24**
HiT Dist0	50.70	52.37	75.73
MiniLM NN	28.65	36.94	39.00
MiniLM Random	28.79	37.87	41.14
MiniLM RandomAnch	28.73	37.76	42.31
HiT NN	52.28	65.13	72.93
HiT Random	50.53	64.41	69.49
HiT RadialAnch	53.16	66.04	76.70
HiT RandomAnch	54.28	67.58	76.30
HiT Dist0 + NN	54.04	67.55	74.05
HiT Dist0 + Random	52.51	66.67	73.65
HiT Dist0 + RadialAnch	54.66	68.91	77.66
HiT Dist0 + RandomAnch	*55.53*	*69.69*	*78.06*

Table 11: Kendall's τ, Pearson r, and Intra-sample Kendall's τ on GRANOLA-EQ. Kendall's τ measures global ordering across all answers, while Intra-sample Kendall's τ measures ordering within individual questions. †: WordNet could only derive a hierarchy for 17.61% of the answers in the test set.

Reference Size	PW Acc.
33	83.37
66	83.16
99	83.13
333	83.30
666	82.44
999	**83.76**
1332	82.47
1665	*83.45*

Table 12: Ablation study on the number of reference anchors on GRANOLA-EQ using Random Anchors. Bold indicates the best result, and italics the second-best result across anchor sizes.

G Ablation Study on the Number of References

We analyze the effect of the number of reference anchors when using the Random Anchor method. Table 12 reports pairwise accuracy on the GRANOLA-EQ test split while varying the anchor set size.

Performance remains largely stable across different anchor sizes, indicating that the method is not highly sensitive to this parameter. The best performance is obtained with 999 anchors, which we therefore use as the default configuration in our experiments.

H Ablation Study on Aggregation Strategy

Because large-scale annotations of granularity for longer texts are difficult to obtain, we use our scientific papers testbed to determine an effective aggregation strategy. We compare several aggregation operators, including *mean*, *weighted mean*, *sum*, *min*, *max*, and *lower quantile mean* (lqm). The lower quantile mean with threshold q averages only the lowest q proportion of unit-level scores within a section (e.g., lqm(0.3) averages the lowest 30% of scores).

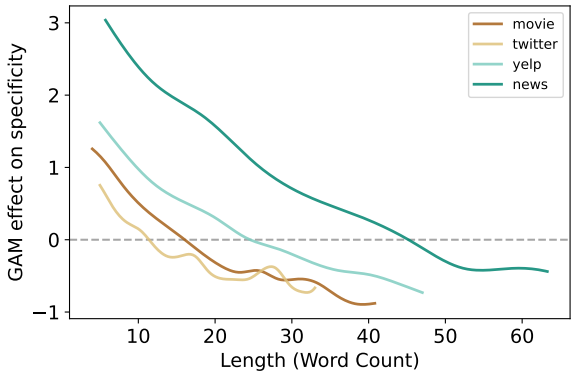

Figure 8: Effect of Length on sentence specificity across domains. Longer sentences correspond to more specific sentences. The plotted range is restricted to the 1st–99th percentiles of Granuscore to avoid sparse-support regions.

Table 13 reports ordering accuracy for the different aggregation strategies. The best performance is achieved by a two-step aggregation procedure. First, we compute a sentence-level Granuscore by averaging the scores of the extracted referential units within each sentence. We then aggregate across sentences by taking the mean of the lowest 80% of sentence-level Granuscores, which reduces the influence of unusually high values.

For completeness, we additionally evaluate all aggregation strategies in generalized additive models predicting sentence specificity from length and granularity. Table 14 reports the corresponding improvements in explained deviance. The rankings induced by ordering accuracy and explained deviance are moderately correlated (Pearson $r = 0.62$), indicating that aggregation strategies that better recover hierarchical ordering also tend to better explain sentence specificity.

I Correlation to Sentence Specificity

The sentence specificity datasets include 920 movie review sentences, 984 Twitter posts, and 845 Yelp reviews from Ko et al. (2019), as well as 573 news sentences from Li et al. (2016).

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: Twitter texts are typically much shorter than those in other domains, while news articles tend to be longer and more descriptive.

J Scientific Papers as Discourse Contexts

We compare paragraphs from the *Introduction* and *Related Work* sections. These sections are selected because their communicative roles are well defined: the Introduction typically presents the research problem and context, whereas the Related Work section situates the contribution within existing literature.

We sample the first 1,000 papers from the S2ORC corpus that contain standard *Introduction*, *Related Work*, and *Conclusion* sections, ensuring a consistent and well-structured discourse layout. Before analysis, we remove bracketed text, URLs, figure captions, and common PDF/OCR artifacts.

To obtain comparable text segments across papers, we apply a simple paragraph selection procedure. For the Introduction, we select the first paragraph containing at least ten referential units. For the Related Work section, we skip the opening paragraph, as it often functions as a brief transition, and instead select the first subsequent paragraph meeting this criterion. If no such paragraph exists, we fall back to the opening paragraph if it also satisfies the requirement. This procedure yields 978 papers for comparison.

K QA Datasets

Question Granularity Beyond the correlation between the Granuscore of gold answers and model correctness reported in section 5, we also analyze the Granuscore of the corresponding questions. Figure 9 shows a similar trend: datasets with lower question Granuscore are associated with lower correctness. This pattern is consistent across models. As for gold answers (Figure 4), the smallest model (QWEN3 0.6B) exhibits a weaker slope and partial saturation, whereas DEEPSEEK V3.2 follow the same overall trend but at higher accuracy levels.

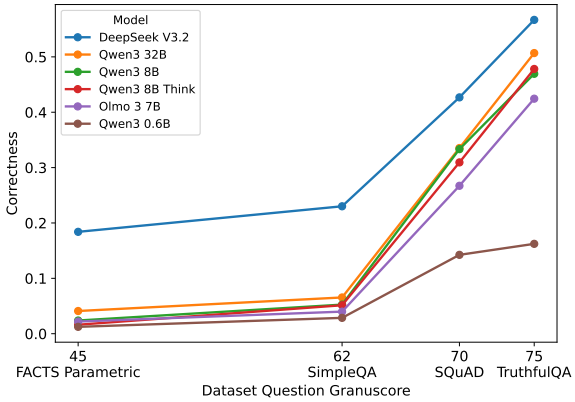

Figure 9: Relationship between dataset-level question Granuscore and model correctness across QA benchmarks. Higher Granuscore datasets are associated with higher correctness across models. All pairwise differences in Granuscore between datasets are statistically significant (Mann–Whitney U, $p \leq 3.2 \times 10^{-30}$).

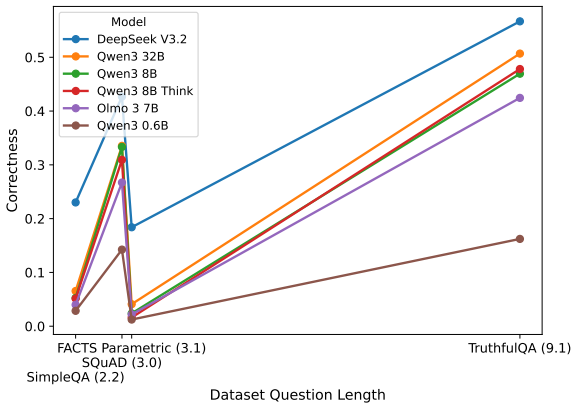

Figure 10: Relationship between dataset-level gold answer length (word count) and model correctness across QA benchmarks.

Alternative Measures For comparison, we also analyze correlations between correctness and simple length-based measures. Figure 10 and Figure 11 show the relationships between correctness and gold answer length and question length, respectively. In contrast to Granuscore, these measures do not exhibit a meaningful relationship with correctness, indicating that the observed effects are not explained by superficial length differences.

L QA Generation and Evaluation

For answer generation, we use a maximum length of 512 tokens for standard and 2,048 tokens for reasoning-based generation, with the temperature

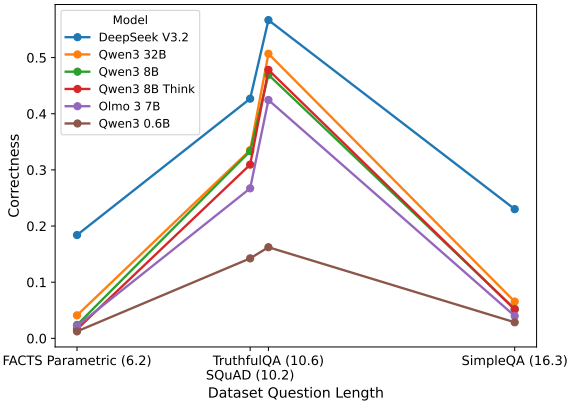

Figure 11: Relationship between dataset-level question length (word count) and model correctness across QA benchmarks.

set to 0. We instruct the model to produce answers of at most five sentences. We retain only responses that terminate before the token limit, ensuring all evaluated outputs are complete and not truncated.

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

User Prompt: Answer Generation

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

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

Aggregation	Ord. Acc.	Aggregation	Ord. Acc.	Aggregation	Ord. Acc.
sent-weighted-mean-pool-sum	58.49	sent-lqm-0.9-pool-sum	60.84	sent-min-pool-sum	60.02
sent-weighted-mean-pool-mean	66.46	sent-lqm-0.8-pool-sum	62.27	sent-min-pool-mean	66.16
sent-weighted-mean-pool-lqm-0.1	62.68	sent-lqm-0.7-pool-sum	62.68	sent-min-pool-lqm-0.1	61.76
sent-weighted-mean-pool-lqm-0.3	66.87	sent-lqm-0.9-pool-mean	*68.10*	sent-min-pool-lqm-0.3	64.01
sent-weighted-mean-pool-lqm-0.5	67.28	sent-lqm-0.8-pool-mean	**68.71**	sent-min-pool-lqm-0.5	65.03
sent-weighted-mean-pool-min	61.55	sent-lqm-0.7-pool-mean	67.59	sent-min-pool-min	59.82
sent-weighted-mean-pool-max	58.79	sent-lqm-0.9-pool-lqm-0.1	63.80	sent-min-pool-max	60.53
sent-sum-pool-sum	47.55	sent-lqm-0.9-pool-lqm-0.3	66.16	sent-max-pool-sum	52.66
sent-sum-pool-mean	45.19	sent-lqm-0.9-pool-lqm-0.5	66.87	sent-max-pool-mean	54.19
sent-sum-pool-lqm-0.1	48.67	sent-lqm-0.8-pool-lqm-0.1	63.80	sent-max-pool-lqm-0.1	53.58
sent-sum-pool-lqm-0.3	47.03	sent-lqm-0.8-pool-lqm-0.3	67.59	sent-max-pool-lqm-0.3	54.91
sent-sum-pool-lqm-0.5	45.71	sent-lqm-0.8-pool-lqm-0.5	67.89	sent-max-pool-lqm-0.5	56.85
sent-sum-pool-min	48.67	sent-lqm-0.7-pool-lqm-0.1	63.29	sent-max-pool-min	52.56
sent-sum-pool-max	44.38	sent-lqm-0.7-pool-lqm-0.3	67.38	sent-max-pool-max	48.47
sent-mean-pool-sum	60.53	sent-lqm-0.7-pool-lqm-0.5	67.59	doc-pool-sum	47.55
sent-mean-pool-mean	67.48	sent-lqm-0.9-pool-min	62.27	doc-pool-mean	66.36
sent-mean-pool-lqm-0.1	62.88	sent-lqm-0.8-pool-min	62.07	doc-pool-lqm-0.1	64.83
sent-mean-pool-lqm-0.3	66.05	sent-lqm-0.7-pool-min	62.07	doc-pool-lqm-0.3	65.64
sent-mean-pool-lqm-0.5	66.16	sent-lqm-0.9-pool-max	59.41	doc-pool-lqm-0.5	66.05
sent-mean-pool-min	62.07	sent-lqm-0.8-pool-max	60.33	doc-pool-min	59.92
sent-mean-pool-max	59.10	sent-lqm-0.7-pool-max	59.82	doc-pool-max	47.75

Table 13: Accuracy of section ordering (Introduction > Related Work) under different aggregation strategies. Aggregation names follow the pattern `scope-aggregation-pool`. `scope` indicates whether aggregation is performed at the document level (`doc`) or sentence level (`sent`). For sentence-level strategies, the first operator aggregates scores across sentences (e.g., `sent-mean`). The `pool` operator specifies how Granuscores of referential units within a sentence are combined (e.g., `sent-mean-pool-sum` first sums Granuscores within each sentence and then averages across sentences). Bold and italics denote the best and second-best results, respectively.

Method	Δ Expl. Dev.	Method	Δ Expl. Dev.	Method	Δ Expl. Dev.
sent-sum-pool-sum	3.67	sent-lqm-0.9-pool-lqm-0.3	9.20	sent-max-pool-sum	2.74
sent-sum-pool-mean	6.75	sent-lqm-0.9-pool-lqm-0.5	9.20	sent-max-pool-mean	7.09
sent-sum-pool-lqm-0.1	6.64	sent-lqm-0.8-pool-lqm-0.1	9.20	sent-max-pool-lqm-0.1	7.81
sent-sum-pool-lqm-0.3	7.27	sent-lqm-0.8-pool-lqm-0.3	9.24	sent-max-pool-lqm-0.3	7.68
sent-sum-pool-lqm-0.5	7.12	sent-lqm-0.8-pool-lqm-0.5	9.24	sent-max-pool-lqm-0.5	7.71
sent-sum-pool-min	6.10	sent-lqm-0.7-pool-lqm-0.1	9.23	sent-max-pool-min	7.28
sent-sum-pool-max	1.51	sent-lqm-0.7-pool-lqm-0.3	9.35	sent-max-pool-max	2.00
sent-mean-pool-sum	2.55	sent-lqm-0.7-pool-lqm-0.5	9.29	doc-pool-sum	3.49
sent-mean-pool-mean	8.34	sent-lqm-0.9-pool-min	8.76	doc-pool-mean	8.56
sent-mean-pool-lqm-0.1	9.19	sent-lqm-0.8-pool-min	8.77	doc-pool-lqm-0.1	*9.84*
sent-mean-pool-lqm-0.3	9.20	sent-lqm-0.7-pool-min	8.78	doc-pool-lqm-0.3	9.52
sent-mean-pool-lqm-0.5	9.20	sent-lqm-0.9-pool-max	1.87	doc-pool-lqm-0.5	9.48
sent-mean-pool-min	8.76	sent-lqm-0.8-pool-max	1.86	doc-pool-min	9.51
sent-mean-pool-max	1.87	sent-lqm-0.7-pool-max	1.85	doc-pool-max	1.88
sent-lqm-0.9-pool-sum	2.55	sent-min-pool-sum	1.95	sent-weighted-mean-pool-sum	2.65
sent-lqm-0.8-pool-sum	2.56	sent-min-pool-mean	8.07	sent-weighted-mean-pool-mean	8.62
sent-lqm-0.7-pool-sum	2.50	sent-min-pool-lqm-0.1	**9.98**	sent-weighted-mean-pool-lqm-0.1	9.14
sent-lqm-0.9-pool-mean	8.34	sent-min-pool-lqm-0.3	9.42	sent-weighted-mean-pool-lqm-0.3	9.41
sent-lqm-0.8-pool-mean	8.33	sent-min-pool-lqm-0.5	9.18	sent-weighted-mean-pool-lqm-0.5	9.54
sent-lqm-0.7-pool-mean	8.33	sent-min-pool-min	9.53	sent-weighted-mean-pool-min	8.69
sent-lqm-0.9-pool-lqm-0.1	9.19	sent-min-pool-max	1.84	sent-weighted-mean-pool-max	1.90

Table 14: Ablation over aggregation strategies measured by the improvement in explained deviance (Δ Expl. Dev., $\times 100$) relative to a length-only baseline for sentence specificity. Aggregation names follow the pattern `scope-aggregation-pool`. `scope` indicates whether aggregation is performed at the document level (`doc`) or sentence level (`sent`). For sentence-level strategies, the first operator aggregates scores across sentences, while `pool` specifies how Granuscores of referential units within each sentence are combined. For example, `sent-mean-pool-sum` first sums Granuscores within each sentence and then averages across sentences. Bold and italics denote the best and second-best results, respectively.