Narrowing the Knowledge Evaluation Gap: Open-Domain Question Answering with Multi-Granularity Answers

Gal Yona
Google Research
galyona@google.com

Roee Aharoni
Google Research
roeeaharoni@google.com

Mor Geva
Tel Aviv University,
Google Research
pipek@google.com

Abstract

Factual questions can typically be answered correctly at different levels of granularity. For example, both "*August 4, 1961*" and "*1961*" are correct answers to the question "*When was Barack Obama born?*". Standard question answering (QA) evaluation protocols, however, do not take this into account explicitly and instead compare a predicted answer against reference answers of a single granularity level. In this work, we propose GRANOLA QA, a novel evaluation setting where a predicted answer is evaluated in terms of accuracy and informativeness against a set of multi-granularity answers. We present a simple methodology for enriching existing datasets with multi-granularity answers, and create GRANOLA-EQ, a multi-granularity version of the ENTITYQUESTIONS dataset. We evaluate models using a range of decoding methods on GRANOLA-EQ, including a new algorithm called Decoding with Response Aggregation (DRAG), that is geared towards aligning the answer granularity with the model's uncertainty. Our experiments show that large language models with standard decoding methods tend to generate specific answers, which are often incorrect. In contrast, when evaluated on multi-granularity answers, DRAG yields a nearly 20 point increase in accuracy on average, which further increases for rare entities, revealing that standard evaluation and decoding schemes may underestimate the knowledge encapsulated in language models.

1 Introduction

Large language models (LLMs) often generate factual errors, especially when the task requires less widely-known knowledge (Mallen et al., 2023; Sciavolino et al., 2021). Such factual errors are commonly attributed to the LM lacking relevant knowledge (Zheng et al., 2023) or over-committing to earlier mistakes (Zhang et al., 2023b).

We conjecture that factual mistakes can stem from a different failure source, when the model prioritizes different textual attributes (e.g., fluency or specific formats that appeared in the training corpora) over factuality. Such failures can result in generated text that mixes both correct and incorrect statements, even when the incorrect parts are not strictly required by the question.

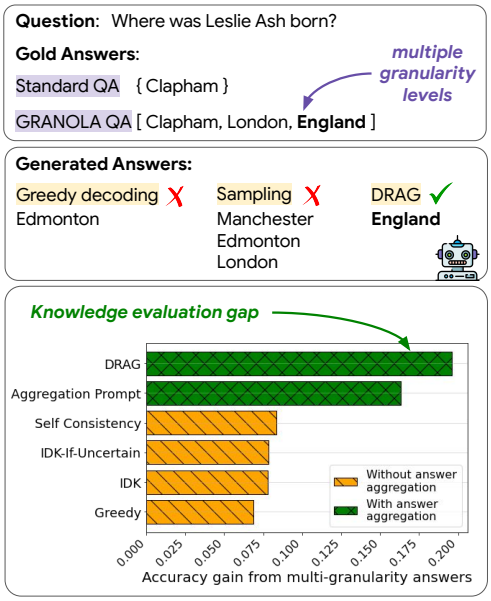

Figure 1: **Top**: GRANOLA QA evaluation with multi-granularity answers. **Middle**: Decoding with Response Aggregation (DRAG) outputs a (potentially coarser) response by aggregating several responses of the model. **Bottom**: Accuracy gain from evaluating using multi-granularity answers for several decoding strategies. DRAG reveals a significant knowledge evaluation gap.

Consider for example the question "*When was Mark Bils born?*". When prompting ChatGPT[1] for answering this question, sampled responses include "*March 22, 1958*", "*May 19, 1958*" and "*August 15, 1958*". This may suggest that the model is confident that Bils was born in 1958 – which is a correct answer in this case, albeit not the most informative one – yet it displays a preference for outputting a more detailed but incorrect response

[1]Responses were obtained by querying ChatGPT 3.5 using the standard Web API in December 2023.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6737–6751

in a specific full-date format.

This example also highlights how factual questions can be answered correctly *at different levels of granularity*. Namely, while the answers "*December 1, 1958*", "*December 1958*", and "*1958*" vary in terms of informativeness, they are all factually correct. However, answer granularity levels are not considered in standard question answering (QA) settings, which typically evaluate a predicted answer based on its similarity to a set of reference answers of the same (usually the most-specific) granularity. Even when different levels of granularity are present, there is no notion in which matching to a more specific answer is "better". As a result, standard QA evaluation may significantly *underestimate* the knowledge encapsulated in LMs, a phenomenon which we refer to as the *knowledge evaluation gap*. Indeed, recent human evaluation suggests that such granularity disparities account for approximately 10-15% of the disagreements between lexical matching and human evaluation (Kamalloo et al., 2023; Zheng et al., 2023).

In this work, we tackle this problem by proposing a novel multi-granularity QA evaluation setting, called GRANOLA QA (short for **GRAN**ularity **O**f **LA**bels). Unlike existing evaluation, in GRANOLA QA questions are labeled with ground-truth answers with multiple levels of granularity and predicted answers are evaluated in terms of both their *accuracy* and *informativeness* (§2). The evaluation is done using two new metrics: GRANOLA Accuracy, which checks if there was a match against *any* of the answers, and GRANOLA informativeness, which is a weighted score prioritizing fine-grained correct answers over their coarse-grained counterparts.

Next, we present a simple and general methodology for augmenting an existing single-granularity QA dataset to the setting of GRANOLA QA, which does not involve any human labor (§3). This process is based on obtaining additional information about entities present in the original questions and answer(s) from an external knowledge graph (KG), and then using an LLM to form multi-granularity answers conditioned on this information. We apply our methodology on the ENTITYQUESTIONS (EQ) dataset (Sciavolino et al., 2021), using WikiData (Vrandečić and Krötzsch, 2014) as the KG. The resulting dataset, GRANOLA-EQ, consists of 12K QA examples with an average of 2.9 multi-granularity answers per question. A manual analysis of a random subset of the data shows that our automatic procedure yields highly-accurate answers.

We evaluate various baselines on GRANOLA-EQ, including greedy decoding and methods that abstain from answering in cases of uncertainty (Yoshikawa and Okazaki, 2023a; Yang et al., 2023a,b; Ren et al., 2023). In addition, we introduce a novel decoding strategy, called Decoding with Response Aggregation (DRAG), that is geared towards aligning the granularity level of a model's response with its uncertainty level (§4). DRAG uses temperature sampling to obtain a set of candidate responses, and then answers the original question based on *an aggregation of these responses*, which we implement using few-shot prompting. Figure 1 depicts an example of DRAG's aggregation of several incorrect responses into a correct coarser answer that matches against the multi-granularity labels.

Our experiments (§5) show that: (1) with standard decoding the gap between GRANOLA accuracy and standard accuracy is small, which corroborates that LMs tend to output detailed responses, even when these are incorrect, (2) with DRAG this gap is high, showing that unlike standard decoding, DRAG outputs coarse answers, (3) GRANOLA accuracy remains high with DRAG even for rare entities, suggesting that LLMs know less *detailed* information about them rather than lacking any knowledge (Mallen et al., 2023), (4) compared to standard decoding and methods that allow the model to abstain from answering ("IDK"), DRAG yields a better trade-off between factuality and response informativeness, and (5) this evaluation gap is not observed when using semantic similarity scores against single-granularity reference answers.

To summarize, we introduce GRANOLA, a new QA evaluation setting that considers both the accuracy and informativeness of predicted answers. We propose a simple automatic procedure for generating accurate multi-granular answers for given QA pairs, and apply it to the ENTITYQUESTIONS dataset to create GRANOLA-EQ. We introduce a new decoding scheme, called DRAG, tailored to modify the response to a level of granularity that fits the model's uncertainty levels. We show that DRAG improves both informativeness and accuracy (relative to standard decoding), and that standard evaluation may significantly under-estimate the knowledge of LMs, especially about rare entities.

2 GRANOLA Question Answering

We formalize the setting of GRANOLA QA and define new metrics for quantifying accuracy and informativeness of QA predictions.

2.1 Problem Setting

In a typical open-domain QA setting (Yang et al., 2015; Voorhees et al., 1999; Kwiatkowski et al., 2019; Joshi et al., 2017; Sciavolino et al., 2021), a model predicts an answer p to a given question q, which is evaluated against an unordered set of gold answers $\mathcal{A} = \{a_1, \ldots, a_k\}$. The evaluation usually relies on lexical matching with standard metrics like exact-match or token-F1 between the predicted answer and each of the gold answers.[2] For example, a possible set of answers to the question "*Where is the headquarter of Guildhall School of Music and Drama?*" would be {Barbican Centre, The Barbican}. Importantly, the gold answers in $\mathcal{A}$ are interchangeable, where matching against either of a_1 or a_2 is equally good.

However, we observe that a question may be answered correctly at different levels of granularity. Namely, "London" is also a correct answer to the question, since the Barbican Centre is located there. If "London" does not appear in $\mathcal{A}$, standard evaluation will render this answer as incorrect, resulting in under-estimating the LM's knowledge. Moreover, if London is included in $\mathcal{A}$, then answering either "London" or "The Barbican" is considered equally correct, despite the fact that the second answer is more specific and arguably more valuable.

Here we propose that QA predictions should be evaluated while considering different granularity levels, a setting which we name GRANOLA QA. Formally, the answer p should be evaluated against an *ordered set of multi-granular* gold answers $\hat{\mathcal{A}} = \{\mathcal{A}_1, \ldots, \mathcal{A}_\ell\}$. Here, $\mathcal{A}_1$ is the set of the most informative correct answers (e.g. {Barbican Centre, The Barbican}) and $\mathcal{A}_\ell$ is the set of least-informative correct answers (e.g. "London" could be in $\mathcal{A}_2$ and "UK" in $\mathcal{A}_3$).

2.2 Evaluation

At a high-level, we will evaluate GRANOLA QA performance across two axes: *accuracy* and *informativeness*. Accuracy is determined based on whether the candidate answer matches against *any* of the GRANOLA answers; informativeness will reward matching against fine-grained answers by using an appropriate weighting scheme:

Definition 1 (GRANOLA Evaluation) *Given a question q, an answer p and GRANOLA labels $\hat{\mathcal{A}}$, accuracy and informativeness are evaluated based on a simple two-step procedure:*

Step 1: Find a match. *Let $i^\star \equiv i^\star(p; q, \hat{\mathcal{A}})$ denote the smallest index $i \in [k]$ for which there is a match between p and $\mathcal{A}_i \in \hat{\mathcal{A}}$ (meaning the F1 score between p and an answer in $\mathcal{A}_i$ exceeds some threshold τ), or $\perp$ if no match is found.*

Step 2: Evaluate. *GRANOLA accuracy is defined as $\mathbf{1}[i^\star \neq \perp]$. Informativeness is defined as $\exp(-\lambda \cdot (i^\star - 1))$, or 0 if no match was found.*

The notion of informativeness relies on a weighting scheme that assigns a weight of 1.0 to the fine-grained answers $\mathcal{A}_1$, and exponentially decreasing weight for answers $\mathcal{A}_{i>1}$. This represents the diminished utility of coarser answers. The parameter λ can be used to control the rate of decrease: as $\lambda \to 0$ coarser answers receive higher weights; see Appendix B for a visualization of how the weights behave as a function of λ.

3 Enriching QA Samples with Multi-Granularity Answers

We turn to the question of constructing GRANOLA QA benchmarks. We observe that multi-granularity answers are in principle *abstractions* of the most-detailed answer. For example (see Figure 3), the answer "Michael Madhusudan Dutta" to the question "*Who translated the play Neel Darpan into English?*" can be abstracted into a higher-level description such as "An Indian Poet". Therefore, one way to generate multi-granularity answers is to start from an existing QA pair and enriching it with multi-granularity answers through abstraction.

Following this approach, we describe a simple and automatic procedure for adjusting factual QA datasets to GRANOLA QA (§3.1). Then, we apply this procedure to the ENTITYQUESTIONS dataset (§3.2), a widely used entity-centric QA dataset (Sciavolino et al., 2021), to create a multi-granularity QA benchmark. Last, we manually analyze the quality of the generated data (§3.3).

3.1 Automatic Answer Generation

We focus on evaluating factual knowledge in LLMs, where the answer to a given question is an entity (e.g., a person or a place). Given an an-

[2] The answers are typically being normalized (i.e. case-folding and removing punctuation and articles).

Figure 2: Our procedure for adding multi-granularity answers to given QA pairs.

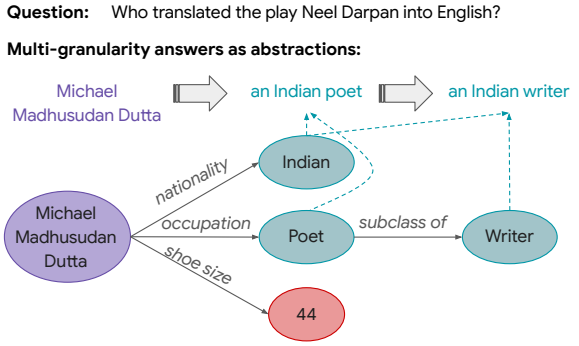

Figure 3: An illustration of multi-granularity answers as entity abstractions. Given an answer entity, we use an external KG to generate coarser answers from its properties (turquoise) in addition to the original answer (purple). Notably, not all KG properties are equally good candidates for multi-granular answers (red).

swer, we propose to generate coarser versions of it by utilizing an external knowledge graph (KG). Specifically, given a KG with facts encoded as subject-relation-object triplets (e.g., the triplet (`Paris`, `capital of`, `France`) would encode the fact that Paris is the capital of France) and an answer entity e, coarser versions of e can be obtained by replacing it with higher-level properties of it in the KG. For example (Figure 3), replacing the answer "Michael Madhusudan Dutta" with its properties of `Nationality` and `Occupation` would create a new coarser answer "Indian Poet".

In principle, however, there are many possible answer properties that can be used – and intuitively, not all of them are key properties of the entity that are useful for evaluating general factual knowledge. For example, answering the original question with Michael Madhusudan Dutta's shoe size is not what we want to capture by coarse answers. Thus, to create a generic methodology for enriching an existing QA dataset with answers, we must be able to automatically determine the relevant properties.

To overcome this challenge, instead of relying on KG triplets directly, we use short textual descriptions that capture the key properties of the entity in the KG. Such descriptions are often offered by knowledge sources such as WikiData. For example, the entity Michael Madhusudan Dutta has the following description: "*Bengali poet and dramatist*".

Overall, our answer generation process has two steps, depicted in Figure 2. Given a QA pair, we first obtain a description of the answer entity and any entities appearing in the question from an external KG. Then, we zero-shot prompt an LLM to generate an ordered list of answers at varying levels of granularity, conditioned on the given QA pair and the entity descriptions. See Table 8 for the exact instruction prompt.

3.2 GRANOLA ENTITYQUESTIONS

We apply the procedure described in §3.1 to enrich the test split of ENTITYQUESTIONS (EQ) (Sciavolino et al., 2021) with GRANOLA answers. ENTITYQUESTIONS is an entity-rich QA dataset, created by converting factual subject-relation-object triples into natural language questions using manually-defined templates. We use PaLM 2-L as the LLM (Anil et al., 2023).

The resulting dataset, which we refer to as GRANOLA-EQ, spans 16 relations and has a total of 12,452 examples. Overall, our procedure yielded 2-3 coarser answers per questions (∼20% have 2 answers overall, ∼60% have 3, and ∼15% have 4 or more; this is distributed relatively uniformly over relations). Examples from GRANOLA-EQ are shown in Table 1. More details are in Appendix C.

3.3 Data Quality

We manually evaluate the quality of a generated answer a with respect to a question q from GRANOLA-EQ across the following axes:

- **Correctness**: We use WikiData to verify whether a is a factually correct answer to q. Notably, while a was generated conditioned on the description, the LLM might produced it while relying on its parametric knowledge rather the information in the description. For example, for the question "*Where did Marcel Gaumont die?*", the model generated the answers "Paris", "Île-de-France", and "France" while the WikiData description of Paris is "*Capital of France*". Therefore, in this case the LLM used its parametric knowledge to add a new granularity level (Île-de-France).

Question	GRANOLA Answers
"*Where was Fiona Lewis born?*"	Westcliff-on-Sea; Essex; England
"*What music label is Courage represented by?*"	Rock Records; a Taiwanese record label
"*Who is August von Hayek's child?*"	Friedrich Hayek; an economist
"*Who is the author of The Adding Machine?*"	Elmer Rice; an American playwright; a playwright
"*Where was Toby Shapshak educated?*"	Rhodes University; Makhanda, South Africa; South Africa

Table 1: Examples from GRANOLA-EQ. Answers are separated by a semicolon and listed fine-to-coarse. The first answer is the original answer in ENTITYQUESTIONS; subsequent answers were generated (see §3.1).

- **Informativeness**: We verify that a is a non-trivial answer to q. We consider an answer as trivial if it could be generated based on the question template alone (i.e., a version of q in which the entity is redacted). For example, "*Earth*" is a trivial answer to the question "*Where was Fiona Lewis born?*" because it could be obtained based on the template `Where was [X] born?`.
- **Granularity**: We assess whether a is coarser than the answers preceding it. For the first GRANOLA answer, we define this as whether the answer is identical to the original answer.

We treat these metrics as binary and manually evaluate a sample of 1% of the data (124 questions and their corresponding 358 answers). Table 2 reports the fraction of examples in each error category with a representative example. Our evaluation reveals that the enriched answers are of high-quality, with all of the generated answers being factually correct Nonetheless, there is headroom for improving our answer generation procedure. E.g., there are examples with useful information in the description that is not utilized by the model, (suggesting the knowledge evaluation gap may be even larger than observed in our results in §5.)

4 Decoding with Response Aggregation

Humans naturally tailor the granularity level of their responses to their uncertainty levels. Consider asking a person A, when another person B was born. The format of the response will depend on the relationship between A and B, and specifically on how much A knows about B. For example, if A is extremely familiar with B (e.g., B is A's son), then we expect the answer to include the full date of birth. If A is only partially familiar with B (e.g., B is a celebrity that A knows), then we expect the answer to be more generic (e.g. only the year or decade). If A is not familiar with B, then we expect A to say that they do not know the answer.

Error type (%)	Example
Informativeness (6%)	Question: What music label is Sarah Buxton represented by? Answers: Lyric Street Records; a music label
Granularity (9%)	Question: Who owns Eccles Coliseum? Answers: Southern Utah University; a public university; a public university in Utah

Table 2: Human evaluation results of GRANOLA-EQ, showing for each error type the fraction of erroneous cases and an example.

In this section, we propose a novel decoding strategy, called Decoding with Response Aggregation (DRAG), that is intended to encourage LMs to do the same. We focus on a fixed (i.e., frozen) LM, and our objective is to improve factuality at inference time by attempting to provide a coarser answer in the place of a fine-grained but incorrect answer. In §5, we will evaluate our proposed decoding strategy against various existing baselines on the GRANOLA QA dataset we constructed.

DRAG consists of two stages:

- **Sampling**: We sample N responses from the model with temperature $T > 0$.
- **Aggregation**: The final output is the most informative response that is consistent with the set of sampled responses. This can be implemented in different ways, e.g. via prompting an LLM.

Revisiting the example question "*When was Mark Bils born?*" (§1), aggregating the sampled responses "*March 22, 1958*", "*May 19, 1958*" and "*August 15, 1958*", should yield "*1958*". Pseudocode for DRAG is provided in Figure 4.

Choice of hyperparameters The sampling temperature T and number of responses N control the trade-off between factuality and informativeness. Intuitively, larger values of T and N encourage more diverse outputs and hence more aggregation, favouring factuality over informativeness.

DRAG vs existing decoding strategies When $N = 1$, the aggregation is trivial and DRAG recovers standard decoding strategies (e.g. greedy decoding or temperature sampling, based on the value of T). Conceptually, DRAG is also a generalization of

Hyperparameters: Temperature $T > 0$; number of samples N
Input: Input x; Model M
Generate $\{r_1, \ldots, r_N\}$ continuations for $M(x)$ at temperature T;
Let $\hat{r} = \texttt{ResponseAgg}(\{r_1, \ldots, r_N\})$;
return The aggregated response $\hat{r}$

Figure 4: Decoding with Response Aggregation (DRAG). We implement ResponseAgg by instructing an LLM to output what $r_1, \ldots, r_N$ have in common, or IDK if they do not share meaningful properties.

other popular decoding strategies that are based on sampling a set of candidate responses. For example, replacing our proposed aggregator with a naive aggregation that outputs the majority response recovers *self-consistency* (Wang et al., 2022).

5 Experiments

We assess how accounting for answer granularity, both in evaluation and during decoding, influences the evaluation of LLM performance on factual questions. After describing our experimental setting (§5.1), we compare between evaluation with standard accuracy and GRANOLA accuracy (§5.2), which reveals that current QA settings underestimate LLMs' knowledge. Then, we show that the gains in accuracy from using GRANOLA cannot be matched by existing semantic similarity scores (§5.3), which highlights the utility of this setting in capturing differences between multi-granularity answers. Last, we use the GRANOLA metrics to evaluate DRAG with respect to baselines in terms of accuracy and informativeness (§5.3), showing its superiority in decoding answers that are tuned towards the LLM's knowledge.

5.1 Experimental Setting

We evaluate DRAG and multiple baselines on GRANOLA-EQ in a closed-book setting, where factual questions must be answered without access to an external knowledge source (Petroni et al., 2019).

For the aggregation stage of DRAG, we instruct an *aggregator* LLM to output what the sampled responses have in common or IDK if the responses have nothing meaningful in common (see Table 8 in Appendix D for the exact prompt).

Baselines We consider the following methods:

- **Standard Decoding**: We evaluated both greedy decoding (Greedy) and temperature sampling (TS), but since TS consistently under-performed Greedy we report results only for Greedy.
- **I don't know (IDK)**: Given the established success of steering model behaviour via prompting (Mishra et al., 2021; Si et al., 2022; Ganguli et al., 2023), we consider two prompt-based IDK variants. In IDK, the model is instructed to either answer the question or output IDK. In IDKIfUncertain, the model is specifically instructed to output IDK if its uncertainty is high.
- **Aggregation-based baselines**: We evaluate DRAG and IDKWithAgg, in which we instruct the model to answer at a level of granularity that matches its uncertainty. As an ablation for the importance of the aggregation step in DRAG we also evaluate SelfConsistency (Wang et al., 2022), where we sample N responses at temperature T and output the majority response.[3] As noted in §4, SelfConsistency can be cast as an instance of DRAG with a simple aggregator (majority rule).

See Table 7 for the prompts used for the baselines.

Evaluation We use *GRANOLA accuracy* and *informativeness* as described in Definition 1. To account for cases of IDK predictions, we adopt the perspective of *selective prediction* (El-Yaniv et al., 2010; Geifman and El-Yaniv, 2017) with recent applications in QA (Kamath et al., 2020) and text generation (Yoshikawa and Okazaki, 2023a). Informativeness is left as is, except that IDK predictions are defined to contribute a score of 0.0, since they are not informative at all. GRANOLA Accuracy is replaced with *selective GRANOLA accuracy*, which is the mean GRANOLA accuracy on the subset of predictions which are not IDK.

Models We use instruction-tuned versions of PaLM 2-M and PaLM 2-L, the medium and large variants of the PaLM 2 LLM (Anil et al., 2023).

5.2 Knowledge Evaluation Gap

Figure 5 shows GRANOLA accuracy as a function of standard accuracy, for the different models and methods. Note that the vertical distance from the $x = y$ line (black) represents the gain in accuracy from evaluating using multi-granularity answers. We observe that this gap is similar and relatively small of ∼5 points (grey dotted line) for methods that do not explicitly incorporate aggregation. This

[3] After case-folding and removing punctuation and articles.

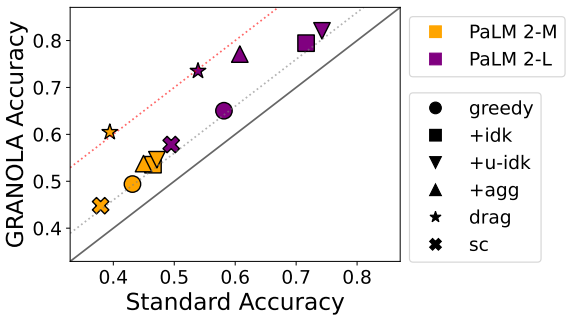

Figure 5: Standard accuracy vs. GRANOLA accuracy for the different models we evaluate.

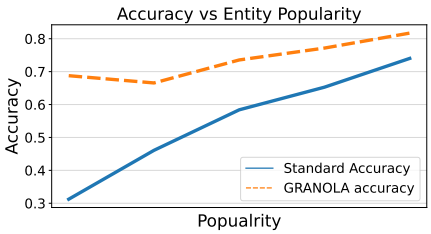

Figure 6: Accuracy vs. entity popularity for PaLM 2-L using DRAG. Unlike standard accuracy, which declines steeply in popularity, GRANOLA accuracy plateaus.

confirms our initial conjecture that standard decoding tends to generate detailed but incorrect responses. In addition, for the aggregation methods, this gap is substantially larger, nearing a ∼20 point increase (red dotted line). This demonstrates that both explicit aggregation (DRAG) and implicit aggregation obtained via prompting can successfully steer the model towards tailoring its response granularity. It also reveals that the knowledge evaluation gap is both a function of existing evaluation practices *and* standard decoding strategies. In Figure 10 in Appendix E we show a breakdown of these results to the different relations in GRANOLA-EQ, revealing that certain relations especially gain from multi-granularity answers.

Next, we consider how this gap behaves as a function of *popularity*.[4] In Figure 6 we stratify GRANOLA-EQ into equally sized bins by popularity (x-axis) and compare standard accuracy (blue) with GRANOLA accuracy (orange, dashed). While standard accuracy steeply declines with popularity, GRANOLA accuracy plateaus. This reveals that models do capture knowledge about even very rare entities (but this knowledge is coarser). In Figure 11 (§B) we show that this behaviour is not demonstrated by standard decoding.

[4]We quantify popularity using Wikipedia page-views for the question entity's page.

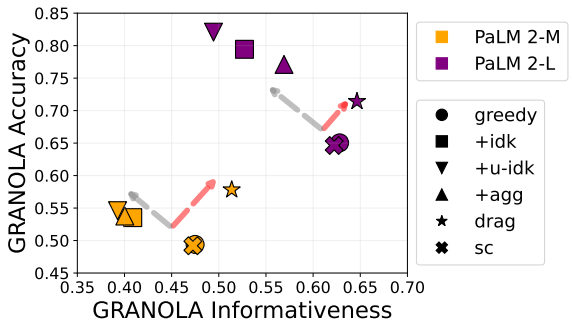

Figure 7: Answer accuracy vs. informativeness when using DRAG compared to the baselines. Behaviour is consistent across model sizes (purple/orange): IDK baselines improve accuracy at the cost of making less informative predictions (grey arrow); DRAG improves both accuracy and informativeness (red arrow).

5.3 Evaluation of DRAG

Figure 7 shows the GRANOLA accuracy and informativeness of DRAG compared to the baselines. The results are consistent across model sizes (purple vs orange). Figure 8 provides a more detailed picture of the distribution of which GRANOLA answer matched against the predicted answers (see Definition 1). We distill several key takeaways:

(1) IDK baselines improve accuracy at the cost of less informative predictions (grey arrows in Figure 7): As expected, abstention (IDK) improves the selective accuracy. However, as evident in Figure 7, this comes at the cost of predictions that are overall less informative. For example, the fraction of errors made by IDK drops from 42% to 31% – but 17% of the predictions are IDK. The number of coarse correct answers is unchanged at ∼5%.

(2) DRAG improves both accuracy and informativeness (red arrows in Figure 7): Compared to standard decoding, DRAG improves both accuracy and informativeness. As evident from Figure 7, this is obtained by a smaller fraction of abstentions (6%) and a significantly larger fraction of coarse correct answers (16%). This result confirms our original conjecture that the dichotomy (know/don't know) underlying IDK methods is too coarse.

5.4 Meta-evaluation

In the previous sections, we showed that multi-granularity answers facilitate a more faithful evaluation of LLM performance on factual questions. Here, we check whether a similar effect could be obtained by evaluating with semantic similarity against single-granularity reference answers.

To this end, we test if semantic similarity against

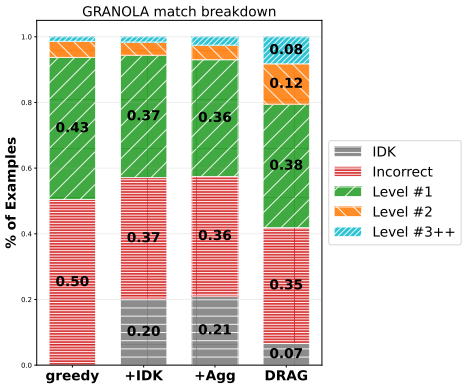

Figure 8: The granularity of answers predicted by PaLM 2-M. Level numbers correspond to the answer index in the ordered set of GRANOLA answers, with 1 being the most fine-grained. While all methods decrease the fraction of errors compared to greedy (from 50% to 35-37%; **red**), DRAG does this with significantly fewer IDK predictions (e.g., 7% vs 20%); **gray**) and more coarse correct answers (e.g. 20% vs 5%).

Standard accuracy	GRANOLA accuracy	% of examples	BLEURT score
✓	✓	49.5	0.83
✗	✓	5.6	0.28
✓	✗	0.0	-
✗	✗	44.9	0.26

Table 3: Mean BLEURT score for PaLM 2-L with greedy decoding on GRANOLA-EQ, stratified by standard accuracy and GRANOLA accuracy.

single-granularity answers can distinguish between answers that GRANOLA accuracy deems correct and incorrect. Concretely, we stratify GRANOLA-EQ according to whether both the standard and GRANOLA F1 scores exceed a threshold τ, and report the mean semantic similarity score for each of the four resulting subsets. Note that, by definition, the standard F1 is a lower bound to GRANOLA F1, so one of the subsets is empty.

Table 3 shows the results when using BLEURT (Sellam et al., 2020) as the semantic similarity metric. The mean BLEURT score is similar for examples that are incorrect according to both metrics and for examples that are correct only according to GRANOLA accuracy (gray rows). This highlights that BLEURT is not a good proxy for matching against multi-granularity answers. Examples from GRANOLA-EQ where GRANOLA accuracy disagrees with both standard accuracy and BLEURT score are provided in Table 9 (Appendix E).

6 Related work

Answer annotation in QA datasets. QA benchmarks, e.g. Natural Question (Kwiatkowski et al., 2019), often have multiple answers per question, which may inadvertently include multi-granularity answers. Min et al. (2020) consider the problem of ambiguous questions, proposing question rewriting to resolve ambiguity. Si et al. (2021) mine answer aliases from a KG and use them to perform "answer expansion" to increase the lexical matching score. Our approach is similar but goes one step further, using the KG and LLMs to add multi-granularity answers vs. simply using aliases.

Granularity-driven evaluation. Granularity of model responses has been evaluated in the context of open-domain chatbots, where informativeness plays a crucial role in building engaging dialogue agents. Adiwardana et al. (2020); Thoppilan et al. (2022) evaluate granularity, but their focus is on conversational language rather than knowledge evaluation. Huang et al. (2022) use WikiData to form masked token prediction tasks, such as "*Toronto is located in [MASK]*", and test whether pretrained models have a preference for more specific completions (e.g. "*Ontario*" vs "*Canada*"). Their approach only accommodates single-token predictions and their evaluation covers smaller models. Conceptually, while their objective is to encourage specific answers, we use granularity to perform more faithful evaluation of LM's knowledge and factuality.

Punting. Abstaining from answering questions is a popular approach for improving factuality (Kadavath et al., 2022; Kuhn et al., 2023; Yoshikawa and Okazaki, 2023b; Chen et al., 2023; Zhang et al., 2023a). Our approach is motivated by the observation that punting may be overly aggressive; when the model has low confidence in a specific answer but is confident in a coarser answer, outputting the coarser answer is preferred over refusing to answer.

7 Conclusion

We highlight a prominent source of factuality errors in modern LMs: generating more detailed responses than their knowledge can support. Using a new QA benchmark, GRANOLA-EQ, with multi-granularity answers, and a novel decoding algorithm, DRAG, we show that taking the answer granularity level into account leads to a dramatic increase in model accuracy. In Appendix A we discuss various directions for future work.

Limitations

Technically, our approach for enriching an existing QA benchmark with multi-granularity answers relies on extracting entities from the original QA pair and matching them to their KG entry. In less-structured datasets this step may be more involved – for example, if the surface form of the entity name differs between the dataset and the KG.

On a more conceptual level, a faithful evaluation of the knowledge of LLMs may also require distinguishing between correct answers based on true knowledge, as opposed to mere educated guesses. This is an issue with QA evaluation in general – but is especially relevant in our setting, since coarser answers are easier to guess correctly. For example, in the question "*Where was [X] born?*", one could guess "*Russia*" if X is a Russian-sounding name (whereas correctly guessing the city X was born in is less likely). This may require additional information (in the form of providing additional information such as reasoning or evidence) but also relates to how one defines knowledge.

Other than that, our work was demonstrated on a set of large-but-specific LMs from the PaLM model family. Further expanding the study to a wider range of models may also be compelling, but beyond the scope of this work.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew R Gormley. 2023. It's mbr all the way down: Modern generation techniques through the lens of minimum bayes risk. *arXiv preprint arXiv:2310.01387*.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with self-evaluation to improve selective prediction in llms. *arXiv preprint arXiv:2310.11689*.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation. *arXiv preprint arXiv:2305.09860*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Can language models be specific? how? *arXiv preprint arXiv:2210.05159*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv:2305.06984*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021. Reframing instructional prompts to gptk's language. *arXiv preprint arXiv:2109.07830*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Roni Rabin, Alexandre Djerbetian, Roee Engelberg, Lidan Hackmon, Gal Elidan, Reut Tsarfaty, and Amir Globerson. 2023. Covering uncommon ground: Gap-focused question generation for answer assessment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 215–227, Toronto, Canada. Association for Computational Linguistics.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. *arXiv preprint arXiv:2312.09300*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. *arXiv preprint arXiv:2109.08535*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. *arXiv preprint arXiv:2109.05289*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

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

Liming Wang, Siyuan Feng, Mark Hasegawa-Johnson, and Chang Yoo. 2022. Self-supervised semantic-driven phoneme discovery for zero-resource speech recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8027–8047, Dublin, Ireland. Association for Computational Linguistics.

Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, et al. 2023a. Uncertainty-aware language modeling for selective question answering. *arXiv preprint arXiv:2311.15451*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023b. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Hiyori Yoshikawa and Naoaki Okazaki. 2023a. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia. Association for Computational Linguistics.

Hiyori Yoshikawa and Naoaki Okazaki. 2023b. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1972–1983.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023a. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023b. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers. *ArXiv preprint, abs/2304.10513*.

A Directions for future work

Question perturbations. Our approach for generating multi-granularity answers relied on abstractions. A complementary approach would modify the question rather than its answer, e.g., altering the question "*When was Mark Bils born?*" to "*In what year was Mark Bils born?*". Such question perturbations could also be coupled with our entity abstraction perspective to generate more broad questions like "*When was a professor from University of Rochester born?*". Another direction considers generating more specific questions to address knowledge gaps (Rabin et al., 2023). However, question perturbations may create new answers and thus would require more complex evaluation.

Improving DRAG. The two stages of DRAG – sampling candidate responses, and response aggregation – could be improved to yield better granularity adjustment. For example, it is possible to replace regular temperature sampling (Ackley et al., 1985) with other sampling strategies that may perform better (Wang et al., 2022; Freitag et al., 2023; Bertsch et al., 2023). Additionally, better aggregators could improve downstream task performance.

Response granularity fine-tuning. While this work focused on improving factuality at inference time, it is interesting to explore fine-tuning with response granularity in mind. For example, DRAG can be used as a reward model for supervised or RLHF finetuning to encourage models to learn how to tailor the their response granularity to their parametric knowledge or the preceding context.

B Additional figures

In Figure 9 we show how the parameter λ impacts the scores used to evaluate informativeness (see Definition 1).

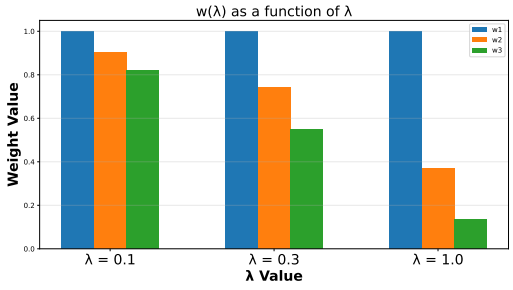

Figure 9: Letting $w_i = \exp(-\lambda(i-1))$, we show how w_1, w_2, w_3 behave for different values of λ.

C Full details: Constructing GRANOLA-EQ

In this section we detail our process of enriching ENTITYQUESTIONS with multi-granular answers.

Entity Questions. ENTITYQUESTIONS (EQ) (Sciavolino et al., 2021) is a entity-rich QA dataset. It was created by selecting 24 common relations from Wikidata and converting fact (subject, relation, object) triples into natural language questions using manually defined templates. We restrict our attention to the test split of ENTITYQUESTIONS. In the original ENTITYQUESTIONS some of the relations have answers that are already coarse (e.g. "*in which continent is country [X]*"). We thus filter examples belonging to such relations. See Table 4 for the full list of ENTITYQUESTIONS and GRANOLA-EQ relations. Additionally, for simplicity, we only keep rows with a unique ground truth example in the original ENTITYQUESTIONS dataset.

Obtaining WikiData descriptions for entities. The entity extraction stage is simple since the entity location is encoded in the template. For each QA pair we extract two entities: the question entity and the subject entity, and then look these up in WikiData to obtain a WikiData qid for each entity. We then use the qid to obtain a free text description of the original entity.

QID disambiguation. In approximately 30% of the cases, there are multiple potential qid matches for the same entity (see Figure 5 for an example). We use a simple heuristic for performing disambiguation: we select the qid with the smallest value.

Data cleaning. There are two sources of noise in the above automatic process: incorrect extracted descriptions (this may occur when there are multiple WikiData entries for the same entity name, and our disambiguation procedure selects the wrong one) and errors in the LLM generated answers. Thus, to ensure the data is of high quality, we apply several automatic cleaning operations. First, we remove rows containing descriptions that are likely to be erroneous. We utilize the observation that when the QID disambiguation heuristic fails and the wrong QID is selected, this failure will typically be evident from the fact that the extracted description is not semantically consistent with the question; see Table 6 for concrete examples. To remove these examples we score each example for how *consistent* the extracted descriptions are with the original

Template	Relation	Included?
Which country is [X] located in?	P17	✗
Where was [X] born?	P19	✓
Where did [X] die?	P20	✓
Who is [X] married to?	P26	✓
Which continent is [X] located?	P30	✗
What is the capital of [X]?	P36	✗
Who is [X]'s child?	P40	✓
Who is the author of [X]?	P50	✓
Where was [X] educated?	P69	✓
What kind of work does [X] do?	P106	✗
Who founded [X]?	P112	✓
Who owns [X]?	P127	✓
Where is [X] located?	P131	✓
What type of music does [X] play?	P136	✗
Where is the headquarter of [X]?	P159	✓
Who was [X] created by?	P170	✓
Who performed [X]?	P175	✓
Which company is [X] produced by?	P176	✓
What music label is [X] represented by?	P264	✓
Where is [X] located?	P276	✓
Which language was [X] written in?	P407	✗
What position does [X] play?	P413	✗
Which country was [X] created in?	P495	✗

Table 4: List of ENTITYQUESTIONS and GRANOLA-EQ relations.

questions, and remove examples for which this predicted score exceeds 0.5. Specifically, we prompt an LLM (with 5 few-shot demonstrations of the intended behaviour) to determine whether the description is consistent ('Yes' or 'No'), and we determine the score as the fraction of 'No' responses (sampled at unit temperature). This process ends up removing 1409 examples (or 9.5% of the dataset). As a second data cleaning step, we remove rows with missing GRANOLA answers or duplicated answers. Finally, we remove GRANOLA answers from a list of hard-coded responses that we define as trivial (such as "person", "university", etc). In total, these steps affected 2378 rows (or 16% of the dataset).

D Prompts

Table 7 details the prompts used for baseline algorithms. Table 8 details the prompts used for the response aggregation sub-routine in DRAG and for generating multi-granular answers to create GRANOLA-EQ.

qid	description
Q64	federated state, capital and largest city of Germany
Q142659	census-designated place in Holmes County, Ohio
Q524646	town in Massachusetts
Q614184	town in Maryland, United States

Table 5: QID matches and descriptions for the free text entity "*Berlin*".

Question	QID	QID Description
Who is the author of Enduring Love?	Q129813	2004 film by Roger Michell
Who performed Orbit?	Q2367904	historical motorcycle manufacturer
Who is the author of Hollywood?	Q34006	neighborhood in Los Angeles, California, United States

Table 6: When QID disambiguation chooses the incorrect entity, the failure is typically evident since the extracted description (rightmost column) does not semantically match the question.

E Additional Results for §5

In this section we include supplementary results from §5.

Baseline	Prompt
`Vanilla`	Question: {question} Answer:
`IDK`	You will be given a question. Answer the question, or output IDK. Question: {question} Answer:
`IDKIfUncertain`	You will be given a question. Answer the question, or, if you are not certain of the answer, output IDK. Question: {question} Answer:
`Agg`	You will be given a question. Answer the question at a level of granularity that fits your uncertainty, or output IDK. Question: {question} Answer:

Table 7: Prompts used in the baselines evaluated in §5.

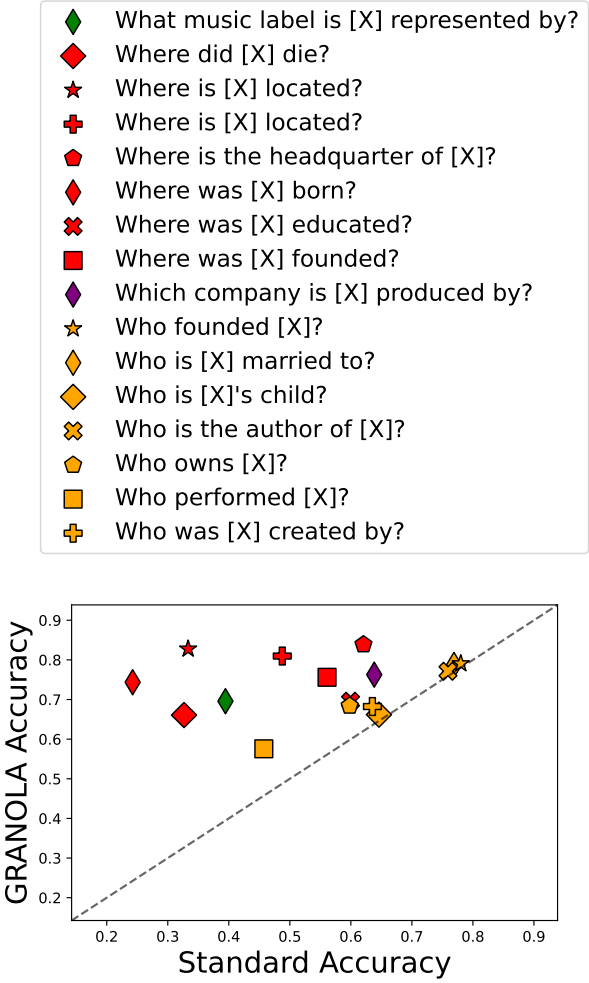

Figure 10: Standard Accuracy (x-axis) vs GRANOLA accuracy (y-axis), stratified by relation, for `DRAG` (PaLM 2-L).

Procedure	Prompt
Forming multi-granular answers	You will be given a pair of question and answer. You will also receive some additional description about the entity in the question and the entity in the answer. Your task is to write NEW ANSWERS for the original question at various levels of granularity. Number these answers starting from 1 (with 1 being the most fine grained answer – the original answer), and larger indices corresponding to coarser answers. The idea is that someone might not know the answer at the most fine-grained level, but perhaps know the answer at coarser levels. Important: STOP generating answers BEFORE you reach trivial answers. For example, given the question "who wrote the book X", answers such as "a writer" or "a person" are considered trivial, as these are completely uninformative and can be guessed even without knowing what X is. In your answers, use the format '1:: answer', etc.
Response Aggregation	You will be given a list of responses; replace them with the most specific answer that is still consistent with all the original responses. If the responses have nothing meaningful in common with respect to the question, output IDK. Here are some examples: Question: Where was [X] born? Responses: - Hamburg - Hamburg - Bonn - Berlin Correct aggregated answer: Germany Incorrect aggregated answer: Hamburg Explanation: These are all different cities in Germany. Hamburg is not a correct aggregation, since it is not consistent with other responses, such as Berlin or Bonn. Question: When was [X] born? Responses: - February 1, 1937 - November 20, 1937 - January 1937 Correct aggregated answer: 1937 Incorrect aggregated answer: November 1937 Explanation: These are all dates in 1937.

Table 8: Prompts used in GRANOLA related procedures in the paper.

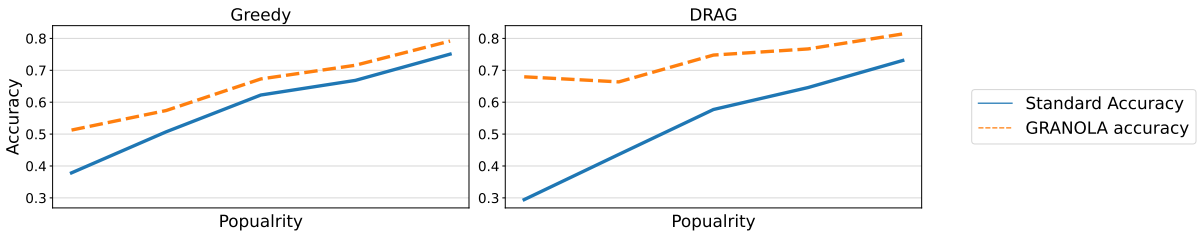

Figure 11: Accuracy (y-axis) vs Entity Popularity (x-axis) for two algorithms: Greedy (left) and DRAG (right). The underlying model is PaLM 2-L. We see that the knowledge evaluation gap is evident for DRAG. The behaviour for PaLM 2-M is identical, except the absolute numbers are smaller.

Question	**Original GT Answer**	**Candidate Answer**	**Matched GRANOLA Answer**	**BLEURT Score (GT vs. Candidate)**	**F1 Score (GT vs. Candidate)**	**F1 Score (GRANOLA vs candidate)**
What is Aline Brosh McKenna famous for?	27 Dresses	screenwriter	being a screenwriter	0.04	0.00	0.67
Where did Tilly Armstrong die?	Carshalton	London	London Borough of Sutton	0.05	0.00	0.40
Where is the headquarter of Guildhall School of Music and Drama?	Barbican Centre	London	City of London	0.06	0.00	0.50
Where is Battersea Park located?	Battersea	London	London	0.06	0.00	1.00

Table 9: Examples from GRANOLA-EQ where GRANOLA accuracy disagrees with standard metrics (lexical matching and semantic matching to the original GT answer). The examples were obtained by filtering for example with low F1 score to the original GT answer but high F1 score to the matched GRANOLA answer, and then sorting by BLEURT scores in ascending order. I.e., they correspond to the points in the top-left corner of Figure 12.

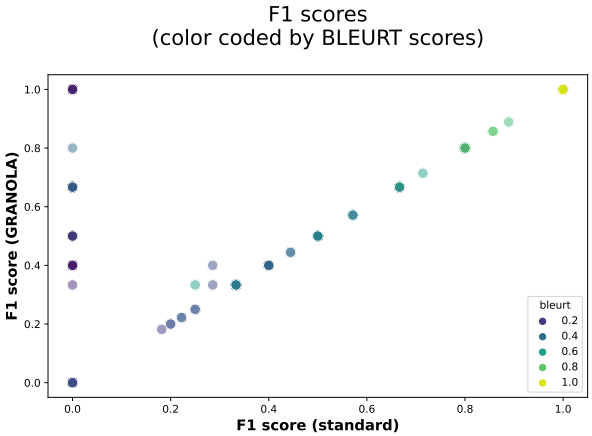

Figure 12: The relationship between standard F1 score (x-axis), GRANOLA F1 score (y-axis) and BLEURT score (Sellam et al., 2020) (color) computed between the original gt answer the candidate answer. The underlying model is Greedy (PaLM 2-L). The figure demonstrates that while there is a strong correlation between standard F1 score and BLEURT scores, this correlation fails specifically for the subset of examples for which GRANOLA accuracy disagrees with standard accuracy. See Table 3 for a quantitative version of this plot. This demonstrates that BLEURT scores can not serve as a replacement for GRANOLA labels.