The Consistency Hypothesis in Uncertainty Quantification for Large Language Models

Quan Xiao[*1] **Debarun Bhattacharjya**[2] **Balaji Ganesan**[2] **Radu Marinescu**[2] **Katsiaryna Mirylenka**[3]
Nhan H Pham[2] **Michael Glass**[2] **Junkyu Lee**[2]

[1]Rensselaer Polytechnic Institute [2]IBM Research [3]Zalando

Abstract

Estimating the confidence of large language model (LLM) outputs is essential for real-world applications requiring high user trust. Black-box uncertainty quantification (UQ) methods, relying solely on model API access, have gained popularity due to their practical benefits. In this paper, we examine the implicit assumption behind several UQ methods, which use generation consistency as a proxy for confidence—an idea we formalize as the *consistency hypothesis*. We introduce three mathematical statements with corresponding statistical tests to capture variations of this hypothesis and metrics to evaluate LLM output conformity across tasks. Our empirical investigation, spanning 8 benchmark datasets and 3 tasks (question answering, text summarization, and text-to-SQL), highlights the prevalence of the hypothesis under different settings. Among the statements, we highlight the 'Sim-Any' hypothesis as the most actionable, and demonstrate how it can be leveraged by proposing data-free black-box UQ methods that aggregate similarities between generations for confidence estimation. These approaches can outperform the closest baselines, showcasing the practical value of the empirically observed consistency hypothesis.

1 INTRODUCTION & RELATED WORK

Large language models (LLMs) have become pervasive due to their state-of-the-art performance for various natural language processing tasks. Despite recent advances, LLMs are known to suffer from important limitations that heavily impact their usability in diverse real-world applications, including lack of traceability, context sensitivity, and difficulty around incorporating domain-specific knowledge and handling rare queries as well as dynamic and evolving data. These issues often lead to unpredictable or unverifiable responses, or outputs that appear authoritative even when they are incorrect, which can undermine user trust in high-stakes domains such as healthcare, finance, and law.

Uncertainty quantification (UQ) approaches provide insights into the reliability of an LLM's predictions by associating them with *confidence* estimates, making them a critical component of many real-world systems. The estimated confidences should ideally be well *calibrated*, as gauged by the degree to which they match the empirical accuracy for that prediction [Murphy and Epstein, 1967, Dawid, 1982]. They can be used to distinguish correct LLM responses from incorrect ones, as well as for *selective generation*, i.e. for rejecting a fraction of the instances that one is least confident about [El-Yaniv et al., 2010]. Robust UQ enables downstream decision-making systems to act conservatively when necessary, and can facilitate human-AI collaboration by allowing users to identify when to trust or question model outputs. Moreover, well-calibrated uncertainty estimates are essential for integrating LLMs into broader pipelines where resource allocation or fallback mechanisms depend on confidence-aware behavior.

There is a growing body of literature on UQ techniques for estimating the confidence of LLM generations. *Verbalized* UQ approaches rely on prompting LLMs to express uncertainty about a generated response in natural language, such as through qualifying phrases or numbers [Lin et al., 2022, Kadavath et al., 2022, Mielke et al., 2022]. For instance, Kadavath et al. [2022] use a self-verification approach for question answering tasks where an LLM is asked to declare whether a provided answer is true or false for a question, and the model's token logit is used to infer the probability that the answer is correct. Although some empirical studies show that verbalized confidences using LLMs trained via the reinforcement learning with human feedback paradigm can yield well-calibrated estimates [Tian et al., 2023], others

[*] This work was conducted when QX was an intern and KM was a researcher at IBM Research. Corresponding authors: QX (quanx1808@gmail.com), DB (debarunb@us.ibm.com).

indicate that LLMs are overconfident when verbalizing their own confidence [Xiong et al., 2024] and that LLMs' metalinguistic judgments are less reliable than quantities derived directly from token-level probabilities [Hu and Levy, 2023].

UQ approaches, regardless of whether they are verbalized or not, can be categorized into two high-level categories – *black-box* techniques only assume API access to an LLM [Lin et al., 2023, Manakul et al., 2023, Cole et al., 2023], whereas *white-box* approaches require other model information such as the model weights [Gal and Ghahramani, 2016, Xiao and Wang, 2019], or internal states such as embeddings and activation spaces [Ren et al., 2023], or only token-level logits [Kuhn et al., 2022, Kadavath et al., 2022]. Some recent white-box methods for calibrating LLM output typically require substantial training data for fine-tuning LLMs [Chen et al., 2023, Kapoor et al., 2024, Ulmer et al., 2024], making them computationally intensive. In comparison, black-box approaches offer practical advantages, including robustness to LLM upgrades, compatibility with proprietary models (even those limited to API-based inputs and outputs), and computational tractability at inference time. As a result, such approaches have gained popularity in generation tasks.

Several UQ approaches, both black-box and white-box, sample multiple generations from an LLM and characterize some notion of variability between generations as a proxy for confidence. For instance, Kuhn et al. [2022] find semantic equivalence between generations to cluster them and then compute the entropy of each cluster, where a lower entropy value implies a more confident generation. Lin et al. [2023] take a spectral clustering approach instead by treating generations as nodes in a graph, but also rely on variability. Various other approaches leverage consistency between generations via different forms of perturbation [Gao et al., 2024], for addressing challenges such as ambiguous questions [Cole et al., 2023] or detecting hallucinated facts [Manakul et al., 2023].

Recent efforts have recommended combining ideas from the various UQ categories, such as an approach that estimates a numeric confidence score for any LLM output by combining an extrinsic consistency-based metric over multiple samples with an intrinsic confidence estimate obtained by prompting the LLM itself [Chen and Mueller, 2023]. Xiong et al. [2024] suggest that using both verbalized confidence and sampling consistency could be promising for attaining more accurate confidence assessments in LLMs.

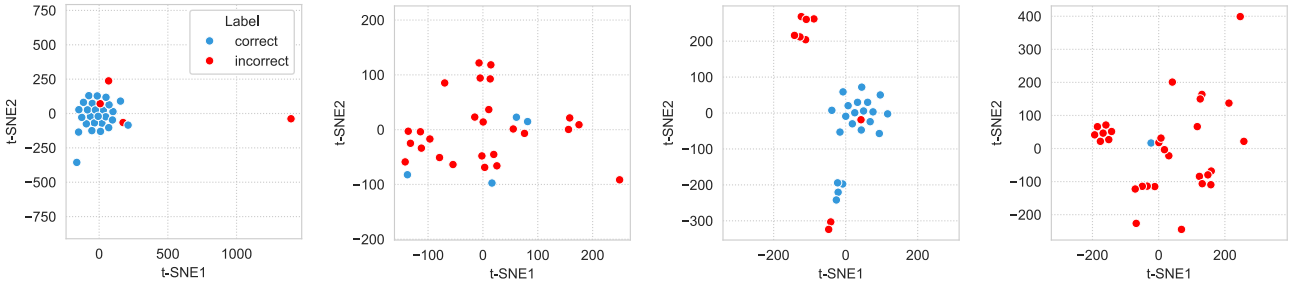
There is an implicit assumption behind all approaches that generate multiple samples and use consistency or variability to infer confidence – informally, when a generated response is more different than others, it is more likely to be incorrect, thus responses that are consistently similar are more likely to be correct. We refer to this assumption as the *consistency hypothesis*, and visualize it using Figure 1, which considers two instances each from two datasets – CoQA [Reddy et al., 2019] for open-ended question answering and Spider Realistic [Deng et al., 2021] for the text-to-SQL generation task. We represent 30 generations for each instance via t-SNE projections of their semantic encodings, distinguishing between generations deemed to be correct and incorrect through color coding. For instances across both types of datasets, we observe that correct generations tend to be visually closer to other generations and to each other as compared to incorrect generations (Figures 1a, 1c). Incorrect generations tend to lie on the border of the correct generations. Furthermore, when there are many incorrect generations, these are often largely spread across the representation space and may be dissimilar (Figures 1b, 1d). This is merely an illustrative visual examination for specific instances, but how can this crucial assumption that forms the basis for so many UQ approaches be formalized and tested? Importantly, **how can one measure the extent to which the consistency hypothesis holds for LLM generations for a dataset**?

In this paper, we continue the burgeoning investigation of consistency-based methods for UQ in LLMs. Our specific **contributions** are as follows:

- We propose a rigorous statistical procedure (as outlined in Algorithm 1) for testing the implicit assumption behind consistency-based UQ approaches for LLMs using notions of similarity between generations. We propose both conceptual and actionable versions of the hypothesis stated verbally, and formulate them as comparisons between sets of similarities.
- We propose metrics to measure the extent to which various versions of the consistency hypothesis hold for LLM generations for a task and dataset.
- We conduct a detailed empirical investigation spanning 8 benchmark datasets over 3 tasks – question answering, text summarization, and text-to-SQL – to verify the extent to which the hypotheses hold for these datasets.
- Based on the empirical findings, we design a black-box UQ method associated with new aggregation functions that performs reasonably well for confidence estimation, further demonstrating the practical value of exploiting the consistency hypothesis for UQ and motivating future exploration into related approaches.
- In addition, our illustrative visualization tools such as those depicted in Figures 1 and 2 can aid data exploration pertaining to the consistency hypothesis (see Appendix B for further examples).

2 CONSISTENCY HYPOTHESIS FORMALIZATION

We begin by clarifying the problem setup with notation, and subsequently propose and formalize various versions of the consistency hypothesis.



(a) QA case 1 (b) QA case 2 (c) Text-to-SQL case 1 (d) Text-to-SQL case 2

Figure 1: T-SNE projections of 30 generations represented by sentence embeddings, for 2 instances each from the CoQA and Spider Realistic dev datasets. Correct and incorrect generations are labeled in blue and red respectively. The plots suggest that correct generations tend to cluster closely to other generations and to each other as compared to incorrect generations, whereas incorrect generations are often more dispersed and exhibit greater dissimilarity.

2.1 PRELIMINARIES & NOTATION

Consider a dataset with queries $x_1, \cdots, x_n$, where x_i is the ith query (a.k.a. instance, such as a question in a question answering task). Consistency-based approaches call an LLM for multiple generations (a.k.a. samples, such as answers in a question answering task) for each instance. Generations corresponding to x_i are denoted $y_i^1, \cdots, y_i^m$. We use $[n] := \{1, \cdots, n\}$ and $[m] := \{1, \cdots, m\}$ as shorthand to denote the set of indices for queries and generations, respectively.

We assume there is a set of responses $\mathcal{Y}_i^*$ that are deemed *correct* for query x_i. Therefore, a generated output y_i^j is correct if $y_i^j \in \mathcal{Y}_i^*$. For tasks such as open-ended question answering and summarization, a generation for a particular input may be viewed as correct if a metric such as Rouge-L with respect to the ground truth output exceeds some predetermined threshold.

We also assume access to a *similarity* metric $s(\cdot)$ defined on the interval $[0, 1]$, to compute pairwise similarities $s_i^{j,k} = s(y_i^j, y_i^k)$, which is the similarity between the jth and k^{th} samples for the ith instance in the dataset. In addition, an *aggregation* function $f(\cdot)$ is required for combining pairwise similarities into a scalar.

2.2 STATEMENTS & FORMALIZATIONS

We propose three different verbal statements as candidates for the consistency hypothesis, and formalize them as mathematical statements. These formal statements could either be at the "instance-level", i.e. at the level of a particular query involving multiple generations, or at the "group-level", i.e. applicable over a set of queries in the dataset. As a preview, we visualize the distributions of two similarity sets, denoted S^C (correct cluster) and S^I (incorrect cluster) in Figure 2; these are keys for formalizing the consistency hypothesis and will be defined subsequently. We highlight that these distributions are clearly distinct across all three verbal hypotheses proposed. Furthermore, the mean of the correct set notably exceeds that of the incorrect set, which will form the basis for our formal statements.

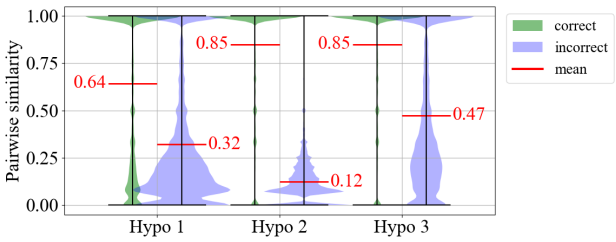


Figure 2: Visualization of the distributions of the Jaccard pairwise similarities of correct cluster S^C and incorrect cluster S^I under different hypotheses from Granite generations of 30 samples each on the CoQA dataset. Complete visualizations for each dataset can be found in Figure 14.

Motivated by Figure 1, we first introduce two natural hypotheses that compare generations with either other correct generations or those within the same category. We then propose a more actionable hypothesis designed for the practical scenario where generation labels are unavailable.

2.2.1 The Sim-Correct Hypothesis

One of the natural verbal statements inspired by Figure 1 is as follows: *A correct generation is more similar to other correct generations for an instance, as compared to an incorrect generation*. We call this statement the *sim-correct* hypothesis, as it makes a claim about similarity only with respect to *correct* generations.

We consider two variants for formalization: one considers similarity of a generation with respect to other individual generations, and the other aggregates similarities with respect to other generations.

Pairwise Consistency. First we consider pairwise similarities between generations. We build two sets of pairwise

similarities for each instance; one is for correct generations and one for incorrect generations: $S_i^C = \{s_i^{j,k} : j, k \in \mathcal{Y}_i^*, k \neq j\}$ and $S_i^I = \{s_i^{j,k} : j \notin \mathcal{Y}_i^*, k \in \mathcal{Y}_i^*\}$ where $k \in [m]$. We construct these sets by selecting the appropriate entries from the matrix formed by computing pairwise similarities between all generations.

The formal statement can now be made as: $\mu_i^C > \mu_i^I$, where these are the means of sets S_i^C and S_i^I respectively. The statement could be verified using a one-sided t-test by rejecting the null hypothesis of $\mu_i^C \leq \mu_i^I$ at a suitable p-value.

Note that the above statements are instance-level as there is a statement for each i-th instance. There may however not be enough data for statistical significance at this level. Group level versions of the sets can be obtained by collecting over all instances in a group: $S_g^C = \bigcup_{i \in G} \{s_i^{j,k} : j, k \in \mathcal{Y}_i^*, k \neq j\}$ and $S_g^I = \bigcup_{i \in G} \{s_i^{j,k} : j \notin \mathcal{Y}_i^*, k \in \mathcal{Y}_i^*\}$, where group g is associated with indices of queries $G \subseteq [n]$ and $k \in [m]$. The formal statement at the group level is $\mu_g^C > \mu_g^I$. When $G = [n]$, there is only 1 group and the statement is at the "dataset level". We denote the correct and incorrect sets at the dataset level as S^C and S^I, respectively.

Aggregated Consistency. Here we consider the aggregated similarity over all other samples. Besides similarity metric $s(\cdot)$, this requires an aggregation function $f(\cdot)$. In this case, we use aggregated similarities instead of pairwise ones to gauge how a generation compares with other generations: $S_i^C = \{\bar{s}_i^j : j \in \mathcal{Y}_i^*\}$ and $S_i^I = \{\bar{s}_i^j : j \notin \mathcal{Y}_i^*\}$, where $\bar{s}_i^j = f(s_i^{j,i_1}, \cdots, s_i^{j,i_m})$ with $\{s_i^{j,i_1}, \cdots, s_i^{j,i_m}\} \in \mathcal{Y}_i^*$ is the aggregated similarity of generation j w.r.t other correct generations for instance i. The formal statement remains $\mu_i^C > \mu_i^I$ using means for the aforementioned sets, and once again one can consider either instance-level statements or those collecting over instances in a group.

The pairwise consistency hypothesis offers more data points for conducting statistical tests, while the aggregated consistency hypothesis is more practical for designing new UQ methods with improved aggregation functions.

2.2.2 The Sim-Separate Hypothesis

According to Figure 1, another candidate verbal statement is: *A correct generation is more similar to other correct generations than an incorrect generation is to other incorrect generations, for an instance*. We refer to this statement as the *sim-separate* hypothesis as its assertion entirely separates correct and incorrect generations.

The main difference between the formalization for this hypothesis and the previous one is how instance-level similarity sets are defined. Here, $S_i^C = \{s_i^{j,k} : j, k \in \mathcal{Y}_i^*, k \neq j\}$ and $S_i^I = \{s_i^{j,k} : j, k \notin \mathcal{Y}_i^*, k \neq j\}$ where $k \in [m]$. Note that the aggregated consistency version of this hypothesis involves aggregating similarities for a correct generation over the set of all correct generations, and for an incorrect generation over the set of all incorrect generations, for a given instance. All other aspects are the same as Sim-Correct.

2.2.3 The Sim-Any Hypothesis

When labels for correct and incorrect generations are unavailable, as is typical, one black-box UQ approach for confidence estimation [Lin et al., 2023] is to leverage this aggregation function directly, e.g. confidence in y_i^j could be computed as $c_i^j = f(s_i^{j,1}, \cdots, s_i^{j,m})$ and interpreted as the probability of the generation being correct, potentially after further calibration. Thus, while conceptually intuitive, the previous two hypotheses are insufficient for providing practical guidance in designing black-box UQ methods. We then propose a novel formalization of the consistency hypothesis by comparing each generation with all other generations, eliminating the need for labels during aggregation.

A verbal statement for this version of the consistency hypothesis is as follows: *A correct generation is more similar to other generations for an instance, as compared to an incorrect generation*. We refer to this statement as the *sim-any* hypothesis, as the statement makes a claim about similarity with respect to *any* other generation or generations, not necessarily those that are correct or incorrect.

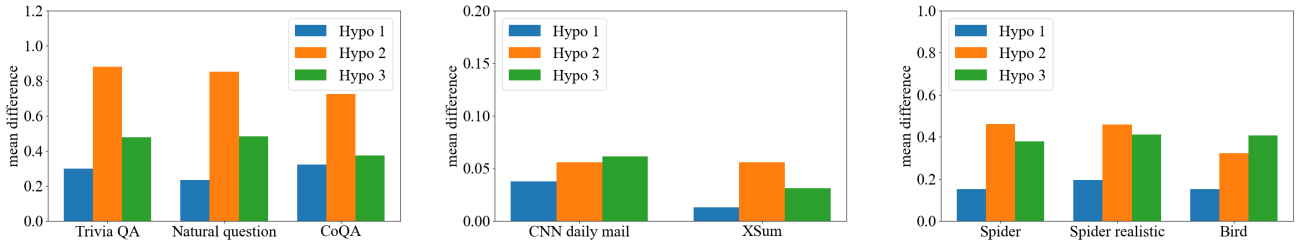
For this hypothesis, instance-level similarity sets are constructed as: $S_i^C = \{s_i^{j,k} : j \in \mathcal{Y}_i^*, k \neq j\}$ and $S_i^I = \{s_i^{j,k} : j \notin \mathcal{Y}_i^*, k \neq j\}$ where $k \in [m]$. The aggregated consistency version of this hypothesis performs aggregation only over the set of all generations for an instance; All other aspects remain the same as Sim-Correct. As the Sim-Any hypothesis is less intuitive from Figure 1, it should be carefully examined by statistical tests. If verified to some extent, the confidence in a generation could potentially be estimated by aggregating its similarity with all other generations.

In subsequent sections (including figures), we also refer to the Sim-Any, Sim-Correct, and Sim-Separate hypotheses as Hypo #1, #2, and #3, respectively, for brevity.

3 CONSISTENCY HYPOTHESIS VERIFICATION

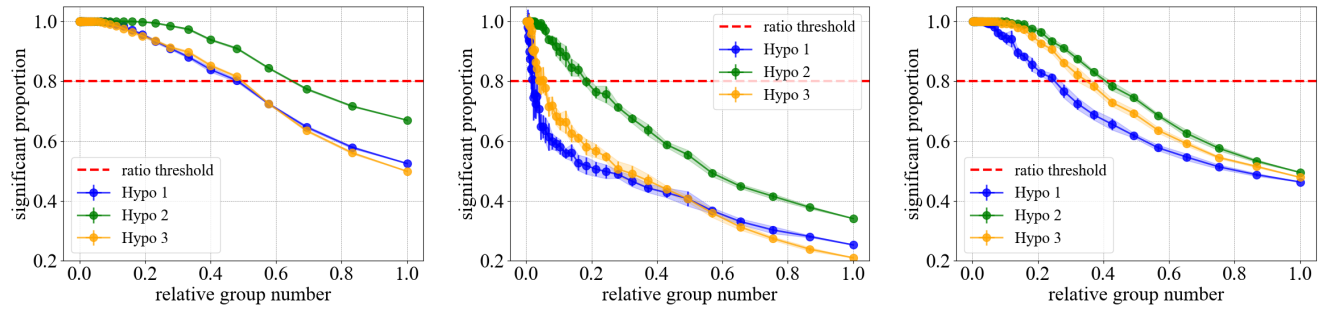
How can one measure the extent to which a dataset conforms to a consistency hypothesis? We propose the following metrics that leverage the formal statements. The metrics rely on Algorithm 1, which is provided the number of groups n_g and returns the fraction of groups that verify the hypothesis, denoted $\rho(n_g)$, as determined by whether the null hypothesis is rejected at p-value ≤ 0.05.

Mean difference $\Delta\mu = \mu^C - \mu^I$. This measures the extent to which the statement holds true across the entire dataset but is problematic as a measure of the efficacy of



(a) QA (b) Summarization (c) Text-to-SQL

Figure 3: Verification of hypotheses using mean difference $\Delta\mu$ between similarity sets with Jaccard pairwise similarity on all 8 datasets for the QA, summarization, and text-to-SQL tasks.



(a) CoQA dataset (b) XSum dataset (c) Spider dataset

Figure 4: Verification of hypotheses (pairwise consistency statements) using the fraction of verified groups $\rho(n_g)$ as a function of the relative number of groups n_g/n with Jaccard pairwise similarity on representative datasets for QA, summarization, and text-to-SQL tasks. Complete results for 8 datasets are included in the Appendix.

consistency-based UQ methods, which necessarily need to be applied at the instance level; this is because it compiles similarity sets using all instances and only applies at the dataset level. There is likely to be sufficient data to test the hypothesis at this level.

Fraction of instances where the hypothesis is verified $\rho(n)$. This is an instance level measure that calls upon Algorithm 1 with a group for each instance, i.e. $n_g = n$. Note that there may not be sufficient data at the instance level for testing the statistical significance of a statement.

Maximum relative number of groups for a thresholded fraction of verified groups: $\theta^* = n_g^*/n$ for $n_g^* = \arg\max_{n_g} I(\rho(n_g) \geq \rho^*)$, where $I(\cdot)$ is the indicator function, $\rho(n_g)$ is the fraction of verified groups obtained from Algorithm 1, ρ^* is some user-specified threshold %, and n is the number of instances in the dataset. We use the relative number of groups $\theta = n_g/n$ for comparability across datasets. This metric is intended to balance the other two metrics by considering both statistical power (significant proportion) and practicality (relative group number) for the purpose of conducting consistency-based UQ. Moreover, this metric is a summary quantity that effectively represents the 'height' of the trade-off curve of statistical power and practicality using a single measure — a higher curve means that a hypothesis is more valid for that dataset. Although we selected $\rho^* = 80\%$ as a threshold in experiments for quantification purposes, our conclusions are based on a comparative analysis of the curves rather than hinging upon this specific value.

These metrics serve as tools for evaluating the verification extent of different consistency hypotheses based on various similarity metrics $s(\cdot)$ and aggregation functions $f(\cdot)$. They provide essential building blocks for comparing metric performance and inspiring the development of new black-box UQ methods based on the verification extent.

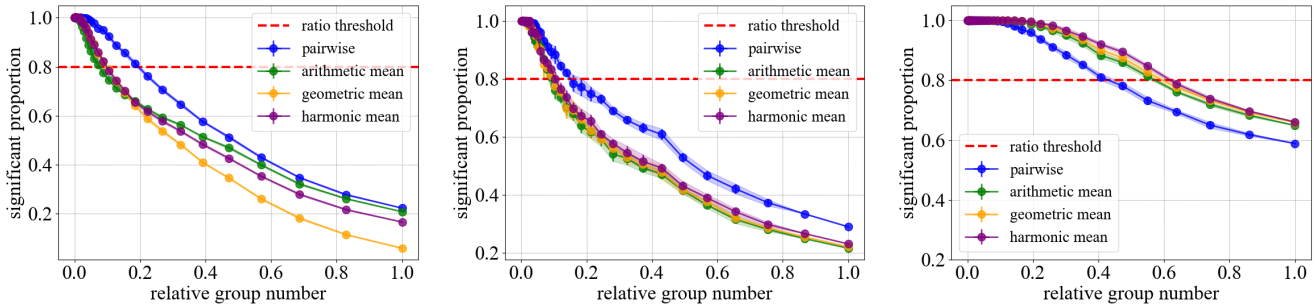
4 EMPIRICAL INVESTIGATION

We conduct experiments for various generative tasks, covering several datasets and generations with representative LLMs to tackle those datasets, with the intent of understanding the extent to which the hypotheses hold and how they could be leveraged for UQ.

QA Task. Following prior work on UQ [Kuhn et al., 2022, Lin et al., 2023], we use the open-book conversational question answering dataset CoQA [Reddy et al., 2019], the closed-book QA dataset TriviaQA [Joshi et al., 2017], as well as the more challenging closed-book QA dataset Natural Questions (NQ) [Kwiatkowski et al., 2019]. We consider the corresponding dev sets, consisting of 7983, 9960, and 3610 questions for CoQA, TriviaQA, and NQ respectively.

Aggregation/ Similarity	Mean difference when $n_g = 10 \uparrow$				Max relative group number when $\rho^* = 0.6 \uparrow$			
	Jaccard	Rouge-1	Rouge-L	Sbert	Jaccard	Rouge-1	Rouge-L	Sbert
pairwise	0.2247 ±0.0009	0.2329 ±0.0008	**0.2331 ±0.0007**	0.1713 ±0.0005	0.1838 ±0.00002	0.1992 ±0.00002	0.1994 ±0.00003	**0.2194 ±0.00003**
arithmetic	0.0816 ±0.0011	0.0897 ±0.0010	0.0898 ±0.0012	0.0858 ±0.0005	0.0464 ±0.00009	0.0587 ±0.00003	0.0548 ±0.00005	0.1096 ±0.00012
geometric	0.1437 ±0.0010	0.1729 ±0.0009	0.1729 ±0.0011	0.1127 ±0.0004	0.1176 ±0.00002	0.1453 ±0.00001	0.1450 ±0.00001	0.1532 ±0.00002
harmonic	0.1417 ±0.0010	0.1722 ±0.0011	0.1722 ±0.0008	0.1425 ±0.0004	0.0410 ±0.00002	0.1468 ±0.00007	0.1536 ±0.00003	0.1714 ±0.00009

Table 1: Ablation study on choice of similarity metric and aggregation functions for the Natural Question dataset. Higher numbers indicate stronger verification of the consistency hypothesis H1.



(a) Trivia QA dataset (b) CNN daily mail dataset (c) Bird dataset

Figure 5: Verification of hypotheses using the fraction of verified groups $\rho(n_g)$ as a function of the relative number of groups n_g/n with Jaccard pairwise similarity and three different aggregation functions on 3 representative datasets for QA, summarization, and text-to-SQL tasks. Complete results for 8 datasets are included in Figure 12 in the Appendix.

We generate responses using **Granite 13B** [Mishra et al., 2024] (default) and **LLaMA 2 70B** [Touvron et al., 2023].

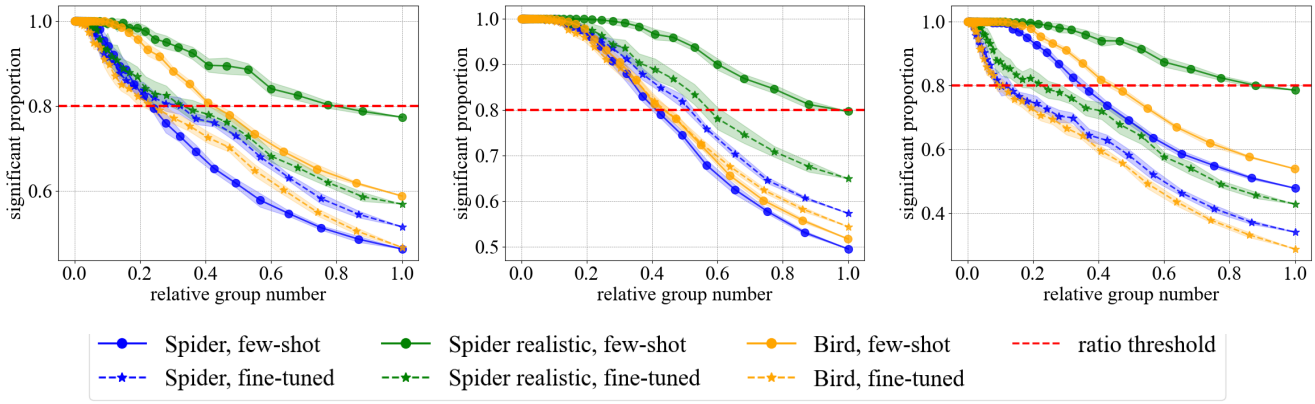
Summarization Task. We use the CNN DailyMail Version 3.0.0 [See et al., 2017, Hermann et al., 2015] and Extreme Summarization (XSum) [Narayan et al., 2018] datasets. For each dataset, we generate summaries using **LLaMA 3 8B** [Meta, 2024] (default) and **Mistral 8x7B** [MistralAI, 2023] for the first 1000 documents.

Text-to-SQL Task. We consider various real-world benchmark datasets for the task of converting natural language queries to SQL. Spider [Yu et al., 2018] is a popular benchmark that covers 138 domains, and Spider-Realistic [Deng et al., 2021] is a more challenging version that modifies the natural language queries to avoid explicit mention of column names. BIRD [Li et al., 2024] is a recent cross-domain benchmark of 95 databases covering more than 37 professional domains. We use the dev sets for all 3 datasets, which include 1034, 508, and 1533 queries respectively. We generate SQL with **Codellama 34 B** [Rozière et al., 2024] (default) and a **Deepseek 33 B** model [Guo et al., 2024] that is further fine-tuned with LoRA [Hu et al., 2022] using Spider's training dataset.

Other Experimental Details. We produce 5 samples each over 6 temperatures (from 0.25 to 1.5 in increments of 0.25) from all LLMs, so as to obtain sufficient variability across generations [Zhu et al., 2024]. An exploration of other sampling approaches and their ablation is provided in the Appendix. Most experiments use the Jaccard coefficient as similarity metric $s(\cdot)$, but we also consider variations of ROUGE metrics such as Rouge-1 and Rouge-L, and the cosine similarity between sentence BERT `https://sbert.net/` (sbert) representations of the generations. Wherever relevant, we consider the arithmetic mean of similarities as the aggregation function $f(\cdot)$, but we also consider simple extensions such as the geometric and harmonic mean. For QA and summarization, a generated response is deemed correct if the Rouge-L score with respect to ground truth is no less than 0.5 and 0.2 respectively; Rouge-L has been used in this fashion for QA in prior work [Kuhn et al., 2022, Lin et al., 2023]. Our results are robust with respect to Rouge-L thresholds that define correctness of generations, as noted by an ablation study in Appendix A.5. For text-to-SQL, a generated SQL is correct if it returns the same result as the ground truth SQL upon query execution on the underlying database. We run Algorithm 1 with $R = 10$ repetitions.

4.1 PAIRWISE CONSISTENCY STATEMENTS

We begin by verifying the pairwise consistency hypotheses 1–3 (denoted H1–H3) using the Jaccard similarity measure



(a) Hypo 1 (Sim-Any) (b) Hypo 2 (Sim-Correct) (c) Hypo 3 (Sim-Separate)

Figure 6: Impact of different datasets and models on verifying Hypothesis 1–3 for the text-to-SQL task.

Algorithm 1 Fraction of verified groups $\rho(n_g)$

Input: Similarity matrix of generations, number of groups n_g, number of repetitions R
for each repetition r from 1 to R **do**
  Split dataset randomly into n_g groups
  Initialize $n_v \leftarrow 0$
  for each group g in n_g **do**
    Construct similarity sets S_g^C and S_g^I
    Conduct a t-test on the means
    Calculate the p-value
    if p-value ≤ 0.05 **then**
      $n_v \leftarrow n_v + 1$
    end if
  end for
  Compute $\rho_r(n_g) \leftarrow \frac{n_v}{n_g}$
end for
Compute $\rho(n_g) \leftarrow \frac{1}{R}\sum_{r=1}^{R} \rho_r(n_g)$
Compute the error bar of $\rho(n_g)$ as the variance
Output: $\rho(n_g)$ with the error bar

across various datasets on the three tasks.

We show the mean difference of correct similarity set S^C and incorrect set S^I in Figure 3 across diverse datasets and tasks. All mean differences are positive, indicating that all hypotheses are validated to some extent, but it is largest for H2. Recall that H2 suggests that compared to incorrect generations, correct generations are more similar to other correct generations. The validation of H3 indicates that incorrect generations have higher variability and are more diversely distributed than the correct generations. Note that in practical scenarios, one does not know the correct and incorrect labels in advance. Therefore, despite the stronger validation of H2 and H3, H1 serves as a practical approximation and its validation provides guidance for confidence estimation of each generation using similarity with respect to all other generations.

In Figure 4, we illustrate the verification of all hypotheses at the group level using a plot of significant proportion (a.k.a. the fraction of verified groups) vs. the relative number of groups. We showcase only one representative dataset for each task in Figure 4 and defer plots for other datasets to Figure 8 in the Appendix. Conclusions drawn regarding the correctness of the hypotheses at the dataset level remain applicable at the group level. We also provide similar plots for the hypotheses using Rouge-L similarity in Figure 9 in the Appendix. The trend is the same as those using Jaccard.

4.2 AGGREGATED CONSISTENCY STATEMENTS

Next we investigate the correctness of the aggregated consistency statements and impact of different aggregation functions across diverse datasets on three tasks: QA, summarization, and text-to-SQL. We provide the verification plots of the aggregated version of three hypotheses in Figure 10 in the Appendix. All hypotheses are again verified to some extent. We choose H1 to study the effect of different aggregation functions versus the pairwise similarity statement due to its practical implications for UQ.

We show results for representative datasets in Figure 5 and defer others to Figure 12 in the Appendix. The conclusions for QA and summarization tasks are identical – since there is less data available for the aggregated consistency statement than the pairwise one, the aggregated hypothesis tends to be statistically less significant. However, this seems to flip for the text-to-SQL task, especially for the harmonic mean. This suggests that aggregated consistency statements may be particularly effective in summarizing useful information from pairwise similarities, at least for some tasks.

4.3 ABLATION STUDY: SIMILARITY METRIC & AGGREGATION FUNCTION

In this section, we conduct an ablation study of similarity measures and aggregation functions for the verification

Dataset (Model)		CoQA (Granite)		Spider (Codellama)		TriviaQA (Granite)	
		AUROC ↑	AUARC ↑	AUROC ↑	AUARC ↑	AUROC ↑	AUARC ↑
Baselines	always 1	0.5±0.0	0.62±0.003	0.5±0.0	0.21±0.009	0.5±0.0	0.35±0.005
	avg. prob	0.75±0.004	0.79±0.006	0.62±0.012	0.26±0.015	0.74±0.002	0.51±0.005
	spec-ecc	0.18±0.008	0.37±0.003	0.31±0.008	0.13±0.006	0.20±0.002	0.17±0.002
	p(True)	0.59±0.026	0.70±0.015	0.54±0.009	0.23±0.011	0.68±0.005	0.47±0.007
Black-box (Agg. sims)	arith-agg	0.82±0.006	0.82±0.007	0.74±0.006	0.34±0.013	**0.80**±0.002	**0.57**±0.004
	geom-agg	**0.85**±0.008	**0.83**±0.007	**0.76**±0.011	**0.35**±0.017	**0.80**±0.003	**0.57**±0.004
	harm-agg	**0.85**±0.008	**0.83**±0.008	**0.76**±0.010	**0.35**±0.016	**0.80**±0.003	**0.57**±0.004

Table 2: Comparing different aggregation approaches with Jaccard similarities for black-box UQ. Error bars for AUROC and AUARC are computed over 5 runs that randomly choose 50% of the dataset for testing.

of H1. The results for QA and text-to-SQL tasks in terms of both mean difference and maximum relative number of groups to achieve ρ^* amount of verified group are shown in Tables 1 and 3. Due to space limitations, we defer Table 3 to the Appendix. For similarity measures, we test Jaccard, Rouge-1, Rouge-L, and the cosine similarity between sbert representation for both tasks, and another metric that identifies the type of SQL output [Pourreza and Rafiei, 2024] for the text-to-SQL datasets. For aggregation functions, we use the arithmetic, geometric and harmonic means.

The key insights from these two tables are: 1) similar to the conclusion in Section 4.2, aggregation degrades the verification extent in the QA task, whereas it enhances validation of the hypothesis in the text-to-SQL task; 2) the hypothesis is more true under Rouge-1 and Rouge-L scores than Jaccard similarity across both tasks; 3) Jaccard, Rouge-1, Rouge-L are robust in both tasks; 4) sbert similarity occasionally outperforms others in the max. relative group number to achieve a certain fraction of verified groups; 5) SQL output type similarity is least reliable for verifying the hypothesis in the text-to-SQL task. Furthermore, since Jaccard similarity is the most computationally efficient measure besides being the most robust, we use it for most experiments.

4.4 ABLATION STUDY: DATASETS & MODELS

The correctness of a hypothesis is a property of both datasets and models. Here we investigate the effect of different datasets and models on the validation of pairwise statements for the text-to-SQL task as shown in Figure 6. We refer the Codellama and the fine-tuned Deepseek model as the few-shot and fine-tuned model, respectively. The Spider Realistic dataset is one where the consistency hypothesis holds to the greatest extent, followed by Bird and then Spider. We observe that all the hypotheses are stronger for few-shot models as compared to fine-tuned models, for both the Spider Realistic and Bird datasets. This may occur because incorrect instances may have less variability after fine-tuning the models for specific tasks.

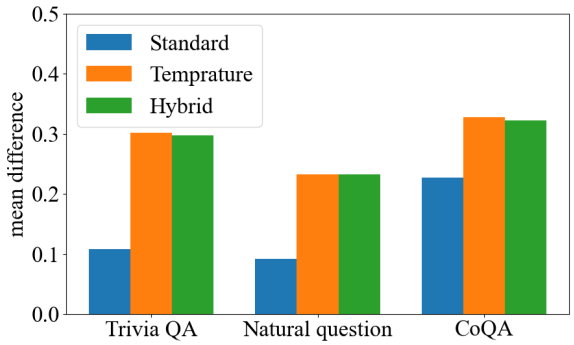


Figure 7: Effects of sampling approaches to the verification of H1 on the QA task with Granite model.

4.5 ABLATION STUDY: SAMPLING METHOD

We evaluate the consistency hypothesis for the QA task using three sampling methods: 1) standard sampling, where tokens are generated through decoding at a fixed temperature of 0.25, 2) temperature sampling, which involves generating one sample at varying temperatures, and 3) hybrid sampling, where multiple samples are generated from multiple temperatures (this is what we used in other experiments). Figure 7 displays the mean difference between correct and incorrect clusters at the dataset level. Sampling at different temperatures (temperature and hybrid sampling) increases the variability of generations, thereby enhancing the validity of the consistency hypotheses.

5 CONFIDENCE ESTIMATION BY AGGREGATION

To demonstrate how rigorous statistical validation of the consistency hypothesis can guide UQ method design, we introduce a new UQ method based on the 'Sim-Any' hypothesis (H1) as an illustrative example.

Given that the consistency hypothesis H1 is generally veri-

fied to various degrees across datasets and models in Section 4, we investigate the use of simple aggregation methods including arithmetic, geometric and harmonic mean as a way of estimating confidence. Specifically, we treat the aggregated similarity of generation j w.r.t other generations $\bar{s}_i^j = f(s_i^{j,1}, \cdots, s_i^{j,m})$ as its confidence. Note that the 'degree' approach used after spectral clustering, as discussed in [Lin et al., 2023], corresponds to what we describe here as arithmetic mean aggregation. In contrast, the geometric and harmonic means are novel aggregation methods introduced in this work, inspired by their superior empirical performance on verifying consistency hypothesis in Table 1. Both methods yield more conservative confidence estimates, typically lower than those produced by the arithmetic mean.

We evaluate confidence using two standard metrics. The *Area Under the Receiver Operating Characteristic* (AUROC) evaluates confidence when used as a probabilistic classifier for correctness, whereas the *Area Under the Accuracy Rejection Curve* (AUARC) is suitable when confidences are used for selective generation, i.e. for rejecting a fraction of the instances that one is least confident about.

Table 2 compares various UQ approaches over 2 datasets for QA (CoQA, TriviaQA) and 1 for text-to-SQL (Spider). We consider 4 baselines: a naive one that always returns a score of 1, a white-box approach that uses the avg. log probability of tokens, a spectral clustering technique using 'eccentricity' [Lin et al., 2023], and a white-box verbalized approach that prompts an LLM to determine whether a generation is correct and uses the logit of the answer (True/False) [Kadavath et al., 2022]. Results suggest that the black-box UQ methods with the geometric and harmonic means outperform the baselines. This highlights the practical significance of our work around verifying the consistency hypothesis and demonstrates that the extent of verification can inform the design of effective new UQ methods.

6 CONCLUSION

In this paper, we proposed three mathematical formulations to formalize a common conjecture about consistency in LLM generations. We substantiated these hypotheses through a comprehensive empirical investigation on eight benchmark datasets (CoQA, TriviaQA, Natural Questions, CNN DailyMail, XSum, Spider, Spider-Realistic, and BIRD) spanning three tasks (QA, Summarization and Text-to-SQL), establishing justification for consistency-based UQ methods. We conducted a thorough ablation study, analyzing factors such as aggregation methods, sampling techniques, datasets, and models to validate the consistency hypotheses. Inspired by the empirical findings, we also proposed a black box UQ method with new aggregation approaches over similarities and demonstrated its superior performance over some baselines, thereby highlighting how our extensive empirical study on consistency can inform practice. An investigation of other aggregation methods as well as a broader empirical study with tasks involving additional complexities are potential avenues for future work.

ACKNOWLEDGMENTS

This work was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration.

References

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in LLMs. In *Conference on Empirical Methods in Natural Language Processing*, pages 5190–5213, 2023.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*, 2023.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Conference on Empirical Methods in Natural Language Processing*, pages 530–543, 2023.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. Structure-grounded pretraining for text-to-sql. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

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

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. International Conference on Machine Learning*, pages 1050–1059, 2016.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 2336–2346, 2024.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li,

Fuli Luo, Yingfei Xiong, and Wenfeng Liang. Deepseek-coder: When the large language model meets programming – the rise of code intelligence, 2024.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proc. Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations*, 2022.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, 2023.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1601–1611, 2017.

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

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proc. International Conference on Learning Representations*, 2022.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. In *Proc. Advances in Neural Information Processing Systems*, volume 36, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Meta. Introducing meta llama 3: The most capable openly available LLM to date. `https://ai.meta.com/blog/meta-llama-3/`, 2024. Accessed: 2024-06-14.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami Seelam, Brian Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David D. Cox, Ruchir Puri, and Rameswar Panda. Granite code models: A family of open foundation models for code intelligence, 2024.

MistralAI. Mixtral of experts. `https://mistral.ai/news/mixtral-of-experts/`, 2023. Accessed: 2024-06-14.

Allan H Murphy and Edward S Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology and Climatology*, 6(5):748–755, 1967.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Mohammadreza Pourreza and Davood Rafiei. DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction. In *Proc. Advances in Neural Information Processing Systems*, volume 36, 2024.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *Proc. International Conference on Learning Representations*, 2023.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

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

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. Calibrating large language models using their generations only. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, 2024.

Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In *Proc. Association for the Advancement of Artificial Intelligence*, volume 33, pages 7322–7329, 2019.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. 2024.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, 2018.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Zhi Jin, and Hong Mei. Hot or cold? Adaptive temperature sampling for code generation with large language models. In *Proc. Association for the Advancement of Artificial Intelligence*, volume 38, pages 437–445, 2024.

The Consistency Hypothesis in Uncertainty Quantification for Large Language Models (Supplementary Material)

Quan Xiao[*1] **Debarun Bhattacharjya**[2] **Balaji Ganesan**[2] **Radu Marinescu**[2] **Katsiaryna Mirylenka**[3]
Nhan H Pham[2] **Michael Glass**[2] **Junkyu Lee**[2]

[1]Rensselaer Polytechnic Institute [2]IBM Research [3]Zalando

Part

Table of Contents

A Additional Experimental Results **12**
A.1 Verification of Pairwise Consistency Statements 12
A.2 Verification of Aggregated Consistency Statements 12
A.3 Ablation Study: Similarity Metric & Aggregation Function 13
A.4 Validation of Hypotheses over Non-default Models 13
A.5 Ablation Study: Rouge-L Score 14

B Visualization of the Distributions of Similarity Sets **14**

A ADDITIONAL EXPERIMENTAL RESULTS

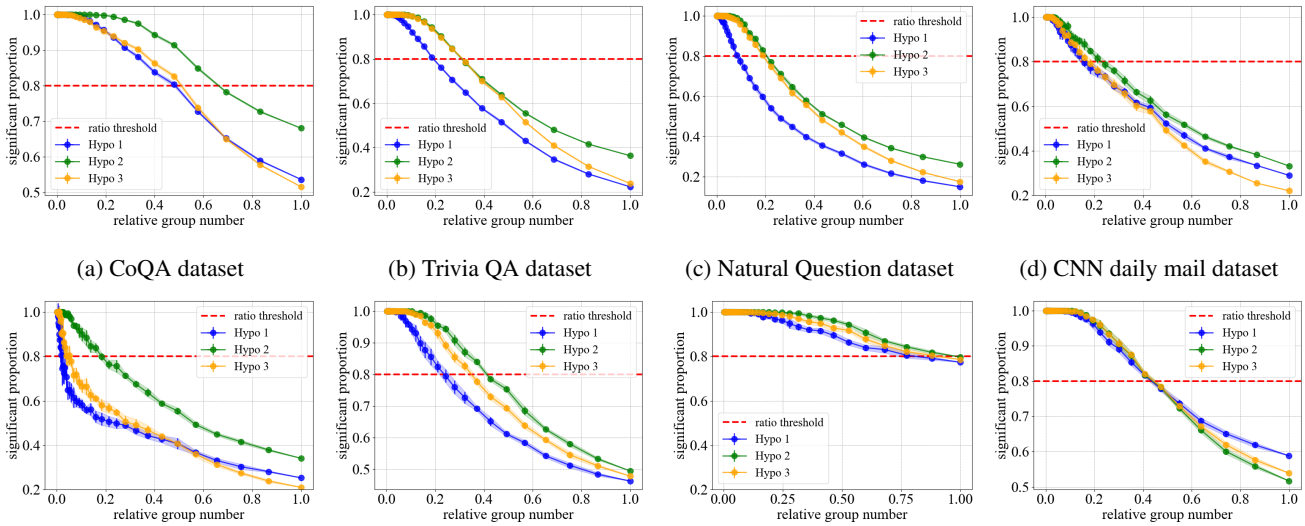
In this section, we include the omitted figures and tables from the main paper.

A.1 VERIFICATION OF PAIRWISE CONSISTENCY STATEMENTS

In this section, we provide the complete verification results for the pairwise consistency hypotheses for all the datasets in Figure 8. We also provide the validation of the pairwise consistency hypotheses using Rouge-L similarity in Figure 9. All of the hypotheses are verified to some extent across all of the datasets. Among those, consistency hypotheses are more true on the QA and text-to-SQL tasks than the summarization task. This may due to the higher viability of correct answers in the text summarization task.

A.2 VERIFICATION OF AGGREGATED CONSISTENCY STATEMENTS

In this section, we provide the complete verification results for the aggregated consistency hypotheses with Jaccard similarity for all the datasets in Figure 10 and Figure 12. It can be seen that all three hypotheses are validated to some extent in all of the datasets. Surprisingly, while the aggregate consistency hypotheses generally exhibit lower validity compared to the pairwise version, especially in QA and summarization tasks due to reduced available similarity pairs, this trend does not hold for the text-to-SQL task. Notably, the harmonic mean aggregation achieves the highest validation performance, indicating its potential as an effective method for classifying the generations based on consistency.



(a) CoQA dataset (b) Trivia QA dataset (c) Natural Question dataset (d) CNN daily mail dataset

(e) XSum dataset (f) Spider dataset (g) Spider Realistic dataset (h) Bird dataset

Figure 8: Verification of three hypotheses using Jaccard pairwise similarity on various datasets for QA, summarization, and text-to-SQL tasks.

Aggregation/ Similarity	Mean difference when $n_g = 10 \uparrow$					Max relative group number when $\rho^* = 0.8 \uparrow$				
	Jaccard	Rouge-1	Rouge-L	Sbert	SQL	Jaccard	Rouge-1	Rouge-L	Sbert	SQL
pairwise	0.1519 ±0.0004	0.1395 ±0.0002	0.1482 ±0.0003	0.0815 ±0.0001	0.099 ±0.0003	0.2472 ±0.0003	0.2820 ±0.0001	0.2728 ±0.0004	0.2847 ±0.0002	0.1269 ±0.0002
arithmetic	0.1625 ±0.0003	0.1964 ±0.0002	0.1956 ±0.0002	0.1410 ±0.0001	0.0656 ±0.0005	0.3261 ±0.0003	0.3933 ±0.0001	0.3854 ±0.0002	**0.4144 ±0.0003**	0.0361 ±0.0001
geometric	0.1365 ±0.0003	0.2081 ±0.0008	0.2007 ±0.0005	0.1653 ±0.0001	0.2010 ±0.0013	0.2858 ±0.0003	0.2951 ±0.0004	0.2983 ±0.0003	0.0010 ±0.0001	0.1348 ±0.0001
harmonic	0.1317 ±0.0003	**0.2161 ±0.0005**	0.2058 ±0.0004	0.1928 ±0.0462	0.2012 ±0.0010	0.3710 ±0.0001	0.3558 ±0.0005	0.3648 ±0.0001	0.0034 ±0.0001	0.1386 ±0.0001

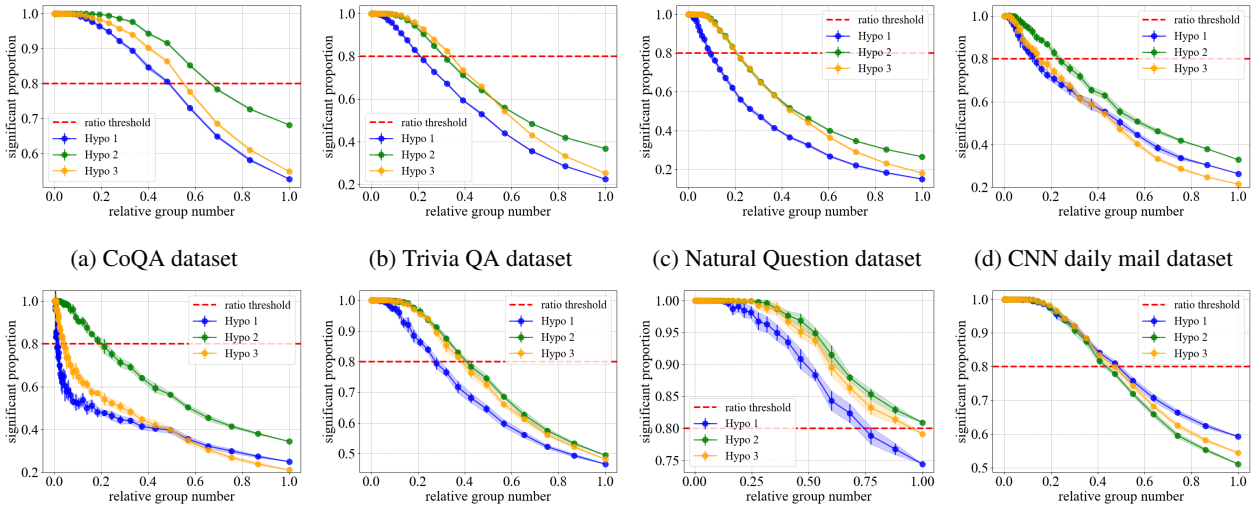
Table 3: Ablation study of similarity metric and aggregation functions on Spider dataset.

A.3 ABLATION STUDY: SIMILARITY METRIC & AGGREGATION FUNCTION

In addition to Figure 12, in this section, we include the ablation study w.r.t the similarity metric and aggregation function on Spider dataset with Codellama model in Table 3. In Spider dataset, some aggregation methods outperform the pairwise similarity, especially for Rouge-1 and Rouge-L score. The SQL output type similarity exhibits the least capability in capturing the viability of incorrect clusters, resulting in the least propensity for validating the consistency hypothesis. This may suggest the poor behavior of SQL output type similarity when applying to UQ methods.

A.4 VALIDATION OF HYPOTHESES OVER NON-DEFAULT MODELS

To show the robustness of the validation of the proposed hypotheses across different LLM models, we plot the mean difference of correct and incorrect clusters on the dataset level for the non-default models (i.e. LLaMA 2 70 B for QA task, Mistral 8x7B for text summarization task, and fine-tuned Deepseek 33B model for text-to-SQL task) in Figure 11. Similar to the results for the default models, all of the mean difference are positive, meaning that all of the consistency hypotheses are true in some sense. Among those, H2 is more true than H3, followed by H1. This suggests that while the validation extent varies across models, the means of the similarity sets for correct generations always exceed those of incorrect ones, indicating that the proposed consistency hypotheses hold to some degree across different models.



(a) CoQA dataset (b) Trivia QA dataset (c) Natural Question dataset (d) CNN daily mail dataset

(e) XSum dataset (f) Spider dataset (g) Spider Realistic dataset (h) Bird dataset

Figure 9: Verification of different hypotheses with Rouge-L pairwise similarity on QA, summarization, and text-to-SQL tasks for 8 datasets.

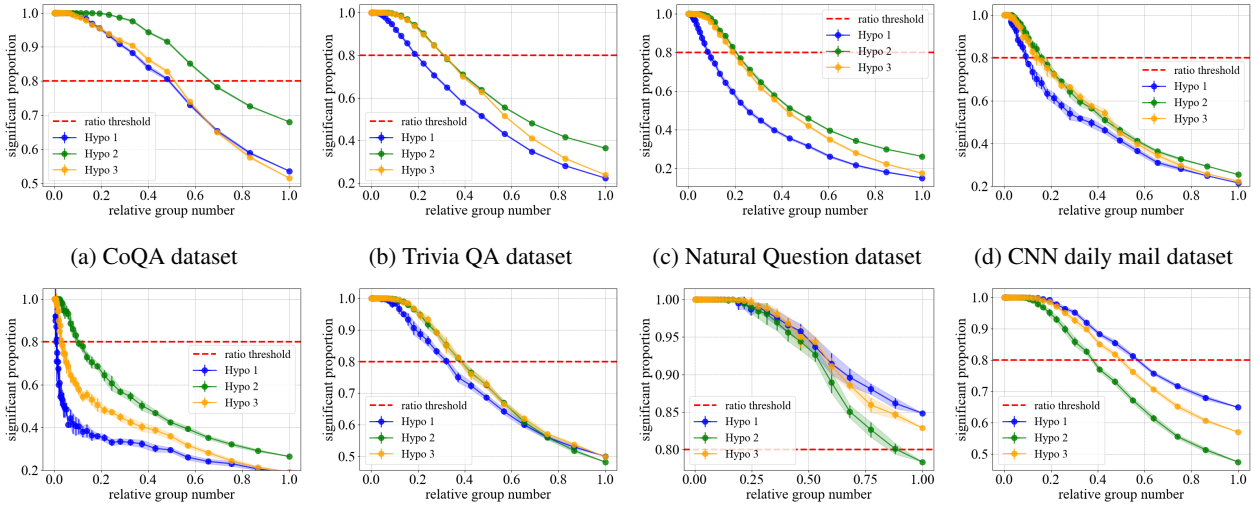
A.5 ABLATION STUDY: ROUGE-L SCORE

As the correctness of generations for QA and summarization task depend on the thresholds for Rouge-L score, we conduct an ablation study over this threshold in this section. The results are shown in Figure 13, demonstrating that the extent of validation for all three versions of the hypotheses are robust to the choice of Rouge-L threshold.

B VISUALIZATION OF THE DISTRIBUTIONS OF SIMILARITY SETS

In this section, we visualize the distributions of correct similarity sets S^C and incorrect similarity sets S^I across various datasets in QA, summarization, and text-to-SQL tasks in Figure 14. The mean values of pairwise similarity sets for correct generations consistently surpass those of incorrect ones across all tasks. The differentiation between these two clusters varies by task. In QA tasks, owing to the short responses, S^C is concentrated around 1, while S^I centers around 0. In text-to-SQL tasks, S^C tends to have higher values, whereas S^I is skewed towards lower values. The most challenging task from this perspective is text summarization, where even correct answers exhibit diverse distributions. However, S^C still maintains a higher mean compared to S^I.

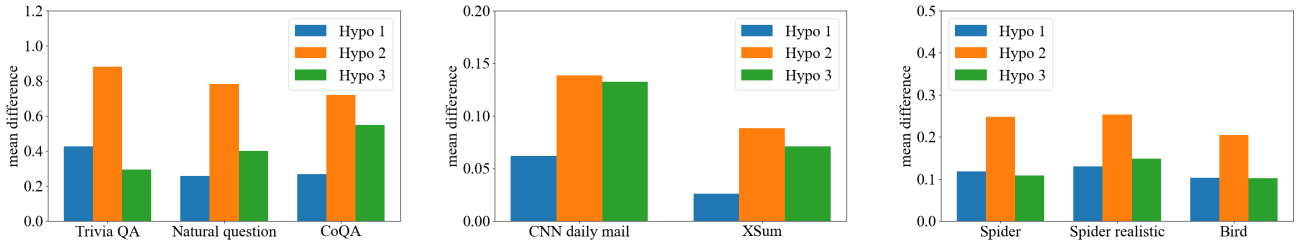
This visualization of the distributions of the pairwise similarity set provides insights into the inherent difficulty levels of different tasks. It is instructive in further classifying generations as correct or incorrect through clustering based on their similarities with other generations.



(a) CoQA dataset (b) Trivia QA dataset (c) Natural Question dataset (d) CNN daily mail dataset

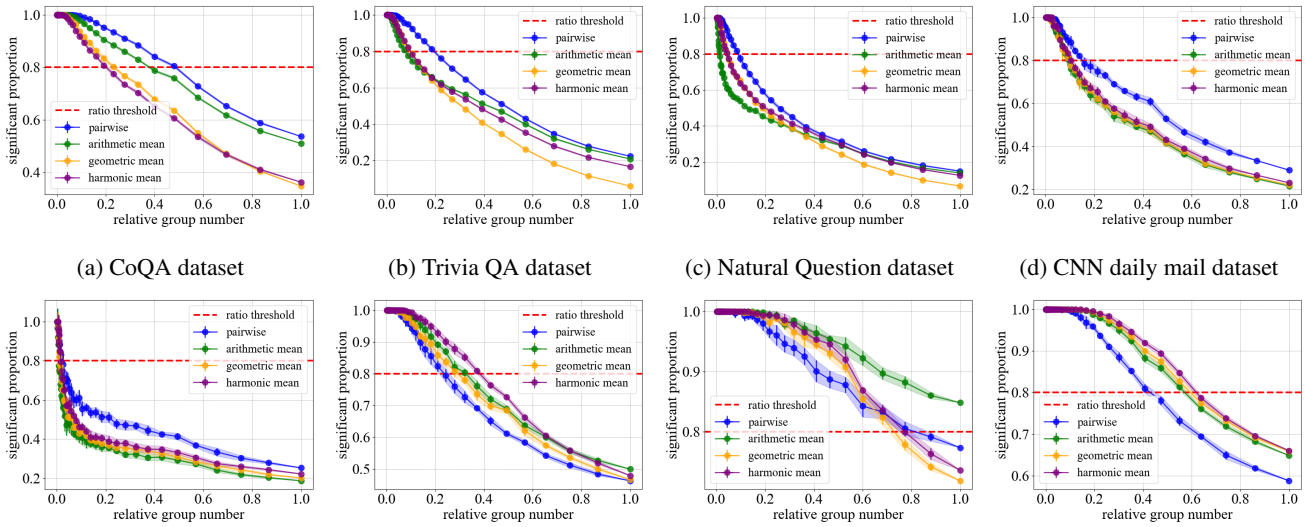
(d) XSum dataset (d) Spider dataset (e) Spider Realistic dataset (f) Bird dataset

Figure 10: Verification of different hypotheses using Jaccard similarity and arithmetic mean as aggregation function on QA, summarization, and text-to-SQL tasks for 8 datasets.



(a) QA (b) Summarization (c) Text-to-SQL

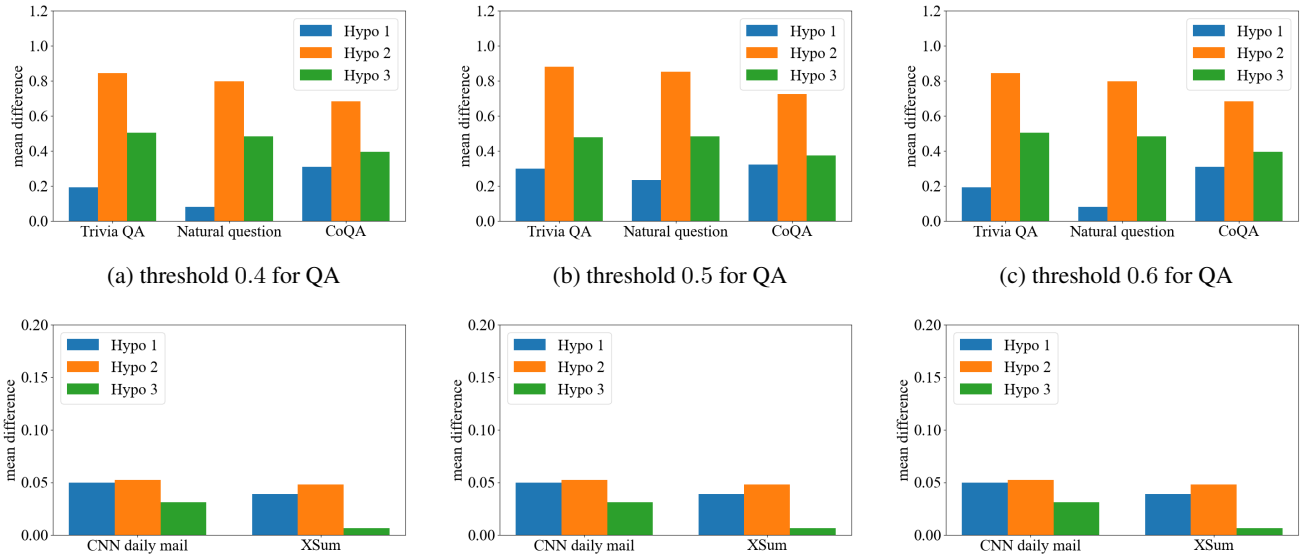
Figure 11: Verification of hypotheses for non-default models using mean difference $\Delta\mu$ between similarity sets with Jaccard pairwise similarity on all datasets for the QA, summarization, and text-to-SQL tasks for 8 datasets.



(a) CoQA dataset (b) Trivia QA dataset (c) Natural Question dataset (d) CNN daily mail dataset

(e) XSum dataset (f) Spider dataset (g) Spider Realistic dataset (h) Bird dataset

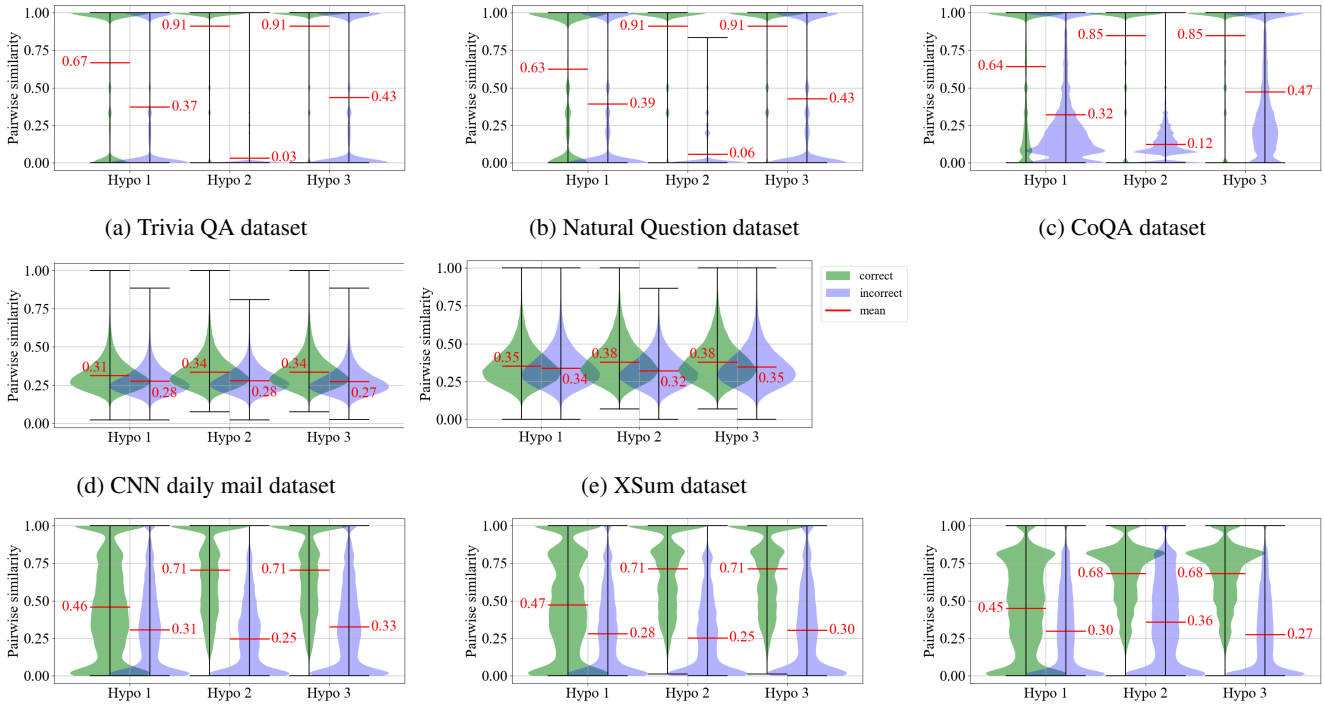
Figure 12: Impact of aggregation functions for verifying aggregation version of Hypothesis 1 on different datasets in QA, summarization, and text-to-SQL tasks for 8 datasets.



(a) threshold 0.4 for QA (b) threshold 0.5 for QA (c) threshold 0.6 for QA

(a) threshold 0.1 for summarization (b) threshold 0.2 for summarization (c) threshold 0.3 for summarization

Figure 13: Verfication of hypotheses under varying Rouge-L score thresholds for the correctness of generations using mean difference $\Delta\mu$ between similarity sets with Jaccard arithmetic similarity on all 5 datasets for the QA and summarization tasks.



(a) Trivia QA dataset (b) Natural Question dataset (c) CoQA dataset

(d) CNN daily mail dataset (e) XSum dataset

(f) Spider dataset (g) Spider Realistic dataset (h) Bird dataset

Figure 14: Visualization of the distributions of the Jaccard pairwise similarities of correct cluster S^C and incorrect cluster S^I under different hypotheses on the datasets of QA, summarization, and text-to-SQL tasks.