SUMIE: A Synthetic Benchmark for Incremental Entity Summarization

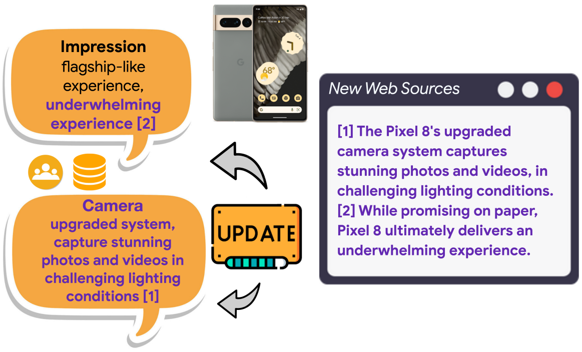
Anonymous COLING 2025 submission

Abstract

No existing dataset adequately tests how well language models can incrementally update entity summaries – a crucial ability as these models rapidly advance. The Incremental Entity Summarization (IES) task is vital for maintaining accurate, up-to-date knowledge. To address this, we introduce SUMIE, a fully synthetic dataset designed to expose real-world IES challenges. This dataset addresses issues like incorrect entity association and incomplete information, capturing real-world complexity by generating diverse attributes, summaries, and unstructured paragraphs with 99% alignment accuracy between generated summaries and paragraphs. Extensive experiments demonstrate the dataset's difficulty – state-of-the-art LLMs struggle to update summaries with an F1 higher than 80.4%. We will open-source the benchmark and the evaluation metrics to help the community make progress on IES tasks.

1 Introduction

Entity Summarization (ES) distills key features of entities (e.g., people, places, organizations) from extensive unstructured data, essential for various NLP applications like question answering (Allam and Haggag, 2012), information retrieval (Kowalski, 2007), and entity comparison systems (Gunel et al., 2023). Traditional ES tasks focus on computing concise summaries for entities, drawing on a size-limited selection of triples (subject-predicate-object statements) within structured RDF data (Liu et al., 2020b, 2021). This work goes further, creating precise and comprehensive structured summaries for entities by leveraging the vast knowledge available in natural language on the web. Structured summaries in various domains, including hotels and restaurants, simplify the comparison of detailed options, helping people make choices that align with their preferences.


Impression
flagship-like experience, underwhelming experience [2]
Camera
upgraded system, capture stunning photos and videos in challenging lighting conditions [1]
UPDATE
New Web Sources

[1] The Pixel 8's upgraded camera system captures stunning photos and videos, in challenging lighting conditions. [2] While promising on paper, Pixel 8 ultimately delivers an underwhelming experience.

Figure 1: Overview of the Incremental Entity Summarization Task. Existing attribute ("Impression") can be updated and new attribute ("Camera") can be augmented.

With the growing amount of information, it's important to update structured summaries automatically. Incremental Entity Summarization (IES) addresses this by enabling updates to entity summaries with new information (Chowdhury et al., 2024), ensuring accurate and comprehensive representation in search engines, as shown in Figure 1. Moreover, IES enables efficient management of the vast and rapidly changing data. Despite its critical importance in organizing massive amounts of information, IES is underexplored. While some work (Goasdoué et al., 2019; Yang et al., 2021; Chowdhury et al., 2024) investigates updating entity summaries using abstractive or extractive techniques, these efforts often lack structured attribute-value organization or suffer from hallucination problems of LLMs. Crucially, there is no dataset specifically designed to test the ability of these models to maintain accurate, up-to-date entity knowledge.

To develop an effective dataset for IES systems, it needs a broad selection of entities with diverse and evolving attributes and values with varied writing styles. Moreover, it requires accurate alignment between web documents and their structured summaries to trace attribute values to their sources. While diverse natural language web sources for various entities are readily available (Ganesan and Zhai, 2012; Asghar, 2016), creating well-maintained and evolving structured summaries from these sources remains both expensive and time-consuming, requiring extensive human verification (Gunel et al., 2023; Chowdhury et al., 2024).

In this paper, we propose a synthetic dataset that captures real-world complexity using LLM by leveraging the empirical finding that LLMs excel at expanding short phrases into descriptive, contextual paragraphs, rather than abstractly summarizing all important components from a longer text. The dataset generation uses a structured approach with LLMs: It begins with producing diverse attributes, values, and paragraphs, and progresses to generating incrementally updated entity summaries. High quality is ensured through an LLM critic, achieving 99% accuracy in human evaluations. In essence, we propose a carefully crafted synthetic dataset designed to be high-quality and complex, effectively simulating real-world scenarios.

Our contributions are as follows:

- We present SUMIE, the first dataset built with high informativeness and diversity for rigorous evaluation of incremental entity summarization methods. We open-source SUMIE to accelerate research in this field, including metrics of evaluation.
- We propose simple but effective LLM-based solutions, **Update** and **Merge** for IES task. These methods provide valuable baselines for future advancements.
- We conduct insightful analyses to pinpoint the limitations of LLM-based entity summarization methods. State-of-the-art LLMs struggle to update summaries with an F1 score higher than 80.4%, highlighting the inherent complexity of this task.

2 Dataset Desiderata

To build a dataset ideal for developing entity summarization systems with incremental generation capability, we outline the following key desiderata:

Diversity of Entities. The dataset should encompass a broad spectrum of entities across domains. This could include businesses (restaurants, hotels), products, events, and more. Diverse entities ensure the model encounters a wide choices of attributes and associated values, expanding its knowledge base.

Complexity of Attributes and Values. Values associated with attributes should demonstrate variation in length, sentiment and subjectivity. Even within the same entity category, attribute values should reflect high diversity to challenge the models' nuanced understanding. Likewise, attributes must range common (e.g. a restaurant's service) to niche and specific interests (e.g. a hiking trail's access to restrooms).

Varied Information Sources. The textual sources should exhibit a rich diversity of real-world styles and origins. Generate a mixture of editorial reviews (which often analyze with authority), user generated contents (informal and potentially biased, found in online forums or social media), and official product descriptions (which frequently use persuasive language focused on features and benefits). Exposing the model to different writing styles and purposes will compel it to adapt to various language patterns.

Inclusion of Misleading Information The dataset should contain subtly misleading details that requires contextual understanding for identification. The goal is to challenge the model's ability to analyze information within the provided context rather than simply relying on basic fact-checking.

Incremental Information Updates. The dataset should include examples where information about an entity evolves over time, simulating updates as new facets or perspectives are revealed. This forces the model to not only add new information but also potentially revise or re-prioritize existing facts. Introducing situations where initial information is incomplete or later contradicted by more supported sources. The model must learn to prioritize well-supported information over time, mirroring a common real-world scenario where our understanding of a subject develops.

Rigorous Alignment between Structured Summaries and Natural Language Paragraphs. Ensure a precise and traceable connection exists between a source paragraph and its corresponding structured summary (i.e. an attribute-value table). Focus on maintaining clear attributions, and ensure the origin of each value is precisely derived from the source paragraph. Avoid introducing information into the structured summary that isn't explicitly supported by the text for a rigorous alignment.

3 Dataset Generation Methodology

We create a synthetic dataset with generated attributes, entity names, and incrementally evolving summary tables (see Figure 2). Accompanying paragraphs mirror the tables, including distracting sentences. We used `Gemini-Ultra` with a temperature setting of 0.8 to generate the dataset. See all LLM prompting instructions in Appendix A.2.

3.1 Attribute and Entity Name Generation

We begin by selecting 20 popular categories (e.g. `Accomodations`) (see Appendix A.1 for all category information). For each, we prompt an LLM to generate attributes (e.g. `Room Quality`) and entity names (e.g. `Canyon Hotel`). To ensure attribute diversity, we retrieve up to 50 common (e.g. `Room quality`) and 50 less-common attributes (e.g. `Honeymoon packages`) typically used to describe entities within that category. For entity names, we generate up to 40 plausible but fictitious names, randomly selecting 10. Each entity is then assigned 30 attributes, with an equal split between common and uncommon descriptors. This process results in a dataset containing 200 entities, which we consider *suitable* for evaluation. The use of random elements in the generation process helps reduce the impact of LLM bias on the dataset. In the final dataset, entity names are replaced with generic ones (e.g., from `Canyon Hotel` to `HOTEL1`) to avoid any unintended claims related to real-world entities.

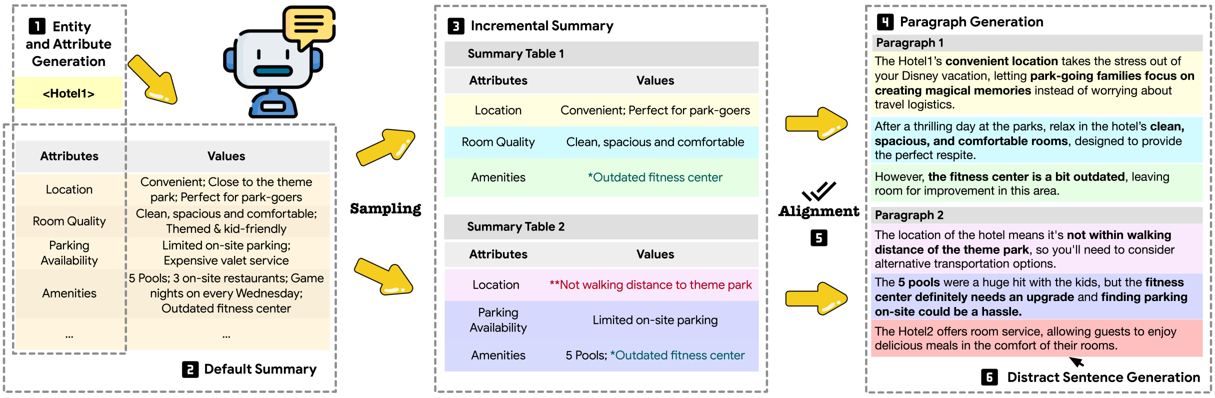
3.2 Summary Table Generation

Default summary table generation. Summary tables provide a structured representation of attributes associated with an entity in a given category. Each row details an attribute and its corresponding value. The goal in this stage is to generate values that meet three criteria: 1) Informative and meaningful, covering both subjective and objective aspects, 2) Diverse in length (one to 10 words), and 3) Varied in sentiment (positive, negative, and neutral). We generate at least three descriptive values per sentiment, resulting in three distinct summary tables for each entity. For instance, when the prompt specifies a positive sentiment, the model is directed to generate favorable descriptors such as "`Spacious and comfortable`" and "`Clean`" for a designated attribute like "`Room Quality`". The final summary tables for each entity combine up to 10 attribute and value pairs, including varied sentiments derived from 3 separate summaries for each entity.

Incremental summary table generation. To assess the LLM's incremental update capabilities, we generate multiple summary tables per entity. The initial summary is the basis from which we sample attributes and values for incremental versions. To simulate real-world scenarios where information evolves, we ensure two criteria are met: 1) Repetition of attributes and values across summaries, and 2) The presence of conflicting attribute information. Conflicting values can be generated by prompting an LLM to produce values that directly oppose the meanings of originally sampled values. We iteratively create K summaries and each iteration combines half the attributes from a previous summary with half from the unused attribute pool, resulting in K summary tables per entity with diverse and potentially contradictory content.

3.3 Paragraphs

Paragraph generation. Building upon the incrementally generated summary tables (Sec. 3.2), we craft aligned paragraphs for each. The fundamental goal is to incorporate all attributes and values from a given table into the text. Additionally, we prioritize diverse writing styles, avoiding overly simplistic language. To achieve this, we define 8 writing categories, including user reviews, official product descriptions, editorial insights, and discussions on online forums, and 6 tones, including optimistic, neutral, pessimistic, sarcastic, humorous, and analytic. Each paragraph is randomly assigned a category and a tone, which guide its genera-



Attributes	Values
Location	Convenient; Close to the theme park; Perfect for park-goers
Room Quality	Clean, spacious and comfortable; Themed & kid-friendly
Parking Availability	Limited on-site parking; Expensive valet service
Amenities	5 Pools; 3 on-site restaurants; Game nights on every Wednesday; Outdated fitness center
...	...

Summary Table 1

Attributes	Values
Location	Convenient; Perfect for park-goers
Room Quality	Clean, spacious and comfortable
Amenities	*Outdated fitness center

Summary Table 2

Attributes	Values
Location	**Not walking distance to theme park
Parking Availability	Limited on-site parking
Amenities	5 Pools; *Outdated fitness center

Paragraph 1

The Hotel1's **convenient location** takes the stress out of your Disney vacation, letting **park-going families focus on creating magical memories** instead of worrying about travel logistics.

After a thrilling day at the parks, relax in the hotel's **clean, spacious, and comfortable rooms**, designed to provide the perfect respite.

However, **the fitness center is a bit outdated**, leaving room for improvement in this area.

Paragraph 2

The location of the hotel means it's **not within walking distance of the theme park**, so you'll need to consider alternative transportation options.

The **5 pools** were a huge hit with the kids, but the **fitness center definitely needs an upgrade** and **finding parking on-site could be a hassle.**

The Hotel2 offers room service, allowing guests to enjoy delicious meals in the comfort of their rooms.

Figure 2: Dataset Generation Methodology Overview: (1) Generate entity names (masked for ethics consideration) and attributes. (2) Create default summary table with diverse values. (3) Sample attributes/values for incremental summaries (* repeated sampling, ** conflicting values). (4) Generate paragraphs with varying tones based on attributes/values. (5) Verify summary table/paragraph alignment. (6) Add distractor sentence. Note that attribute values and sentences in the same color should be aligned and bold texts in paragraphs are the evidences for corresponding attribute values.

tion. We also integrate citation numbers that directly link each sentence to the attribute-value pair it reflects in the summary table. This process results in 7 paragraphs per entity, showcasing a variety of styles, tones, and embedded citations for easy reference.

Paragraph-Summary table alignment verification. While the sentences in paragraphs are created based on summary tables, the generated paragraphs do not guarantee that all values are reflected in sentences. To make sure that sentences include the attribute-value pairs in the given summary table, we break paragraphs down into sentences and LLM verifies if the attribute-value pair (e.g. (`Room quality`, `Clean`)) is accurately represented in each sentence. If the value is correctly included and its meaning is not misrepresented (e.g. `With its impeccable clean rooms...`), no change is needed. If the value is missing or misrepresented (e.g. `While its cleanliness of the rooms are debatable...`), the sentence should be adjusted to incorporate the value accurately.

After the automated critique and revision step, we performed a human evaluation of all sentences across all paragraphs, totaling more than 11K sentences, along with their corresponding summary tables. Three human annotators checked for misaligned attribute-value pairs in the paragraphs based on the summary tables. Our dataset achieved 98.7% accuracy, where 98.2% of samples reached a 100% agreement rate. This high accuracy, coupled with a strong human agreement ratio, reaffirms the effectiveness of our automated critique and revision process. More details of the human verification can be found in Appendix A.4.

3.4 Distracting Sentences

After ensuring paragraph-summary table alignment, where all sentences contain attribute-value pairs, we introduce distractor sentences to test the LLM's focus. Since LLMs perform well in finding relevant contexts, we need to challenge their ability to identify and ignore incorrect entity associations. We do this in two ways: first, by generating sentences about irrelevant entities, explicitly including their generic names (e.g. `HOTEL2 boasts a vibrant atmosphere, perfect for...`), and second, by creating metaphorical sentences that describe a human using properties of the given entity's category (always including the word "HUMAN") (e.g. `HUMAN's empathy is a sprawling garden, teeming with vibrant blooms of compassion...`). These distractors allow us to analyze two crucial aspects of LLM performance: entity focus (avoiding irrelevant information) and adjective sensitivity (understanding adjectives even in unrelated contexts).

3.5 Dataset Statistics

We present our dataset statistics for the entity level and paragraph level in Appendix A.1. Overall, the dataset contains 200 entities (20 for each of the 10 categories) and each entity contains an average of 22 attributes and 42 values across all paragraphs, which we believe, achieves *sufficient complexity* for evaluation.

Entities within the same category display a significant amount of diversity. They have approximately 14 distinct attributes (64%) and 41 distinct values (97%) on average. This demonstrates a high degree of variation in their attributes and values. In paragraph statistics (in Appendix A.1), we find that number of "same", "conflict", and "new" attribute values in each paragraph are around 3.7, 3.5, and 2.3, respectively, meaning that same, conflict, and new attribute-value pairs are reasonably distributed across paragraphs. Average number of sentences in paragraphs is 12, with roughly 4 sentences acting as distractors. This indicates that our paragraphs offer sufficient length and incorporate a reasonable amount of distractor sentences.

We show 5 dataset examples in the Appendix from Figure 21 to 25 in 5 categories.

4 Experiments

4.1 Baseline Methods

Our dataset evaluation utilizes two prompt-based approaches, UPDATE and MERGE, designed to assess the LLM's ability to handle new information and conflicts.

Update. LLMs struggle to create comprehensive, high-quality summary tables from large amounts of text due to limited recall (Gunel et al., 2023). We address information overload and reduce the LLM's processing burden by feeding it one paragraph at a time. The first iteration involves generating a summary table from a single paragraph. Afterwards, the LLM receives a new paragraph (potentially containing overlapping, new, or conflicting information) and the previously generated summary table. Its goal is to produce an updated summary table, accurately incorporating relevant details from the new paragraph. Prompts for this method can be found in Appendix A.3.

Merge. This approach breaks down the UPDATE process into two steps, designed to enhance the LLM's understanding. The first iteration remains the same as the UPDATE, with the model generating a summary table from a new paragraph. In later iterations, the model first creates a summary table solely from the new paragraph and then merges it with the existing table. This promotes a clear understanding of the two-step process of retrieving information and updating the summary, potentially reducing the LLM's cognitive load. Prompts for this method can be found in Appendix A.3.

4.2 Evaluation Metrics

We evaluate the performance of the aforementioned approaches to the incremental entity summarization task using precision, recall, and F1. An extraction comprises three components – the attribute, its corresponding value, and the supporting evidence. A successful extraction is one that is also found in the set of goldens corresponding to the input paragraph.

We determine *true positives* via two methods. **Exact matching** checks for a direct match between the predicted value or evidence and the golden set. **LLM-based evidence finding** leverages an LLM to detect if the predicted attribute and value find support within the larger golden set (see Appendix A.3 for prompt). If a predicted extraction fails to match exactly or through the LLM-based evidence prompt, it's marked as a *false positive*. *False negatives* are tracked by noting goldens unmatched to any prediction. While exact matches are simple, LLM-based matches are trickier. The LLM outputs the matched golden row (attribute, value, evidence), but it may not precisely align with the table due to the LLM's generative nature. To address this, we evaluate the cosine similarities between a sentence encoding (we use Universal Sentence Encoder (Cer et al., 2018)) of the response's evidence to the sentence encodings of all the evidences in the golden set to find the highest likelihood golden.

To check its effectiveness in identifying evidence linking predicted and gold-standard attribute values, we manually checked up to 3 paragraphs under the 3 categories, which include more than 210 attribute-value pairs to evaluate. We count incorrectly classified pairs in true positive, false positive, and false negative sets. The `Gemini-Pro` model achieves 90.4% accuracy in evidence detection with a standard deviation of 1% across categories, proving its suitability as an evidence detector between predicted and gold values.

Redundancy and hallucination are crucial metrics requiring evaluation. Redundancy, where models repeatedly extract the same correct value, can artificially inflate F1 scores and hinder fair performance comparisons. Moreover, LLMs are prone to hallucinations, where they generate incorrect values from extracted evidences. Though these hallucinations negatively impact precision and F1 scores, we still want to explicitly measure its severity. For a thorough analysis, we employed two human experts to manually assess these issues within the predicted summary tables (See Sec. 4.4).

4.3 Experimental Setup

We experiment with `Gemini-Pro` (Team et al., 2023), `GPT3.5` (Ouyang et al., 2022), and `Gemini-Nano` (Team et al., 2023) models. The temperatures for all models are set to 0.7. With each entity having 7 paragraphs, we aggregate summary tables iteratively, reporting average precision, recall, and F1 scores across all entities.

4.4 Results and Discussion

Overall performance, Table 1. Table 1 shows overall performance of `Gemini-Pro`, `GPT3.5`, and `Gemini-Nano` on our dataset. At first glance, all models have a large room for improvement, highlighting our dataset's complexity. In particular, `Gemini-Nano` model performs significantly worse than `Gemini-Pro` and `GPT3.5` in both UPDATE and MERGE methods, with an average F1 score gap of 40.3 for UPDATE and 38.6 for MERGE. The performance gap between models is largely due to differences across iterations. `Gemini-Nano`'s recall scores drop significantly from the second iteration, with a decline of over 20 points for both methods. This indicates that as context complexity increases, smaller LLMs struggle to integrate new information effectively. Additionally, Nano has difficulty understanding long prompt instructions, resulting in up to 13% invalid answers, such as repeating the input prompt. In contrast, `Gemini-Pro` and `GPT3.5` produce substantially fewer invalid answers (around 0%) even with complex inputs.

Interestingly, while `Gemini-Pro` and `GPT3.5` show comparable performance, `Gemini-Pro` tends to produce higher precision scores, suggesting that it prioritizes confident and accurate answers. On the other hand, `GPT3.5` achieves better recall scores, indicating that it explores a broader range of answer choices. This becomes more evident in later iterations. While `GPT3.5` model produces relatively stable performance in both precision and recall scores across all iterations, Gemini models exhibit a trade-off between precision and recall scores, prioritizing generating reliable results given a complex context. None of these advanced LLMs exceeded an F1 score of 80.4%, supporting the empirical finding that LLMs excel at generating content, but struggle with abstractly summarizing lengthy texts, consistent with West et al. (2024).

Difference across methods, Table 1. We observe that models perform better with MERGE method than UPDATE approach. This confirms our hypothesis that breaking down UPDATE method into two steps gives a better understanding of our task to LLMs. The MERGE method is particularly beneficial for maintaining recall scores. This is likely because it first extracts attributes and values from the given new paragraph, which are then presented to the model for merging with the existing knowledge. By making the information we want to add explicit in the prompt, the model can more easily make use of the given knowledge.

Difference across categories and tones, Figure 3, 20. Figure 3 presents the F1 scores achieved by the model across different categories, along with their standard deviations. As the figure shows, the model exhibits consistent performance across all categories. There are no significant outliers, implying that the performance of models on our dataset is not biased towards certain categories. Figure 20 in the Appendix shows the performance across paragraph tones and we observe the similar trends to the performance across categories. We also note that standard deviations of `Gemini-Nano` models are considerably larger than those of `Gemini-Pro` models in most cases, reconfirming the challenging nature of our dataset.

Effect of distractor sentences, Table 2, Figure 4. Table 2 shows the performance of `Gemini-Pro` model with UPDATE method after removing distractor sentences in paragraphs. We find that precision scores achieve up to 97 point when the distractor sentences are removed. This proves that our distractor

	Model	Metric	Turns 1	2	3	4	5	6	7	Avg.
UD	Gemini-Pro	Precision	80.0	81.9	82.6	82.5	83.8	84.1	84.3	82.8
		Recall	82.5	76.2	73.2	70.4	69.7	68.4	67.3	72.5
		F1	**80.7**	**78.4**	**77.2**	75.3	75.5	74.8	74.2	**76.6**
	GPT3.5	Precision	78.7	78.2	79.3	79.4	79.8	79.7	80.0	79.3
		Recall	81.6	78.1	75.7	74.8	74.8	74.9	75.1	76.4
		F1	79.5	77.6	77.0	**76.7**	**76.8**	**76.9**	**77.1**	77.4
	Gemini-Nano	Precision	58.7	52.6	49.0	47.0	46.1	46.0	45.9	49.3
		Recall	65.4	43.5	31.0	25.0	21.1	18.6	16.2	31.5
		F1	60.7	46.4	37.2	31.9	28.3	26.0	23.5	36.3
MG	Gemini-Pro	Precision	79.4	79.7	80.4	80.6	80.8	81.8	83.1	80.8
		Recall	82.1	84.0	83.2	82.0	81.1	78.7	74.8	80.8
		F1	**80.1**	81.4	**81.4**	**80.9**	**80.5**	**79.9**	78.3	**80.4**
	GPT3.5	Precision	75.7	76.4	76.1	77.4	76.4	76.2	77.7	76.6
		Recall	83.3	88.3	87.8	85.3	84.2	82.9	82.5	84.9
		F1	78.8	**81.6**	81.3	80.8	79.8	79.1	**79.8**	80.2
	Gemini-Nano	Precision	60.0	51.0	53.7	54.0	56.7	57.9	57.1	55.8
		Recall	66.5	47.4	37.0	32.0	29.6	25.5	22.0	37.1
		F1	62.1	47.9	42.1	38.4	37.5	33.9	30.5	41.8

Table 1: Performance with `Gemini-Pro`, `GPT3.5`, and `Gemini-Nano` models across different turns. UD denote UPDATE and MG denote MERGE. Best F1 scores are in **boldface**.

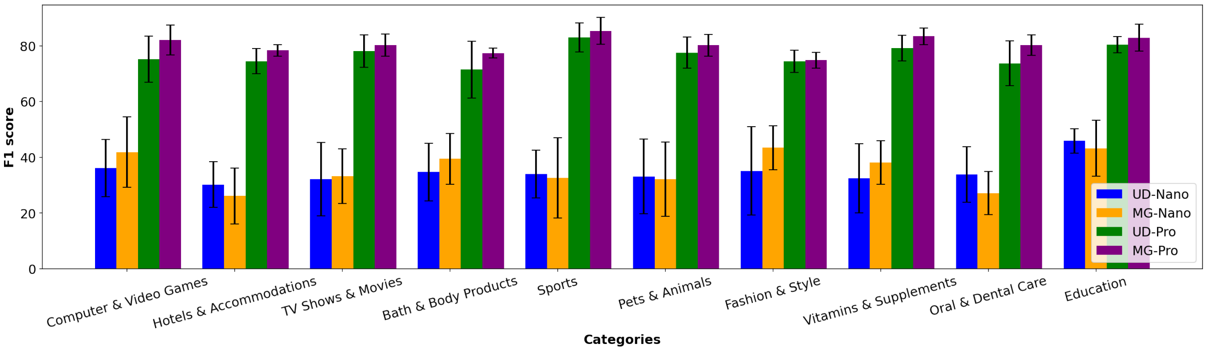


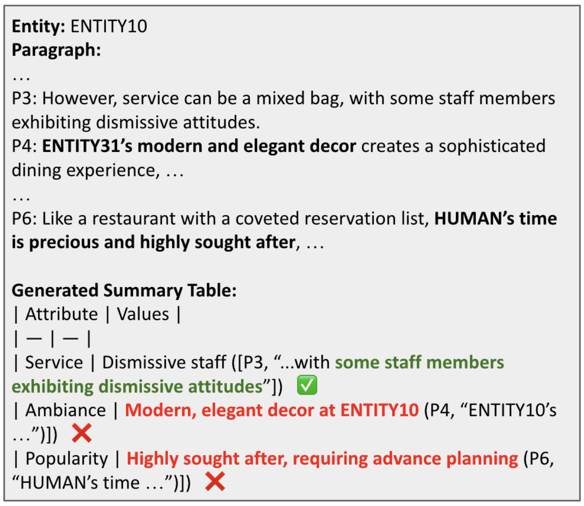
Figure 3: F1 scores across 10 categories (see Appendix A.5 for the rest.).

	Metric	Turns first	last	Avg.
w/ distractor	Precision	80.0	84.3	82.8
	Recall	82.5	67.3	72.5
	F1	80.7	74.2	76.6
w/o distractor	Precision	97.2	96.8	96.8
	Recall	84.3	70.1	74.5
	F1	89.9	81.0	83.8

Table 2: Precision, Recall, and F1 score after removing distractor sentences.

sentences are effectively confusing LLMs and LLMs struggle in strictly focusing on the context relevant to the specific entity. Moreover, it further indicates that our evaluation method based on LLMs works reasonably well in detecting evidence between generated attribute and value pairs and gold attribute, value, and sentence pairs. Figure 4 shows an example of incorrect output from `Gemini-Pro` with distractor sentences. We find that LLMs can easily be misled by information that include several adjective words and also struggle in distinguishing the context crucial to the specific entity.

Human evaluation for checking value redundancy. In addition to F1 scores, we perform two human evaluations to assess how well the model consolidates similar attribute-value pairs (redundancy checking) and to check how well the extracted evidence supports the values. For redundancy checking, two annotators are presented with 30 randomly selected attributes with more than two distinct values generated by `Gemini-Pro`. They indicate 'yes' if the values for each attribute are redundant (e.g., Attribute: `Location`, Values: [`Walking distance from Downtown`, `easy access to Downtown`]), and 'no' otherwise. This evaluation is crucial because an excessive number of similar values for the same attribute can inflate true positives, resulting in artificially high precision and recall scores. We find that, on average, 45% of values are deemed redundant with a 73% agreement rate, indicating that the LLM struggles with identifying



Entity: ENTITY10
Paragraph:
…
P3: However, service can be a mixed bag, with some staff members exhibiting dismissive attitudes.
P4: **ENTITY31's modern and elegant decor** creates a sophisticated dining experience, …
…
P6: Like a restaurant with a coveted reservation list, **HUMAN's time is precious and highly sought after**, …

Generated Summary Table:
| Attribute | Values |
| — | — |
| Service | Dismissive staff ([P3, "...with **some staff members exhibiting dismissive attitudes**"]) ✅
| Ambiance | **Modern, elegant decor at ENTITY10** (P4, "ENTITY10's …")]) ❌
| Popularity | **Highly sought after, requiring advance planning** (P6, "HUMAN's time …")]) ❌

Figure 4: An example of an LLM distracted by irrelevant information.

and merging synonyms into a single value.

Human evaluation for hallucination between value and evidence. Similarly to redundancy checking, two annotators are tasked with assessing the alignment between extracted evidence and attribute values. They are provided with 30 randomly selected attributes, along with their corresponding values and evidences. The annotators mark 'yes' if the attribute and values are supported by evidence, and 'no' otherwise. This allows us to assess the faithfulness of LLMs in extracting evidences to support attribute and values. We find that an average of 25% of the samples are marked as 'no', meaning that evidence does not support the values, with 90% of agreement ratio. An example where the value is not supported by evidence is the attribute-value pair "Guest Privileges" with the value "ability to earn points that can be redeemed for free nights," and the evidence provided by the model is "Your loyalty will be rewarded," where the evidence does not explicitly mention earning points or free nights. This suggests that LLMs often generate attributes and additional details that are not directly supported by the source information.

5 Related Work

Techniques for Incremental Entity Summarization Current ES research has largely focused on summarizing entities from RDF data by selecting key triples (Wei et al., 2019; Liu et al., 2020a, 2021), aiming for compact summaries. Our approach, in contrast, seeks to harness unstructured web text for more comprehensive summaries. While Formal Concept Analysis shows promise in structured knowledge bases (Yang et al., 2021), it struggles with the complexity of web information. Existing datasets (Liu et al., 2020b; Gunaratna et al., 2015, 2016) fall short in testing LLMs' capabilities for web-driven, incremental summary generation. The ENTSUM dataset (Maddela et al., 2022) aids in controllable summarization but is limited in assessing structured or incremental summary creation. Our work broadens the definition of IES, investigating the construction of comprehensive and precise structural summaries using advanced generation models such as LLMs, and introducing a dataset tailored for this innovative field.

Addressing Knowledge Updates and Conflicts The main challenge in IES is enabling LLMs to handle knowledge updates and resolve information conflicts. Solutions like CoverSumm (Chowdhury et al., 2024) and the KNOWLEDGE CONFLICT dataset (Wang et al., 2023) offer ways to update summaries and test conflict resolution, albeit without the necessary complexity for IES. Similarly, the FreshQA benchmark (Vu et al., 2023) tests LLM factuality but doesn't cater specifically to evolving summaries. Our dataset fills this gap, demanding LLMs to identify and adjust to conflicts in entity summaries, with a focus on evidence-based claim reprioritization, aligning closely with the unique requirements of IES.

6 Conclusion and Future Work

In this paper, we introduce SUMIE, a novel benchmark, specifically created to assess the ability of LLMs to generate incremental summaries of entities. SUMIE's synthetic nature ensures data quality and diversity while minimizing the need for extensive human annotations. While our initial baselines demonstrate the dataset's challenges, future work can include: preventing knowledge loss during LLM updates, refining attribute and value recognition to minimize hallucinations, and extending the task to multi-entity comparison summaries. Overall, we aim to spark future research on the task of maintaining up-to-date and comprehensive knowledge.

7 Limitations

Although the evaluation uses three LLMs (including Gemini and GPT-3.5), incorporating additional open-source models would strengthen the findings. Additionally, the chosen LLM-based evaluation metrics can be computationally expensive and time-consuming to execute.

8 Ethics Statement

Our dataset is primarily meant to serve as a diagnostic tool to evaluate LLMs' ability of resolving knowledge conflicts incrementally and generating faithful responses. In addition, the LLMs we used for creating the dataset are trained on a large-scale web corpus and may also bring some bias when generating sentences.

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Nabiha Asghar. 2016. Yelp dataset challenge: Review rating prediction. *arXiv preprint arXiv:1605.05362*.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *Preprint*, arXiv:1803.11175.

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Manzil Zaheer, Andrew McCallum, Amr Ahmed, and Snigdha Chaturvedi. 2024. Incremental extractive opinion summarization using cover trees. *arXiv preprint arXiv:2401.08047*.

Kavita Ganesan and ChengXiang Zhai. 2012. Opinion-based entity ranking. *Information retrieval*, 15(2):116–150.

François Goasdoué, Pawel Guzewicz, and Ioana Manolescu. 2019. Incremental structural summarization of rdf graphs. In *EDBT 2019-22nd International Conference on Extending Database Technology*.

Kalpa Gunaratna, Krishnaparasad Thirunarayan, and Amit Sheth. 2015. Faces: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Kalpa Gunaratna, Krishnaprasad Thirunarayan, Amit Sheth, and Gong Cheng. 2016. Gleaning types for literals in rdf triples with application to entity summarization. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings 13*, pages 85–100. Springer.

Beliz Gunel, Sandeep Tata, and Marc Najork. 2023. Strum: Extractive aspect-based contrastive summarization. In *Companion Proceedings of the ACM Web Conference 2023*, pages 28–31.

Gerald J Kowalski. 2007. *Information retrieval systems: theory and implementation*, volume 1. springer.

Qingxia Liu, Yue Chen, Gong Cheng, Evgeny Kharlamov, Junyou Li, and Yuzhong Qu. 2020a. Entity summarization with user feedback. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*, pages 376–392. Springer.

Qingxia Liu, Gong Cheng, Kalpa Gunaratna, and Yuzhong Qu. 2020b. Esbm: an entity summarization benchmark. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*, pages 548–564. Springer.

Qingxia Liu, Gong Cheng, Kalpa Gunaratna, and Yuzhong Qu. 2021. Entity summarization: State of the art and future challenges. *Journal of Web Semantics*, 69:100647.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. Entsum: A data set for entity-centric summarization. *arXiv preprint arXiv:2204.02213*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia

Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao

Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan

Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Dongjun Wei, Yaxin Liu, Fuqing Zhu, Liangjun Zang, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Esa: entity summarization with attention. *arXiv preprint arXiv:1905.10625*.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: "what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Erhe Yang, Fei Hao, Yixuan Yang, Carmen De Maio, Aziz Nasridinov, Geyong Min, and Laurence T Yang. 2021. Incremental entity summarization with formal concept analysis. *IEEE Transactions on Services Computing*, 15(6):3289–3303.

A Appendix

A.1 Detailed Dataset Stats

We present our dataset statistics for the entity level and paragraph level in Table 3 and 4. Table 3 details the average number of attributes and values associated with individual entities. It also shows the average number of unique attributes and values observed across all entities, considering all paragraphs associated with each entity. Table 4 shows the average number of "same", "conflict", "new" attribute and value pairs, and an average number of sentences and distractor sentences in each paragraph.

A.2 Dataset Generation Prompts

Figure 5 and 6 show prompts for generating attributes and fake entity names, respectively. Figure 7 presents a prompt for generating values as a summary table format. Figure 8 and

Categories	# attr	# val	# diff attr	# diff val
Bath & Body Products	23.70	44.00	12.06	43.17
Bedding & Bed Linens	22.10	40.40	13.50	37.11
Books & Literature	22.30	43.80	16.78	43.00
Computer & Video Games	23.20	42.60	14.78	42.44
Computers & Electronics	22.50	43.80	15.72	42.22
Drugs & Medications	19.60	39.70	10.83	38.33
Education	23.20	45.20	14.56	44.50
Fashion & Style	23.50	45.60	15.17	45.11
Fruits & Vegetables	22.30	40.80	13.72	39.22
Hobbies & Leisure	22.80	44.50	16.06	44.22
Hotels & Accommodations	22.70	40.40	16.06	38.50
Household Supplies	21.60	40.90	13.50	38.39
Music Equipment & Technology	21.90	44.30	13.17	43.06
Oral & Dental Care	22.40	44.80	13.44	43.56
Pets & Animals	22.70	41.70	15.56	39.72
Restaurants & Bars	22.30	40.30	14.89	39.22
Skin & Nail Care	22.90	40.80	15.94	39.56
Sports	22.40	42.70	16.17	42.61
TV Shows & Movies	21.50	39.50	15.83	38.39
Vitamins & Supplements	20.80	39.70	13.44	38.89

Table 3: Entity level statistics. # attr: average number of attributes per entity, # val: average number of values per entity, # diff attr: average different number of attributes across entities, # diff val: average different number of values across entities.

9 are prompts for generating paragraphs and for aligning summary tables to paragraphs, respectively.

```
TASK: List the top 50 attributes when
people summarize entity of a given
category.  The attributes should be
common or rare according to the request.
Attributes should be separated by ’; ’.
```

Figure 5: Generate Attribute Instruction.

```
TASK: Generate 45 fake plausible entity
names in the given category.
Make sure that entity names are unique.
Entities should be separated by ’; ’.
```

Figure 6: Generate Entity Name Instruction.

```
TASK: Create a descriptive summary table
for a given entity focusing on the
following attributes and the given type.
For each attribute, generate at least
three descriptive values that are:

1.  Meaningful and informative.
2.  Diverse in length, ranging from one
word to a maximum of ten words.
3.  Varied in style, offering a mix
of user reviews, official product
descriptions, and editorial insights.
4.  Type:  "Fact" should not contain
any words that can be interpreted as
positive or negative properties of the
given entities (e.g.  restrooms are
well-maintained, family-friendly).

The summary table should have two columns:
attributes and values.  Ensure the
values are separated by ’; ’ to clearly
distinguish between them.
```

Figure 7: Generate Default Summary Instruction.

A.3 LLM-based Evaluation Prompts

Figure 18 shows a prompt used for LLM evidence finding (in Sec 4.2). In UPDATE method, we use prompts for GENERATE at 1st iteration, which are a combination of Figure 12 and 13. Afterwards, we use prompts for UPDATE in Figure 14 and 15 for the subsequent iterations. Similarly, in MERGE method, we use prompts for GENERATE at 1st iteration, which are a combined version of Figure 12 and 13. For the subsequent iterations, we employ two prompts for GENERATE (Figure 12 and 13) MERGE (Figure 16 and 17).

A.4 Human Verification of Dataset Alignment

We present details of the human verification for the alignment between generated tables and paragraphs. Figure 11 and Table 5 describes the instructions and examples that we share

Categories	# same attr-val	# conflict attr-val	# new attr-val	# sent	# dist
Bath & Body Products	3.45	3.58	2.43	12.28	4.00
Bedding & Bed Linens	3.50	3.87	2.23	12.33	4.00
Books & Literature	3.68	3.55	2.30	11.93	4.00
Computer & Video Games	3.57	3.35	2.48	12.13	4.00
Computers & Electronics	3.65	3.62	2.33	12.55	4.00
Drugs & Medications	4.17	3.22	1.95	12.12	4.00
Education	3.40	3.62	2.43	12.02	4.00
Fashion & Style	3.67	3.52	2.55	12.70	4.00
Fruits & Vegetables	3.93	3.02	2.38	11.88	4.00
Hobbies & Leisure	3.57	3.52	2.38	12.72	4.00
Hotels & Accommodations	3.45	3.67	2.38	12.33	4.00
Household Supplies	3.43	3.78	2.20	11.85	4.00
Music Equipment & Technology	3.58	3.72	2.28	12.25	4.00
Oral & Dental Care	3.40	3.88	2.20	11.80	4.00
Pets & Animals	4.00	3.30	2.32	12.18	4.00
Restaurants & Bars	4.02	3.45	2.18	12.23	4.00
Skin & Nail Care	3.77	3.55	2.42	12.42	4.00
Sports	3.87	3.27	2.32	12.35	4.00
TV Shows & Movies	3.55	3.42	2.22	11.97	4.00
Vitamins & Supplements	3.48	3.63	2.07	12.08	4.00

Table 4: Paragraph level statistics. # same attr-val: average number of same attribute-value pairs between paragraphs, # conflict attr-val: average number of conflicting attribute-value pairs between paragraphs, # new attr-val: average number of new attribute-value pairs between paragraphs, # sent: average number of sentences per paragraph, # dist: average number of distracting sentences per paragraph.

with the annotators for the verification task. In total, the dataset includes 11,551 sentences for verification. Each sentence is verified by three annotators. The averaged annotation time on each sentence is 17.59 seconds. The annotation task costs $1, 100. Out of 11,551 questions, 98.19% showed high consensus with a 3/3 agreement rate, while only 1.81% had a 2/3 agreement rate.

Sense checks: On a live basis, sense checks are conducted by Leads (Experts) to validate the responses given by raters. After validation, the expert inputs the correct answer in the ”Expert Answer” column and provides feedback to the raters, highlighting any errors made. This feedback mechanism assists in enhancing the overall quality of the responses.

A.5 Performance across paragraph tones and categories

Figure 19 shows F1 scores across additional 10 categories. Figure 20 presents F1 scores across paragraph tones.

A.6 Example data points of SUMIE dataset

Figure 21, 22, 23, 24 and 25 show examples of (attribute, value, sentence) triples and distractor sentences exist in our dataset in 5 different categories.

Entity	Category	Attribute	Value	Sentence	Annotation
ENTITY0	Computer & Video Games	Cultural impact	Spawned a Netflix anime series	But hey, at least it inspired a Netflix anime series! Now you can watch the game instead of playing it.	1
ENTITY1	Hotel	Social Spaces; Celebrity sightings	Crowded and noisy; Rare celebrity sightings	While the social spaces might get a bit crowded, who knows, you might just spot a celebrity or two!	1
ENTITY2	Restaurants & Bars	Service	Prompt and efficient service	While the craft beer selection may be limited, the happy hour specials are a steal , offering great value for families on a budget.	0

Table 5: Human Verification Examples: 1 represents that the attribute value is covered by the sentence, while 0 is the opposite.

```
TASK: Create a paragraph for a given
entity focusing on the following
attributes and values.
For each attribute and value, generate at
least one sentence that is:

1.  Meaningful and informative, including
both subjective opinions and objective
facts.
2.  Writing style should follow the given
paragraph writing style.
3.  Make sure to cite index number
in summary table when generating the
sentence.
4.  Make sure to include diverse
sentiments and attribute and values in
the summary table.
5.  Make sure not to change the core
meaning of attribute and value pair due to
writing style and sentiment.

The paragraph should include all index
numbers, attributes, and values in the
summary table.  Split sentences with a new
line.
```

Figure 8: Generate Paragraph Instruction.

```
TASK: Verify whether the given attributes
and values are described in the sentences
and whether corresponding index number is
cited correctly.

The inputs contain multiple lines, each
of which starts with multiple (index,
attribute, value) pairs, and a sentence
can be followed or not.
Please output True/False for each line.

These are two conditions of being False:
1.  Given attribute and value pairs do not
followed by a sentence.
2.  The context around citation number
does not match with the index number in
the attribute and value pairs.
3.  Sentiment of the given attribute and
value pair is incorrectly reflected in the
sentence.

If False is outputted, please provide
an explanation and revise the original
sentence or generate a new sentence to
describe the attribute and value for an
entity and its category.
Revised sentence should not include
any new information other than provided
attribute and value.
Ensure that all attribute and value pairs
are completely mentioned.
Make sure to include the index number of
the attribute and value pair using square
braces (e.g.  [index]).
Do not make up any citation numbers that
are not provided in (index, attribute,
value) pairs.
The format should be as follows:
"[(index1, attribute1, value1), (index2,
attribute2, value2), ...];;;True;;;"
or "[(index1, attribute1, value1),
(index2, attribute2, value2),
...];;;False;;;Explanation;;;Revised/New
sentence".
```

Figure 9: Critique for Summary-Paragraph Alignment Prompt.

```
TASK: Generate 10-15 complicated sentences
that describe the given entity and
category using the given attributes.
Generated sentences should:

1. Be meaningful and informative,
including both subjective opinions and
objective facts.
2. Be varied in style, offering a
mix of user reviews, official product
descriptions, and editorial insights.
3. Make sure to include entity name in
the sentence.

Split sentences with a new line.
```

Figure 10: Generate Irrelevant Sentence Instruction.

```
Objective: Determine if a given
attribute-value pair for an entity is
explicitly or implicitly covered by a
sentence.

Definitions:
Entity: The subject or topic (e.g.,
hotel, restaurant) to which the
attribute-value pair pertains.
Attribute-Value Pair: A specific
characteristic (attribute) and its
description (value) related to the entity.
Sentence: The text in which the
attribute-value pair coverage is to be
determined.
Coverage: An attribute-value pair is
considered "covered" if the sentence
directly or indirectly references the
attribute and matches or closely relates
to the given value.

Annotation Labels:
1 (YES): The sentence covers the
attribute-value pair.
0 (NO): The sentence does not cover the
attribute-value pair.

Annotation Guidelines:
Direct Match: 1 (YES) if the sentence
directly mentions the attribute value.
Indirect or Implied Match: 1 (YES) if
the sentence indirectly references the
attribute or describes the value in a
related way.
No Coverage: 0 (NO) if the sentence does
not mention or imply the attribute or the
value.
```

Figure 11: Human Verification Instruction.

```
Task Overview:
Your task involves synthesizing
information from detailed descriptive
paragraphs about a specific entity into a
summary table.
This table will highlight key attributes
of the entity along with their detailed
descriptions derived from the given texts.

Instructions:
* Extract Descriptive Values: Focus on
extracting specific, detailed information
rather than general or vague adjectives
like "good" or "bad." Ensure that
descriptions are precise and informative.
* Present a Balanced View: The table
should reflect a balanced perspective,
including positive, negative, and neutral
attributes. For attributes with mixed
reviews, indicate the sources supporting
each viewpoint.
* Attribute Selection:
- Commonly Interested Attributes: Include
attributes that are generally of interest
for the type of entity being described.
- Unique Attributes: Also identify and
include unique attributes that are
specifically mentioned in the provided
descriptions.
* Citations and Evidence: Each attribute
listed in the table should be supported by
citations from the source paragraphs.
Keep evidence concise but ensure it
substantiates the listed values.

Structure of the Summary Table:
* The table should be organized into two
columns: Attribute and Value.
* List attributes with their corresponding
values, including citations indicating the
source paragraph and relevant excerpts for
substantiation.
* Citation and evidence should be paired
in a [] and separated by ';'. If an
attribute has multiple values, then each
value should be separated by '&&&'.
```

Figure 12: Instruction prompt for GENERATE.

```
Example:
Entity: San Jose Marriott Hotel

Paragraphs:
P1. Great room and service, but breakfast
was lacking. We loved the spacious room
and friendly staff, but the breakfast
options were limited. There are two
pools.
P2. Poor customer service overshadowed
the beautiful location. The beachfront
view was amazing, but dealing with
unhelpful staff was frustrating. Room
is comfortable.
P3. Exceptional dining and comfortable
beds, but noisy at night. The restaurant
was five-star, and the beds were very
cozy, but there was a lot of street noise.

Summary Table:
| Attribute | Value |
| --- | --- |
| Room Quality | Spacious and comfortable
rooms ([P1, "spacious room"]; [P2, "Room
is comfortable"]) |
| Amenities | Two pools ([P1, "There are
two pools"]) |
| Service | Friendly staff ([P1, "friendly
staff"]) &&& overshadowed by unhelpful
staff ([P2, "Poor customer service
overshadowed the beautiful location"])
|
| Location | Beautiful beachfront view
([P2, "The beachfront view was amazing"])
|
| Food & Beverage | Exceptional dining
experience ([P3, "Exceptional dining"])
&&& limited breakfast options ([P1, "but
breakfast was lacking"]) |
| Noise Level | Notable street noise at
night ([P3, "but there was a lot of street
noise"]) |

Your Task:
Generate a similar table based on the
following descriptions of the specified
entity.
Entity: < entity name >

Paragraphs:
< paragraph >

Proceed to generate the summary table.
Output summary table format should follow
the above example of Summary Table.
```

Figure 13: Prompt that describes GENERATE task with one example.

```
Task Overview:
You are tasked with refining and expanding
an existing summary table based on new
descriptive paragraphs about an entity.
This involves updating the table to
include new information, modify existing
details without removing any, and ensuring
all entries are supported by evidence from
the text.

Instructions:
* Update Descriptive Values: Carefully
read the new paragraph(s) and identify
any information that should be added to
the current table entries or modify them.
Focus on specific, descriptive details,
avoiding vague adjectives.
**Do not remove any existing attributes or
values**, but rather add to or revise them
as necessary.
* Maintain a Balanced View: Ensure the
updated table continues to present a
balanced perspective, incorporating
positive, negative, and neutral values.
For any attribute with mixed evidence,
update the count of sources supporting
each view.
* Maintain a Balanced View: Ensure the
updated table continues to present a
balanced perspective, incorporating
positive, negative, and neutral values.
For any attribute with mixed evidence,
update the count of sources supporting
each view. All original attributes and
values must be preserved in the table,
with modifications only to reflect new
insights or corrections based on the
latest information.
* Attribute Revision and Addition:
- Commonly Interested Attributes: Update
or add attributes that are of general
interest for the type of entity being
described, based on the new information.
- Unique Attributes: Identify and
incorporate any unique attributes
mentioned in the new paragraphs that
were not previously included in the table.
* Evidence and Citations: For each
updated or new attribute entry, provide
citations from the new paragraphs. Strive
for concise evidence that directly
supports the attribute values.

Structure of the Updated Summary Table:
* Retain the two-column format: Attribute
and Value.
* For each attribute, list the updated or
new values along with citations indicating
the source paragraph and relevant excerpts.
Original attributes and values should
remain listed, with additional information
appended as necessary.
* Citation and evidence should be paired
in a [] and separated by ';'. If an
attribute has multiple values, then each
value should be separated by '&&&'.
```

Figure 14: Instruction prompt for UPDATE.

```
Example
Entity:  San Jose Marriott Hotel
New Paragraph:
"P4.  The hotel has recently renovated its
lobby, which now features a modern design.
Guests have also noted improvements in
breakfast variety and quality."

Given Existing Summary Table:
| Attribute | Value |
| --- | --- |
| Room Quality | Spacious and comfortable
rooms ([P1, "spacious room"]; [P2, "Room
is comfortable"]) |
| Amenities | Two pools ([P1, "There are
two pools"]) |
| Service | Friendly staff ([P1, "friendly
staff"]) &&& overshadowed by unhelpful
staff ([P2, "Poor customer service
overshadowed the beautiful location"])
|

Updated Summary Table:
| Attribute | Value |
| --- | --- |
| Room Quality | Spacious and comfortable
rooms ([P1, "spacious room"]; [P2, "Room
is comfortable"]) |
| Amenities | Two pools ([P1, "There are
two pools"]) |
| Food & Beverage | Exceptional dining
experience ([P3, "Exceptional dining"])
&&& limited breakfast options ([P1, "but
breakfast was lacking"]) &&& improved
breakfast variety and quality ([P4,
"improvements in breakfast variety and
quality"])|
| Lobby Design | Modern design ([P4,
"recently renovated its lobby, which
now features a modern design"])|

Your Task:
Update the summary table with the given
new descriptions of the specified entity.
Entity:  < entity name >
New Paragraph:
< paragraph >

Given Existing Summary Table:
< existing summary table >

Proceed to update the summary table.
Output summary table format should follow
the above example of Summary Table.
```

Figure 15: Prompt that describes UPDATE task with one example.

```
Task Overview:
You are tasked with combining two
summary tables based on existing and
new descriptive paragraphs about an entity
and generating an updated summary table
that contains information from both tables
(existing summary table and new summary
table).
This involves updating the table to
include new information, modify existing
details without removing any, and ensuring
all entries are supported by evidence from
the text.

Instructions:
* Update Descriptive Values:  Carefully
read the new paragraph(s) and identify
any information that should be added to
the current table entries or modify them.
Focus on specific, descriptive details,
avoiding vague adjectives.
**Do not remove any existing attributes or
values**, but rather add to or revise them
as necessary.
* Maintain a Balanced View:  Ensure the
updated table continues to present a
balanced perspective, incorporating
positive, negative, and neutral values.
For any attribute with mixed evidence,
update the count of sources supporting
each view.
* Maintain a Balanced View:  Ensure the
updated table continues to present a
balanced perspective, incorporating
positive, negative, and neutral values.
For any attribute with mixed evidence,
update the count of sources supporting
each view.  All original attributes and
values must be preserved in the table,
with modifications only to reflect new
insights or corrections based on the
latest information.
* Attribute Revision and Addition:
- Commonly Interested Attributes:  Update
or add attributes that are of general
interest for the type of entity being
described, based on the new information.
- Unique Attributes:  Identify and
incorporate any unique attributes
mentioned in the new paragraphs that
were not previously included in the table.
* Evidence and Citations:  For each
updated or new attribute entry, provide
citations from the new paragraphs.  Strive
for concise evidence that directly
supports the attribute values.

Structure of the Updated Summary Table:
* Retain the two-column format:  Attribute
and Value.
* For each attribute, list the updated or
new values along with citations indicating
the source paragraph and relevant excerpts.
Original attributes and values should
remain listed, with additional information
appended as necessary.
* Citation and evidence should be paired
in a [] and separated by ';'.  If an
attribute has multiple values, then each
value should be separated by '&&&'.
```

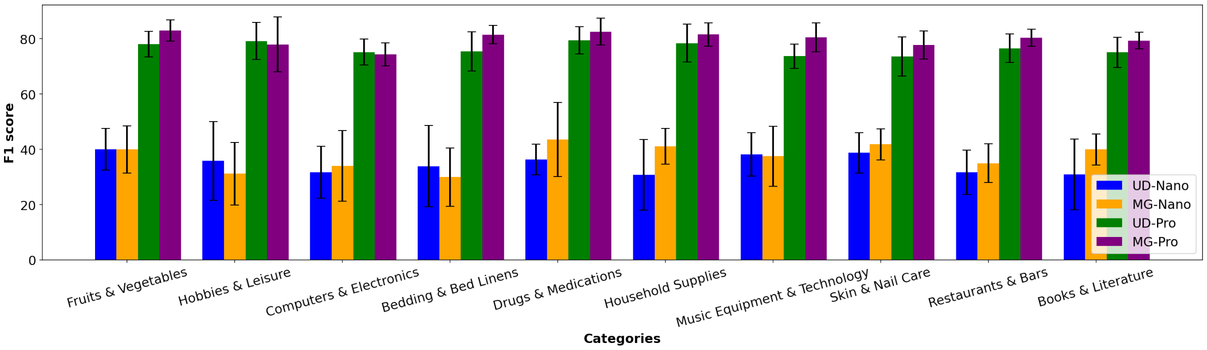
Figure 16: Instruction prompt for MERGE.

```
Example
Entity: San Jose Marriott Hotel

Given Existing Summary Table:
| Attribute | Value |
| --- | --- |
| Room Quality | Spacious and comfortable
rooms ([P1, "spacious room"]; [P2, "Room
is comfortable"]) |
| Amenities | Two pools ([P1, "There are
two pools"]) |
| Service | Friendly staff ([P1, "friendly
staff"]) &&& overshadowed by unhelpful
staff ([P2, "Poor customer service
overshadowed the beautiful location"])
|
| Food & Beverage | Exceptional dining
experience ([P3, "Exceptional dining"])
&&& limited breakfast options ([P1, "but
breakfast was lacking"])

New Summary Table:
| Attribute | Value |
| --- | --- |
| Food & Beverage | improved breakfast
variety and quality ([P4, "improvements in
breakfast variety and quality"])|
| Lobby Design | Modern design ([P4,
"recently renovated its lobby, which
now features a modern design"])|

Combined Summary Table:
| Attribute | Value |
| --- | --- |
| Room Quality | Spacious and comfortable
rooms ([P1, "spacious room"]; [P2, "Room
is comfortable"]) |
| Amenities | Two pools ([P1, "There are
two pools"]) |
| Food & Beverage | Exceptional dining
experience ([P3, "Exceptional dining"])
&&& limited breakfast options ([P1, "but
breakfast was lacking"]) &&& improved
breakfast variety and quality ([P4,
"improvements in breakfast variety and
quality"])|
| Lobby Design | Modern design ([P4,
"recently renovated its lobby, which
now features a modern design"])|

Your Task:
Combine existing and new summary tables
of the specified entity and generate a new
output summary table.
Entity: < entity name >

Given Existing Summary Table:
< existing summary table >

New Summary Table:
< new summary table >

Proceed to combine the two summary tables
and generate a new output summary table.
Output summary table format should follow
the above example of Summary Table.
```

Figure 17: Prompt that describes MERGE task along with one example.

```
You will be given two summaries: a
reference summary table (gold standard)
and a generated summary table. Your task
is to check if the gold standard contains
the information in the generated summary.
Please output with Yes/No.

Requirements for Yes:
- Meaningful Correspondence: Each
attribute-value pair in the generated
table should capture the core meaning of
its corresponding pair in the reference
table, even if worded differently.
- Partially relevant evidence is okay:
While the evidence in the generated table
does not have to be exactly match with
its corresponding attribute-value pair's
evidence, it should not be completely off
base.
```

Figure 18: LLM-based redundancy checking prompt.



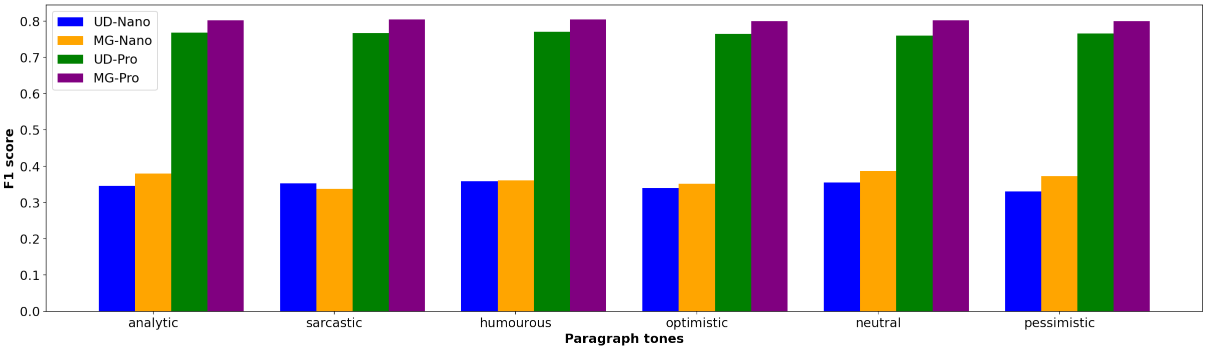


Figure 20: F1 scores across paragraph tones.

```
Category: Computer & Video Games

Examples of (Attribute, Value, Sentence):
Attribute: Memorability of characters
Value: Limited
Sentence: GAME1 offers limited memorable
characters , making it a forgettable
gaming experience.

Attribute: Story
Value: Lackluster and predictable
storyline
Sentence: And don't even get me started
on the story - it's so predictable, I
could write it in my sleep!

Attribute: Microtransactions and in-game
purchases
Value: Optional microtransactions
Sentence: The game features optional
microtransactions , so you can choose not
to spend any additional money.

Examples of distractor sentences:
- HUMAN's empathy is a healing potion,
allowing them to connect with others and
understand their virtual and real-life
struggles.
- GAME10's characters are complex and
relatable, drawing players into the game's
world and making them care about the fate
of Aloy and her companions.
```

Figure 21: Examples of (attribute, value, sentence) triples and distractor sentences in Computer & Video Games category.

```
Category: Vitamins & Supplements

Examples of (Attribute, Value, Sentence):
Attribute: Brand
Value: Longstanding history in the
industry
Sentence: Vitamin Company1 boasts a
long-standing history in the industry ,
ensuring credibility and trust for their
products.

Attribute: Price
Value: Not covered by insurance
Sentence: But hey, at least it's not
covered by insurance .

Attribute: Side Effects
Value: May cause mild gas or bloating
Sentence: But be warned, this fiber party
comes with a side of gas and bloating.

Examples of distractor sentences:
- HUMAN's optimism is a probiotic,
maintaining a healthy balance in their
outlook and promoting a positive gut
feeling about the future.
- The technology-enabled tracking
feature of Vitamin Company10 allows
users to monitor their caffeine intake
conveniently.
```

Figure 22: Examples of (attribute, value, sentence) triples and distractor sentences in Vitamins & Supplements category.

```
Category:  Restaurants & Bars

Examples of (Attribute, Value, Sentence):
Attribute:  WiFi Access
Value:  Convenient for business meetings
or working lunches
Sentence:  This spot offers convenient
WiFi access, making it perfect for
business meetings or working lunches.

Attribute:  Catering Services
Value:  Delicious and customizable menus
Sentence:  And if you're feeling fancy,
hit up their catering service.

Attribute:  Noise Level
Value:  Excessively loud and distracting
Sentence:  Just be warned, it can get loud
AF , so if you're trying to have a deep
convo, forget about it.

Examples of distractor sentences:
- HUMAN's determination is a bustling
coffee shop, where the aroma of ambition
permeates the air.
- RESTAURANT10's edible garden on-site
provides fresh, seasonal ingredients that
add a touch of vibrancy to their dishes.
```

Figure 23: Examples of (attribute, value, sentence) triples and distractor sentences in Restaurants & Bars category.

```
Category:  Books & Literature

Examples of (Attribute, Value, Sentence):
Attribute:  Overall Quality
Value:  A masterpiece of literature
Sentence:  Step into BOOK1, a literary
masterpiece that will transport you to
another realm.

Attribute:  Language
Value:  Written in prose
Sentence:  AndImmerse yourself in the
author's exquisite prose , which paints
vivid imagery on the canvas of your mind.

Attribute:  Binding
Value:  Unattractive and unappealing
Sentence:  While the binding may not
be its strong suit , the power of the
story within far outweighs its aesthetic
shortcomings.

Examples of distractor sentences:
- HUMAN's life is a masterpiece, a unique
and captivating story that is still being
written with every passing day.
- BOOK10's books are not only visually
stunning but also intellectually
stimulating, inviting readers to engage
with complex themes and ideas.
```

Figure 24: Examples of (attribute, value, sentence) triples and distractor sentences in Books & Literature category.

Category: Education

Examples of (Attribute, Value, Sentence):
Attribute: School Culture
Value: Lack of support and camaraderie among students
Sentence: Welcome to The Evergreen School, where the competition is fierce and the support is scarce .

Attribute: Learning Environment
Value: Innovative teaching methods
Sentence: But hey, at least they'll be exposed to innovative teaching methods (if they can keep up with the breakneck pace).

Attribute: Study Abroad Opportunities
Value: Immersive experiences in diverse cultures
Sentence: If you're looking for immersive experiences in diverse cultures, this school offers study abroad programs that will expand your horizons.

Examples of distractor sentences:
- HUMAN's mind is a fertile ground where ideas bloom and take root, transforming into a thriving garden of understanding.
- EDUCATION10's strong industry partnerships provide students with valuable internships and networking opportunities, preparing them for successful careers.

Figure 25: Examples of (attribute, value, sentence) triples and distractor sentences in Education category.