HotelCheckSpan: A Benchmark Dataset for LLM Faithfulness

Patrícia Schmidtová[1], **Ondřej Dušek**[1], **Saad Mahamood**[2]
[1]Charles University, Faculty of Mathematics and Physics [2]Shopware
{schmidtova, odusek}@ufal.mff.cuni.cz saad@saad.me.uk

Abstract

Hallucinations are among the most persistent and challenging issues in large language model (LLM) outputs. This particularly holds in domains that combine both objective and subjective content, such as hotel descriptions, that are intended to be enticing advertisements for the hotel. Distinguishing between factual errors and interpretative exaggeration is often subtle, complicating both human and automated evaluation. To address this, we present HotelCheckSpan, the first span-level faithfulness dataset for the hotel domain. Each example aggregates one or more hotel descriptions, and human-annotated summaries are labeled with three error types: *Incorrect*, *Misleading*, and *Not Checkable*. By marking the precise spans where errors occur, the dataset captures fine-grained information about the nature of hallucinations and factual inconsistencies. In addition to human annotations, we collect span-level judgments from multiple LLMs, enabling direct human–model comparisons. Our analysis shows that inter-annotator agreement varies substantially across aggregation levels: example-level agreement can mask subtle span-level disagreements, while soft and hard F1 variants highlight discrepancies in both span placement and error categorization. HotelCheckSpan provides a benchmark for studying ambiguity and disagreement, validating automatic faithfulness metrics, and evaluating LLMs as judges, offering a rich resource for research on faithfulness, subjectivity, and annotation practices in mixed-content domains.

Keywords: dataset, summarization, span labeling, subjectivity, metaevaluation

1. Introduction

Large language models (LLMs) have become increasingly proficient at generating fluent text across a variety of domains. However, one of their most persistent and concerning shortcomings is hallucination—the production of content that is unfaithful to the source or factually incorrect (Ji et al., 2023; Maynez et al., 2020). Hallucinations are particularly problematic in domains that combine both objective information with subjective assessments, such as hotel descriptions, where subtle nuances and interpretative statements occur frequently. In these contexts, the distinction between a true factual error and a minor exaggeration often complicates both human and automated evaluation.

To capture this subtle information, we adopt a span-level annotation approach (Thomson and Reiter, 2020a), which identifies the exact spans in generated summaries that are erroneous and labels the type of error. This approach provides rich information about what exactly the errors are, rather than simply whether a summary is overall faithful or not. By analyzing the span-level patterns of agreement and disagreement among annotators, we can illuminate both the nature of the errors and the inherent ambiguity in subjective domains.

We introduce HotelCheckSpan, the first span-level faithfulness dataset in the hotel domain. The dataset aggregates multiple accommodation descriptions per example and contains human-annotated summaries labeled for three types of errors: *Incorrect* (factually wrong information), *Misleading* (technically true, but open to misinterpretation), and *Not Checkable* (cannot be supported nor refuted by the input description). Table 1 illustrates an example description–summary pair with this span-level annotation approach. In addition to human annotations, we collect comparable span-level judgments from multiple LLMs to enable direct human–model comparisons. We release the code[1] and the dataset[2] publicly.

The main contributions of this work are as follows:

1. **A new dataset for span-level faithfulness assessment**, HotelCheckSpan, covering a mix of subjective and objective information in the hotel domain. This dataset is, suitable for studying ambiguity, annotator disagreement, and for validating automated metrics, including LLM-based evaluators (Kasner et al., 2025; Pagnoni et al., 2021).
2. **Comprehensive analysis of agreement and disagreement**, showing how different aggregation levels (example-level vs. span-level) provide complementary perspectives on inter-annotator consistency and ambiguity.
3. **Systematic human–LLM comparison**, which

Work done on PS' internship with trivago, supervised by SM while at trivago.

[1] https://github.com/patuchen/hotelcheckspan

[2] https://huggingface.co/datasets/patuchen/hotelcheckspan

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9973–9987
11–16 May 2026.

Description: *Just a 5-minute walk from Mall of the Emirates, DoubleTree by Hilton Hotel and Residences Dubai offers modern accommodations. [...] The hotel is 7.0 km from Dubai Marina and 12.1 km from Dubai Mall. Dubai International Airport is 30 minutes away by car.*
S1: *Shop in the Mall of the Emirates thanks to the hotel's convenient location.*
S2: *Enjoy wonderful views across the Hudson River to New Jersey and Liberty Island from select suites.*

Table 1: An example of a hotel description with span-level annotation. Note that in the actual dataset, each description is paired with exactly one summary. Here, S1 demonstrates a faithful summary, while S2 illustrates an *Incorrect* error (highlighted in red) used specifically as an attention check during annotator qualification.

demonstrates varying levels of model reliability and highlights the challenges of using LLMs as faithfulness judges.

4. **Recommendations for IAA reporting:** We demonstrate that relying on a single inter-annotator agreement (IAA) metric can mask underlying disagreements. To prevent this distortion, we recommend evaluating agreement across multiple granularities: dataset-level, example-level, and span-level.

2. Related Work

The news domain dominates the existing summarization datasets with dialog being the second largest domain (Dahan and Stanovsky, 2025). This lack of domain diversity means that as researchers, we might be inaccurately assessing the utility of summarization methods due to lack of insight. We therefore structure our coverage of related work into two subsections: datasets for summarization evaluation and datasets in the hotel domain.

2.1. Faithfulness Evaluation

Research on evaluating faithfulness and hallucinations in abstractive summarization has led to the creation of several benchmark datasets. SummEval (Fabbri et al., 2021) remains the most widely used reference corpus, containing human evaluations of summaries from multiple models on the CNN/DailyMail dataset with judgments of coherence, consistency, fluency, and relevance. FactCC (Kryscinski et al., 2020) introduced a factual consistency benchmark derived from news summaries, accompanied by a BERT-based verification model trained to detect contradictions. More recent resources such as AggreFact (Tang et al., 2023) aggregate annotations from multiple factuality datasets, enabling cross-domain evaluation of consistency metrics. FaithBench (Bao et al., 2025) extends faithfulness evaluation to large language models, providing annotations of hallucinations of existing datasets at both sentence and span level without a domain focus.

2.2. Datasets in the Hotel Domain

Several datasets exist for textual modeling in the hotel and travel domain. HotelRec (Antognini and Faltings, 2020) is a large-scale TripAdvisor corpus containing 50 million reviews with metadata for recommendation research. Smaller collections are also publicly available on Kaggle,[3] Hugging Face (Alam et al., 2016), and other sites, including a 2012 dataset of 887K of TripAdvisor hotel reviews from 4,333 hotels[4], which provide tens to hundreds of thousands of user reviews paired with numerical ratings across multiple aspects. These datasets are typically used for sentiment classification or opinion mining rather than summarization.

A few resources directly address summary generation. The SPACE corpus (Amplayo et al., 2021) contains user reviews of hotels and other venues, with human-written summaries for a subset of entities. More recently, LFOSum (Nayeem and Rafiei, 2024) paired professional critic summaries with user reviews for over 500 hotels, supporting controllable long-form summarization. A smaller dataset by Kamath et al. (2024) introduced model-generated highlights of hotel descriptions categorically annotated on the document level for hallucinations and contradictions, providing the first example of hallucination evaluation in this domain. We use an alternative error span annotation that allows for more qualitative insights on the sources and the nature of the errors.

3. Dataset Overview

3.1. Task Definition

The objective of HotelCheckSpan is to describe the unique aspects of a given accommodation without the need for users to parse verbose accommodation descriptions and contradictory reviews. Each instance consists of an input text (a hotel

[3] https://www.kaggle.com/datasets/waseemalastal/hotel-reviews-dataset

[4] Four Cities Hotel-Reviews Dataset: https://www.cs.cmu.edu/~jiweil/html/hotel-review.html

description) and a corresponding short summary focused on a single aspect generated by a large language model. The dataset enables span-level analysis of factual consistency between summaries and their source texts, supporting both human and automatic evaluation.

3.2. Data Source and Selection

The source texts were sampled from publicly available accommodation descriptions written in English, primarily sourced from various online travel agencies (e.g., Booking.com, Expedia) and direct hotel providers. Depending on availability, each example aggregates one or more descriptions for a single property. The sampled accommodations span 44 countries across all inhabited continents, with the most prominently represented being the US (approx. 30%), the UK (8%), and Japan (7%). The properties are predominantly standard hotels (80%), alongside a mix of motels, bed and breakfasts, and apartments. Because our study strictly focuses on textual modeling and evaluation, geographical and property metadata were decoupled during processing and are not included in the released dataset.

To mitigate redundancy where properties provide nearly identical information across multiple platforms, we concatenated the available descriptions and applied an automated de-duplication step. Redundant segments were filtered out if they satisfied any of the following conditions against an existing segment: a BLEU score greater than 0.8, or a ROUGE-L F1 score greater than 0.8. Because the source texts consist of promotional material, they are typically fluent and internally consistent, though they may contain seasonal, outdated, or exaggerated marketing claims. For the purposes of this dataset, the aggregated description serves as the absolute ground truth. The annotation task strictly evaluates the generated summary's alignment with this provided text, isolating the model's generation fidelity from the real-world accuracy of the accommodation's claims.

We then generated short summaries using Gemini 1.5 Flash (Gemini Team, 2024), emphasizing a specific topic mentioned in the description (e.g., location, amenities, or dining options). The model was selected as a direct architectural upgrade to the PaLM 2 (Team, 2023) models utilized in prior work (Kamath et al., 2024). The generation of the summaries closely follows the methodology described by Kamath et al. (2024). These generated summaries were intentionally left unedited to preserve the natural distribution of model hallucinations for the subsequent annotation phase. Each summary is paired with its corresponding source text, forming a single evaluation instance.

3.3. Data Composition

The final version of HotelCheckSpan comprises a total of 496 description–summary pairs, each consisting of a single summary generated for an aggregated set of one hotel's descriptions. To allow for analysis with more controlled variables, the dataset has two splits of equal size: shorter and longer. On average, the input descriptions contain 716.5 words (68.4 sentences), ranging from 20 to 1,517 words. The summaries are highly concise, averaging 14.7 words (1 sentence) and ranging from 7 to 24 words. We note that the dataset contains short outliers. Refer to Table 1 for an illustrative example showing span-level annotation of unsupported content.

3.4. Availability

The base HotelCheckSpan dataset (the unannotated description–summary pairs), the annotation guidelines, and the processing code are publicly available.

However, given the extensive issues with test set contamination in LLM evaluation (Balloccu et al., 2024), the collected human and model span annotations are purposefully withheld from the public repository to preserve the integrity of the benchmark. These annotations will be made readily available to researchers upon request. Both the public dataset and the gated annotations are distributed under the **Creative Commons BY–NC 4.0** license, permitting non-commercial use with attribution.

4. Annotation Process

4.1. Annotation Schema

Each summary in HotelCheckSpan was annotated for factual errors at the span level. Annotators were instructed to select the smallest text fragment in the summary that constitutes a factual error with respect to the input description. When a single summary contained multiple distinct factual issues (e.g., an incorrect facility description alongside an unverifiable location claim), annotators were instructed to highlight each error as a separate, independent span. Overlapping spans were not allowed.

Each span was assigned exactly one of three error types: *Not Checkable*, *Misleading*, or *Incorrect*. We derived these three categories from Kasner et al. (2025), retaining only those relevant to the hotel domain based on our initial data analysis. We deliberately opted for this concise schema rather than a more granular one (e.g., separating omissions from exaggerations) to keep the cognitive load manageable for crowdworkers. As noted by Thomson and Reiter (2020b), highly complex guidelines often degrade inter-annotator agreement without yielding

higher-quality insights. The categories are defined as follows:

- *Not Checkable* errors denote statements that cannot be verified based on the source text (e.g., mentioning an amenity that is never described).
- *Misleading* errors occur when the summary distorts the meaning of the source, often by exaggeration or ambiguous phrasing (e.g., referring to a hotel as "beachfront" when the beach is several kilometers away).
- *Incorrect* errors capture clear factual contradictions, such as wrong entities or numerical mismatches.

We acknowledge that the boundary between *Misleading* and *Not Checkable* can sometimes be subjective, depending on whether a statement is perceived as merely a distortion of existing text or an entirely ungrounded claim. To anchor these judgments, annotators were provided with clear definitions and one representative example for each error class, which are fully disclosed in Appendix B. Furthermore, annotators were explicitly asked to ignore stylistic differences, omissions, or subjective intensifiers (e.g., *"wonderful location"*) that do not affect factual accuracy.

In addition to the error spans, the annotators were asked for their overall impression of the summary on a scale from 1 to 7. The 7-point Likert scale has been deliberately chosen as van der Lee et al. (2019) find it optimal based on the analysis of prior studies.

4.2. Annotation Interface and Guidelines

Span annotations were collected using the *Factgenie* span annotation interface (Kasner et al., 2024) without further modifications. Prior to the task, annotators received detailed written guidelines explaining the annotation procedure. They were instructed to highlight the smallest text fragment in the summary that constituted a factual error, assigning one label per span, and to mark no spans when a summary contained no errors by checking the corresponding box. The instructions emphasized selecting minimal spans that, if removed or replaced, would correct the error while preserving grammaticality. Examples illustrating correct and incorrect annotations were provided, together with clarifications that stylistic differences and subjective expressions should not be annotated.

Annotation time was automatically tracked, and submissions flagged as implausibly quick post-submission were manually reviewed. Low-quality annotations were rejected and subsequently replaced with other annotators. A small number of workers were found to delay their Prolific submissions to bypass time checks; these cases were identified manually and excluded from the final dataset.

4.3. Annotators

The annotation campaign was conducted in two stages, preceded by three pilot rounds to verify the clarity of the guidelines and estimate annotation time.

Pilots Three pilots were conducted prior to the main annotation campaigns to validate the experimental design, gain feedback on the clarity of guidelines, and estimate the time complexity. The first pilot involved eight core team members familiar with the data and task, the second included two colleagues external to the project, and the third involved eleven crowd workers recruited via Prolific[5], representing the intended target population. The first two pilots were voluntary; the third was paid under the same conditions as the main annotation.

Stage 1. A total of 62 annotators were recruited on Prolific to annotate batches of ten examples, each containing two attention checks (an error-free example and one with an obvious *incorrect* error). Annotators were required to be native English speakers located in the United Kingdom or the United States and to have a prior Prolific task approval rate above 95%. Each example was double-annotated by workers randomly assigned to two groups (A and B) differing only in the phrasing of the "no errors" checkbox: group A saw phrasing “I did not find any errors in the summary” while group B saw a less personal “There were no errors in the summary”.

Stage 2. To improve annotation quality, a second stage was carried out on a subset of the data. For this, we used the half (248 examples) of the dataset whose descriptions were shorter and thus required less time to annotate. We further set aside 3 examples from this subset to serve as additional questions in the qualification task, resulting in 245 examples receiving two additional annotations. In addition to the filtering criteria in Stage 1, we added a filter for the minimal number of tasks completed on Prolific to 200. Potential annotators first completed a qualification task consisting of five examples (the two original attention checks and three new ones). Each participant was scored on a 0–5 scale based on their accuracy; each example had a reference error and was worth one point. Partial point assignment was possible for the consistent use of the “no

[5]Prolific - https://www.prolific.com

Error Type	Count	Percentage
Not Checkable	367	46%
Misleading	314	40%
Incorrect	113	14%

Table 2: Distribution of annotated error types in HotelCheckSpan.

errors present" checkbox. Those scoring at least three points were invited to continue. Out of forty candidates, only twenty-eight qualified and were subsequently considered trusted annotators. They could participate in multiple batches, each containing sixteen examples. All annotators were paid £9 per hour, with a performance-based bonus for passing attention checks that raised their effective pay to £12.60 per hour.

4.4. Dataset Statistics

A total of 496 description–summary pairs were annotated in the main campaign. Each example was labeled independently by at least two annotators, and 245 of these were annotated by two more annotators to enable subsequent reliability analysis. In total, the dataset contains 794 annotated error spans.

On average, summaries contain 0.54 annotated spans per record. When considering only examples with at least one annotation, this corresponds to 1.31 spans per summary. Annotated spans are typically short, with an average length of 4 words, a median of 3 words, with lengths ranging from 1 to 18 words. Based on this information, annotators generally marked minimal text segments representing discrete factual issues, as requested in the guidelines.

The distribution of annotated error types is shown in Table 2. The majority of the annotation spans are labeled as *Not Checkable* (46%), followed by *Misleading* (40%) and *Incorrect* (14%). This pattern suggests that unverifiable content is the most frequent source of faithfulness concerns in model-generated hotel summaries.

Inter-annotator agreement and qualitative annotation trends are discussed in Section 6.

5. LLM Annotation Collection

To complement the human annotations described in Section 4, we additionally collected span-level faithfulness annotations from large language models (LLMs). The setup followed the procedure of Kasner et al. (2025), using the same *Factgenie* interface to ensure comparability between human and model annotation campaigns. We used three different models: GPT-4o (OpenAI Team, 2024), o3-mini,[6] and Gemma-3 (Gemma Team, 2025). Each model was run on the full set of 496 description–summary pairs annotated by humans in Stage 1.

Prompt design The prompting protocol was adapted from Kasner et al. (2025), with modifications to include dataset-specific examples identical to those used in the human annotation guidelines. For `gpt-4o` and `o3-mini`, we ran two variants: a *default* prompt and a *less strict* variant (abbreviated as **LS**), which explicitly instructed the model to mark only errors with higher confidence or stronger factual contradictions.[7]

Results overview Table 3 summarizes the resulting annotation statistics across all of the LLM campaigns. The table reports total span counts, average span density, and the distribution across error categories.

Across models, annotation density and error-type balance varied considerably. The `gpt-4o` model and its less strict variant produced the highest number of spans (272 and 295, respectively), exhibiting a strong tendency to **over-annotate** relative to human annotators. Because the identical prompt structure proved effective in prior work (Kasner et al., 2025), we attribute this behavior to the model's inherent difficulty adapting to the subjective nuances of the hotel domain, rather than a purely prompt-driven artifact. Conversely, `gemma-3` exhibited a more conservative annotation style, generating fewer spans per record but a notably higher share of *Misleading* errors (65%). We omitted a less strict variant for `gemma-3` as preliminary observations indicated it already captured the general error distribution adequately, and prompt relaxations did not meaningfully improve its span boundary precision. Finally, the reasoning model `o3-mini` demonstrated severe **under-annotation**; it had high precision on the few spans it identified, but very low recall, finding at most 4 annotations across the entire dataset.

Qualitative Validation Beyond the raw span counts, a manual qualitative inspection of the LLM-generated annotations confirmed these quantitative limitations. While the few spans identified by the `o3-mini` variants were semantically reasonable and precise, annotations from `gpt-4o` and `gemma-3` frequently violated the core instruction to select minimal spans, often highlighting entire sentences. Furthermore, the rationales generated by these models frequently contradicted themselves

[6] https://platform.openai.com/docs/models/o3-mini

[7] Verbatim prompt templates will be provided in the appendix of the camera-ready version.

LLM	Total	Spans/Rec.	Spans/Err.Rec.	Avg Len.	Incorrect	Misleading	Not Check.
gemma-3	177	0.36	1.19	4.7	47 (27%)	116 (65%)	14 (8%)
gpt-4o	272	1.10	1.18	3.8	111 (41%)	70 (26%)	91 (33%)
gpt-4o (LS)	295	1.19	1.19	3.9	104 (35%)	89 (30%)	102 (35%)
o3-mini	2	0.00	1.00	3.0	1 (50%)	1 (50%)	0 (0%)
o3-mini (LS)	4	0.01	1.00	5.5	3 (75%)	1 (25%)	0 (0%)

Table 3: Statistics of span annotations marked by LLMs. "LS" = *Less Strict* prompt variant. All campaigns were run on the full set of 496 examples.

or penalized valid stylistic choices instead of factual errors. This confirms that the models struggle fundamentally with the semantic reasoning required for this specific task, validating the necessity of human-in-the-loop evaluation.

This parallel collection of human and LLM-generated span annotations enables a direct comparison of annotation behavior, agreement patterns, and error-type tendencies, which we examine in the following section (Section 6).

6. Inter-Annotator Agreement

All span annotation agreement methods have their pros and cons which will be described in the following subsections. We point out that our dataset contains a higher amount of subjectivity which lowers the agreement. The impact of subjectivity on the agreement will be discussed in Section 6.4 in light of Plank (2022) who points out that label variation in human annotations can be a signal rather than noise.

6.1. Dataset-Level Agreement

Before analyzing fine-grained inter-annotator consistency, we examine agreement trends at the dataset level by comparing aggregate error distributions across annotation campaigns. Table 4 summarizes the number and type of spans produced in the human and LLM annotation rounds.

To assess annotation robustness, we compared outcomes between the two annotation stages and across the two UI phrasing groups (A and B). Overall, we observed no systematic conceptual shifts between Stage 1 and Stage 2. The procedural updates and qualification filtering introduced in Stage 2 resulted in a marginal increase in inter-group agreement (+4.8 points) and a slight reduction in overall annotation volume (–8.5%). This confirms that the Stage 2 annotators became more selective and internally consistent in their label assignments without fundamentally altering their interpretation of the guidelines.

However, a moderate systematic difference emerged between the annotator groups themselves. Across both stages, Group B produced roughly 20% more error annotations than Group A. When Group B was in the minority during a disagreement, it was almost exclusively due to flagging additional errors (over 90% of cases) rather than missing them. We attribute this mild tendency toward over-annotation to the subtle variation in the "no errors" checkbox: Group A saw the personalized phrasing "I did not find any errors in the summary," which seemingly prompted a more conservative, cautious stance, whereas Group B saw the objective "There were no errors in the summary." These findings highlight how even minor UI variations can influence annotator confidence and error thresholding in subjective tasks.

Among the LLM-based campaigns, `gpt-4o` yields substantially more spans than either human group, suggesting a general tendency to **over-annotate** relative to human annotators. Its less strict (*LS*) variant produces an even higher span count, though with a similar error-type balance. In contrast, *o3-mini* produces only a handful of annotations, demonstrating severe **under-annotation** and limited detection capability. The `gemma-3` model sits between these extremes, generating a moderate number of spans with a strong bias toward the *Misleading* category.

The dataset-level agreement results serve as a sanity check and important context for interpreting the more granular findings. However, while the counts of error spans and distributions of errors are similar across campaigns, they do not inform us whether the annotators in fact found the errors in the same examples.

6.2. Example-Level Agreement

We next examine agreement patterns at the example level, focusing on how many annotators marked the presence of each error type for a given summary. Each example in Stage 2 was independently annotated by four qualified annotators, allowing us to observe the distribution of "error votes" (i.e., the number of annotators marking an error). Note that this includes any errors within the example and two annotators might mark two distinct errors within a single example.

Table 5 reports how often a given number of annotators (0–4) identified each error type—*Not*

Campaign	Total Spans	Incorrect	Misleading	Not Checkable
St. 1A	131	27 (21%)	50 (38%)	54 (41%)
St. 1B	148	14 (9%)	66 (45%)	68 (46%)
St. 2A	102	4 (4%)	50 (49%)	48 (47%)
St. 2B	125	12 (9%)	47 (38%)	66 (53%)
gem.3	92	27 (29%)	53 (58%)	12 (13%)
4o	269	109 (41%)	70 (26%)	90 (33%)
4o (LS)	292	102 (35%)	89 (30%)	101 (35%)
o3-m (LS)	4	3 (75%)	1 (25%)	0 (0%)
o3-m	1	0 (0%)	1 (100%)	0 (0%)

Table 4: Dataset-level comparison of total annotated spans and their distribution across error types for human (Stage 1/2) and LLM campaigns, calculated on the subset of 245 Stage 2 examples. Multiple error spans can be identified within a single example. "LS" = Less Strict prompt variant.

Checkable (NC), *Misleading* (M), and *Incorrect* (I)—as well as whether any error was marked (*Any*). The proportions are computed over all 245 examples annotated by four annotators.

E	NC	M	I	Any
0	110 (45%)	109 (45%)	207 (84%)	51 (21%)
1	86 (35%)	88 (36%)	34 (14%)	68 (28%)
2	32 (13%)	39 (16%)	4 (2%)	63 (26%)
3	14 (6%)	8 (3%)	0 (0%)	45 (18%)
4	3 (1%)	1 (0%)	0 (0%)	18 (7%)

Table 5: Example-level annotation agreement across four Stage 2 annotators. "Error Votes (**E**)" = the number of annotators voting for the presence of an error in a given example. The following columns show example count and prevalence: **NC** = *Not Checkable*, **M** = *Misleading*, **I** = *Incorrect*, **Any** = presence of any error. For example, there are 110 examples where none of the annotators found an NC error. Percentages are rounded to whole numbers.

Observations Table 5 highlights several clear patterns. The *Incorrect* category shows the strongest consensus, with 84% of examples containing no such error according to all four annotators, confirming that annotators are highly conservative when marking content as factually wrong. In contrast, both *Misleading* and *Not Checkable* errors display a more even distribution across 0–2 votes, suggesting a higher degree of subjectivity with more ties.

When considering any error type (*Any*), only 7% of examples were unanimously judged to contain an error (4 votes), while 21% received none (0 votes). Most examples fall into the middle bands (1–3 votes), underscoring the inherent variability of faithfulness judgments at the summary level.

LLM Performance When considering example-level cases with clear human consensus (3–1 or 4–0 agreement), we further examined how LLMs align with these majority decisions. The results reveal that although the highest-scoring models (`o3-mini` and `o3-mini (LS)`) achieve over 95% agreement with human judgments when evaluated per error type, this pattern does not generalize when the evaluation is aggregated across all error categories. When measured in terms of overall binary error detection (any error vs. none), their alignment drops to around 68–69%. A similar but weaker discrepancy appears in other models as well, with `gemma-3` decreasing from 86% to 63%, and `gpt-4o` models falling from around 65–67% to near 35%. This systematic gap arises because humans overwhelmingly agree on the absence of specific error types, while agreement on the presence of an error is rare. As a result, per-type majority alignment artificially inflates the performance of models that simply default to the majority class baseline (i.e., the absence of a specific error). The `o3-mini` variants achieve near-perfect per-type scores entirely by consistently predicting "no error," while `gpt-4o` models, despite lower overall agreement, are the only ones that actually identify some of the rare true positives. This conservative alignment paradox demonstrates that standard per-type agreement metrics can severely overstate LLM–human consistency, emphasizing the need to evaluate models on their ability to detect the minority class (actual errors) rather than just rewarding their alignment with human consensus on error-free text.

6.3. Span-Level Agreement

Evaluating agreement on span annotations is nontrivial, as spans can vary in length, boundaries, and degree of overlap. To ensure comparability with prior work, we adopt the conventions introduced by Kasner et al. (2025), following the definitions established in Da San Martino et al. (2019).

We compute pairwise agreement between annotators using precision, recall, and their harmonic

mean (F1). Taking one annotator as the reference and the other as the hypothesis, a span is counted as a true positive when a span in the hypothesis overlaps with a span in the reference. We use the same logic to measure recall and F1. This span-based formulation evaluates only overlaps between annotated spans—it does *not* assign credit for cases where both annotators leave a segment unmarked.

We report two variants for each metric: **Hard** (which considers both span boundaries and the assigned error type) and **Soft** (which considers only span boundaries, ignoring the error type). The difference between them (e.g., ΔF1 = $F1_{soft}$ – $F1_{hard}$) reflects disagreement arising from label assignment rather than span detection.

Analysis. Agreement between human annotators remains modest, with soft F1 scores around 0.25–0.30 and hard scores around 0.10–0.15. Stage 2 annotators show the highest internal consistency (**$F1_S$=0.28**), suggesting improved calibration after pilot rounds. Among LLMs, `gpt-4o` and `gpt-4o (LS)` reach human-level agreement in the soft setting, while `gemma-3` performs slightly worse and `o3-mini` exhibits near-zero alignment, consistent with its strong underannotation tendency. The persistent gap between soft and hard agreement (ΔF1 $\approx$ 0.15) indicates that disagreements arise primarily from label categorization rather than identifying which parts of the summaries are problematic. Higher recall but lower precision for LLMs reflects their tendency to mark more potential issues than human annotators, while human–human pairs tend to be more conservative yet consistent in labeling.

6.4. Error and Disagreement Analysis

To better understand the nature of the annotations and sources of disagreement, we conducted a semantic clustering analysis of all annotated error spans. This approach complements the span-level agreement metrics by examining what kinds of claims tend to be marked as errors—and where annotators diverge most.

Methodology. We embedded 464 annotated spans using the `all-MiniLM-L6-v2` sentence transformer (Wang et al., 2020) and performed K-means clustering ($k = 15$). The resulting clusters represent semantically coherent error types discovered directly from the data, rather than predefined categories. Each cluster was manually inspected by the first author and labeled according to its dominant linguistic or factual pattern. Full implementation details and hyperparameters for the clustering are provided in Appendix A.

Overall Taxonomy. The clustering yielded fifteen interpretable groups, summarized in Table 7. The most frequent types of erroneous or disputed claims concern location and proximity, amenities, and subjective or experiential qualities. Proximity-related expressions (e.g., “near the city center,” “short walk,” “close to attractions”) account for roughly one third of all annotations, reflecting the prevalence of vague or unverifiable location language in the dataset. Amenity mentions (e.g., “spa centre,” “fitness room,” “terrace”) constitute another substantial share, while more subjective descriptions such as “beautiful,” “relaxing,” or “memorable” cluster under quality and experience claims.

Relation to Disagreement. Low-consensus annotations—cases where annotators disagreed on whether a span was erroneous—were found to concentrate in proximity and quality-related clusters. These often involve *soft factuality* claims, such as walking distances, view quality, or convenience, which blur the line between factual and subjective statements. In contrast, clusters involving concrete facilities or services (e.g., “business centre,” “fitness room”) showed higher agreement, suggesting clearer factual grounding.

Error Type Breakdown. When grouped by error type, the taxonomy reveals systematic tendencies:

- **Non-checkable errors** frequently describe subjective or unverifiable qualities, such as *Walking Distance*, *Facility Description*, and *Experience* claims.
- **Misleading errors** tend to involve overstated amenities or proximity (e.g., “easy access to attractions,” “spa centre with a hot tub”). While many of these statements might be technically true, they leave a wide margin for interpretation which can lead to misunderstanding and disappointment.
- **Incorrect errors** mostly concern concrete facility or feature mentions (e.g., “business centre,” “in-room service”), indicating factual inaccuracies rather than interpretative uncertainty.

Interpretation. Overall, the embedding-based taxonomy suggests that most faithfulness disagreements arise not from purely factual contradictions, but from linguistic vagueness and soft factuality expressions. Terms implying distance, convenience, or experiential quality often depend on contextual or subjective interpretation, making consistent annotation particularly difficult. Conversely, concrete claims about facilities or services elicit higher agreement and more consistent error labeling.

Comparison Type	Prec (H)	Rec (H)	F1 (H)	Prec (S)	Rec (S)	F1 (S)
Group A – Group B (Stage 1)	0.10	0.10	0.10	0.29	0.28	0.28
Group A – Group B (Stage 2)	0.22	0.19	0.15	0.33	0.24	0.28
Stage 1 – Stage 2	0.15	0.11	0.12	0.29	0.21	0.24
Human–LLM (Gemma3)	0.10	0.07	0.08	0.24	0.18	0.21
Human–LLM (GPT-4o)	0.09	0.13	0.11	0.20	0.39	0.26
Human–LLM (O3-Mini)	0.20	0.01	0.02	0.62	0.03	0.05

Table 6: Average span-level agreement between human annotators and large language models on the Stage 2 subset. "H" = type-sensitive (hard), "S" = type-insensitive (soft). Precision (Prec), Recall (Rec), and F1 scores are averaged over all pairwise comparisons.

Cluster Name	Count	(%)
Central Location	60	14.1
Amenity Availability	50	11.8
Proximity (var.)	127	29.9
Facility Description	31	7.3
Access Convenience	29	6.8
Subjective Quality	29	6.8
Experience/Atmosphere	25	5.9
Business Facility	24	5.6
Dining/Food	22	5.2
Scenic Quality	16	3.8
Walking Distance	15	3.5
Amenity Availability (Fitness)	12	2.8

Table 7: Overview of embedding-based clusters representing typical error types in annotations.

6.5. Relation Between Errors and Impression Scores

To examine how factual and subjective errors influence the perceived quality of model outputs, we analyze the relationship between error presence and the overall impression scores assigned by annotators. Impression scores range from 1 (very poor) to 7 (excellent) and were available for all annotated summaries across the four campaigns.

The results show a strong and statistically significant relationship between the presence of errors and lower impression scores ($t = 11.9$, $p < 0.001$). Summaries without detected errors received an average score of 4.79, while those containing at least one error averaged 3.71. Moreover, impression scores decline steadily with the number of detected errors (correlation $r = -0.31$), dropping below 2 when more than four distinct errors were identified. This suggests that even a small number of factual issues substantially affects perceived quality.

Breaking down the analysis by error type reveals that *Incorrect* errors have the most severe impact (mean 2.58, a drop of 2.21 points from the no-error baseline), followed by *Not Checkable* (mean 3.48, drop 1.31) and *Misleading* errors (mean 3.94, drop 0.85). The trend indicates that annotators penalize factual inaccuracies more heavily than ambiguous or interpretive issues. This penalty hierarchy observed in the annotators' impression scores mirrors anticipated business impact. As demonstrated in our prior work (Schmidtová et al., 2025), end users reading these summaries for practical decision-making are less likely to be severely impacted by *Misleading* statements compared to outright *Incorrect* claims. Consequently, distinguishing between these error types is critical for automated evaluation, as they carry vastly different risks of reputational or legal damage.

Finally, impression score distributions reinforce this relationship: only 17.6% of high-rated summaries (scores 6–7) contain errors, compared to over half of those in the mid and low ranges. Together, these findings show that faithfulness errors matter for the overall impression of the summary.

7. Conclusion

In this work, we introduced HotelCheckSpan, the first span-level faithfulness dataset for the hotel domain, combining objective and subjective information across multiple accommodation descriptions. By annotating specific spans in model-generated summaries and categorizing errors as *Incorrect*, *Misleading*, or *Not Checkable*, we provide fine-grained insights into the nature of hallucinations and factual inconsistencies in LLM outputs.

Our analysis revealed that agreement between annotators varies substantially depending on the aggregation level: example-level agreement often masks nuanced disagreements visible at the span level, while span-level F1 and soft/hard variants highlight subtle differences in both span placement and error classification. Comparing human annotations with multiple LLMs further demonstrated that models differ widely in their ability to capture or replicate human judgments, underscoring the importance of reporting multiple perspectives on agreement.

HotelCheckSpan is designed to support a range of research applications, including the study of ambiguity and disagreement, the evaluation of LLMs as faithfulness judges, and the validation of automatic metrics in subjective domains.

8. Limitations

Unfortunately, we cannot release the prompt used for generating the summaries in the dataset due to proprietary intellectual property constraints. However, the full LLM error labeling prompt is available in the appendix.

9. Ethical Considerations

Human Annotations The recommended Prolific wage of £9 per hour base rate paid out to all annotators regardless of their annotation quality.[8] We paid out a bonus of £4.60 per hour to workers who passed our attention check in stage 1 and workers who received more than 3 points out of 5 in the qualification task. This ensured compliant workers received the UK living wage of £12.60 per hour.[9]

Model Inference The total cost to run the LLMs for span annotations (2-3 runs on 500 examples per model to optimize the prompt) through APIs was roughly $100.

Use of AI We used AI-assisted coding (i.e. Copilot) with the bulk being human-written, namely for the conversion of the dataset from factgenie files into a HuggingFace-friendly format. For writing, AI was used to check grammar mistakes and improve clarity and flow.

10. Acknowledgements

This research was co-funded by the European Union (ERC, NG-NLG, 101039303), the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO, and by Charles University projects GAUK 252986 and SVV project 260 821.

11. Bibliographical References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Noam Dahan and Gabriel Stanovsky. 2025. The state and fate of summarization datasets: A survey. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7259–7278, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Gemma Team. 2025. Gemma 3 technical report.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zdeněk Kasner, Ondrej Platek, Patricia Schmidtova, Simone Balloccu, and Ondrej Dusek. 2024. factgenie: A framework for span-based evaluation of generated texts. In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 13–15, Tokyo, Japan. Association for Computational Linguistics.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. 2025. Large language models as span annotators.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

OpenAI Team. 2024. Gpt-4o system card.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

[8] https://researcher-help.prolific.com/en/article/2273bd

[9] https://www.livingwage.org.uk/

Human Language Technologies, pages 4812–4829, Online. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrícia Schmidtová, Ondrej Dusek, and Saad Mahamood. 2025. Real-world summarization: When evaluation reaches its limits. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25014–25026, Suzhou, China. Association for Computational Linguistics.

PaLM 2 Team. 2023. Palm 2 technical report.

Craig Thomson and Ehud Reiter. 2020a. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 158–168, Dublin, Ireland.

Craig Thomson and Ehud Reiter. 2020b. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

12. Language Resource References

Md. Hijbul Alam, Woo-Jong Ryu, and SangKeun Lee. 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. *Information Sciences*, 339:206–223.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antognini, Diego and Faltings, Boi. 2020. *HotelRec: a Novel Very Large-Scale Hotel Recommendation Dataset*. European Language Resources Association.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Srinivas Ramesh Kamath, Fahime Same, and Saad Mahamood. 2024. Generating hotel highlights from unstructured text using LLMs. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 280–288, Tokyo, Japan. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mir Tafseer Nayeem and Davood Rafiei. 2024. Lfosum: Summarizing long-form opinions with large language models.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

A. Clustering Implementation Details

For the semantic clustering analysis described in Section 6.4, we generated span embeddings using the finetuned `sentence-transformers/all-MiniLM-L6-v2` model (Wang et al., 2020). The resulting embeddings were clustered using the `scikit-learn` implementation of K-means. To ensure robustness and reproducibility, we utilized `k-means++` initialization with 10 restarts and a fixed random seed. The optimal number of clusters ($k = 15$) was determined by computing and maximizing the silhouette score across a search grid of $k \in [2, 30]$.

B. Human Annotation Guidelines

As described in Section 4.3, annotators were divided into two groups (A and B) to test for potential bias in checkbox phrasing regarding the absence of errors. We found no significant difference in annotation behavior that could be attributed to this factor. Below, we reproduce the exact instructions and interface guidelines provided to the crowdworkers.

! For technical reasons, please use a different browser than Safari !

You will see a **collection of texts describing a hotel** and a **short summary of the texts focusing on a specific aspect of the hotel**. Your task will be to read both the texts and the summary and identify parts of the summary that contain the errors described below.

Definitions and Examples of the Errors
We present these in Table 8. In the original interface, they were presented to the annotators using Markdown formatting.

Guidelines for identifying the parts that contain an error
To mark a part of the sentence that contains the error, drag your cursor to highlight the text. **Aim to select the smallest span** that, if removed or replaced, would correct the error while allowing the rest of the sentence to remain intact.
Some summaries will not contain any errors, in such case you are expected to not annotate any spans and instead check the box saying "*I did not find any errors in this summary*" [Group A] / "*There were no errors in this summary*" [Group B], rate your overall impression and move on to the next example.

Example
Text: *Immerse yourself in Florida's culinary heritage with Latin fusion flavors at our restaurant, Blue Matisse, or sip craft cocktails at Nau Lounge.*
Summary: Experience the vibrant flavors of Latin cuisine with a modern twist at Blue Matisse restaurant. *(No span is selected)*
Explanation: This summary contains no errors, so instead of selecting any spans, just confirm the lack of errors in the checkbox below the text.

C. LLM Evaluation Details

C.1. Model Implementations

We accessed the most recent versions of the OpenAI models (GPT-4o and o3-mini) via their API in April 2025. We ran Gemma-3 locally using the Ollama framework[10] with the `gemma3:27b` checkpoint.

C.2. Main Evaluation Prompt

The default prompt used for all three models was adapted from Kasner et al. (2025).

Given the hotel descriptions: `{data}`

Annotate all the errors in the following summary: `{text}`

Output the errors as a JSON list "annotations" in which each object contains fields "reason", "text", and "annotation_type". The value of "text" is the text of the error. The value of "reason" is the reason for the error. The value of "annotation_type" is one of {0, 1, 2} based on the following list:

- 0: Not checkable: The fact in the text cannot be checked in the data.
- 1: Misleading: The fact in the text is misleading in the given context.
- 2: Incorrect fact: The fact in the text contradicts the data.

The list should be sorted by the position of the error in the text. Make sure that the annotations are not overlapping.

Example:

Data: "The closest major airports to Bomontist Suit are: Istanbul (SAW-Sabiha Gokcen Intl.) - 17.5 km / 10.9 mi Istanbul (IST-Ataturk Intl.)."
Summary: "Schiphol Airport is just a 15-minute drive from the hotel."
Output:

```
{
  "annotations": [
    {
      "reason": "Schiphol Airport is
                 incorrect as
                 the accommodation
```

[10] https://ollama.com/

Error Type	Definition	Example
Not Checkable	The summary contains information that is not mentioned anywhere in the original text. This information could either be objective (such as the presence of a swimming pool) or subjective (such as quietness).	**Text:** *A fun-filled vacation or relaxing business trip awaits you at the Holiday Inn Express & Suites Tampa Airport nestled on the beautiful waters of Tampa Bay at Rocky Point. Our hotel is minutes from the beautiful waterfront views of Tampa's Famous Riverwalk featuring miles of shops, artists and Tampa's premier dining. Our friendly and knowledgeable staff invite you to relax in the outdoor pool.* **Summary:** Enjoy stunning views of Tampa Bay and the beautiful waterfront from this pet-friendly hotel. **Explanation:** It was not mentioned whether the hotel is pet-friendly, thus this information is Not Checkable.
Misleading	The summary presents information that appears in the original text, however, it does so in a way that changes the perceived meaning. This can be due to subjective judgments (is an attraction 10 km away "close"?) or due to a word that can have multiple meanings (pool as in swimming pool or the game requiring a pool table).	**Text:** *Sheraton Düsseldorf Airport hotel is directly connected with the Terminal - in the unique location on the roof of car park P3, surrounded by 10,000m² greenery. [...] Relax from your travels or prepare for your meeting with green views.* **Summary:** Enjoy breathtaking views from the rooftop terrace and garden, offering a relaxing escape. **Explanation:** Terrace and garden are Misleading. The hotel seems to be on the roof, but there is no mention of a terrace. At the same time, 10,000m² seems unlikely to be a garden.
Incorrect	The summary contains information that either contradicts a statement from the original text (i.e the text mentioning the hotel is NOT pet-friendly, but the summary stating it is) or contains a severe error, such as using a wrong entity (e.g. place or a person), or a wrong number (for example confusion of different numbers or kilometers vs miles).	**Text:** *The closest major airports to Bomontist Suit are: Istanbul (SAW-Sabiha Gokcen Intl.) - 17.5 km / 10.9 mi Istanbul (IST-Ataturk Intl.).* **Summary:** Schiphol Airport is just a 15-minute drive from the hotel. **Explanation:** Schiphol Airport in Amsterdam is Incorrect, since the accommodation is clearly in Istanbul. In addition, 15-minute drive is Not Checkable in this context, because even though we know the distance, we don't know the expected speed of the journey.

Table 8: Definitions and Examples of Error Types presented to annotators.

```
                is in Istanbul",
      "text": "Schiphol Airport",
      "annotation_type": 2
    },
    {
      "reason": "15-minute drive cannot
                 be verified as we
                 only know the distance,
                 not journey time",
      "text": "15-minute drive",
      "annotation_type": 0
    }
  ]
}
```

C.3. Less Strict (LS) Prompt Variant

After observing that GPT-4o tended to over-annotate, we trialed a "less strict" prompt variant for the OpenAI models. This variant introduced an error-free example and explicit instructions not to penalize stylistic choices or omissions. However, as noted in Section 5, this variant ironically caused the models to annotate even more spans.

Given the hotel descriptions: `{data}`

Annotate all the errors in the following summary: `{text}`

[... JSON formatting instructions and Error Type list identical to Main Prompt ...]

Example 1: *[... Istanbul Airport example identical to Main Prompt ...]*

Example 2:

Data: "Immerse yourself in Florida's culinary heritage with Latin fusion flavors at our restaurant, Blue Matisse, or sip craft cocktails at Nau Lounge."

Summary: “Experience the vibrant flavors of Latin cuisine with a modern twist at Blue Matisse restaurant.”
Output:

```
{
  "annotations": []
}
```

Note that some details may not be mentioned in the text: do not count omissions as errors. Also do not be too strict: some facts can be less specific than in the data (rounded values, shortened or abbreviated text, etc.), do not count these as errors. Sometimes, stronger adjectives will be used to make the summary more exciting, these are also not errors. If there are no errors in the text, “annotations” will be an empty list.
