Annotating Hallucinations in Data-to-Text NLG: Implementing a Logical Framework in Different Domains

Anonymous ACL submission

Abstract

Hallucinations are a persistent challenge in natural language generation, including data-to-text. van Deemter (2024) introduced a unifying framework based on logical consequence, aiming to categorize all hallucinations through a single formal relation. We examine whether human annotators and large language models are able to apply the framework, in two data-to-text domains. Results suggest that the framework is applicable, but they also show up significant domain-dependent variation and discrepancies between human and model judgments. We also uncover several challenges that inform future work on hallucination annotation.

1 Introduction

Hallucinations, in the sense of factual inaccuracies in generated texts, are a well-documented challenge in natural language generation (NLG) (e.g., Rawte et al., 2023; Huang et al., 2025). While NLG evaluation traditionally emphasized factors like fluency and clarity (Gatt and Krahmer, 2018; Howcroft et al., 2020, *i.a.*), the growing concern over hallucinations is causing researchers to place greater emphasis on content evaluation.

Numerous efforts have been made to define and classify hallucinations, including in the context of traditional data-to-text NLG (Reiter and Dale, 2000; Narayan and Gardent, 2020; Osuji et al., 2024), whose aim is to convert input structured data, e.g., from sensors (Gatt et al., 2009), knowledge bases (Colin et al., 2016), or tables (Parikh et al., 2020), into natural language (see §2).

van Deemter (2024) offered a critique of these analyses and proposed a categorization of hallucinations based on the logical consequence ($\models$) relation that can exist between the input and the output of a data-to-text NLG system (§3.1 for details). The idea is to compare the truth-conditional content of the input and output with each other, asking whether they "match" each other (i.e., whether each is a logical consequence of the other), and if not, then why not. The resulting analysis covers all types of hallucination in terms of a single relation. It promises to enhance our understanding of factual inaccuracies committed by large language models (LLMs) and human authors alike and to offer a principled starting point for addressing questions of error severity (van Miltenburg et al., 2020) and hallucination mitigation (Ji et al., 2024).

This paper presents the first practical implementation of this logic-based framework for hallucination analysis in real-world data-to-text domains, with a particular focus on the challenges involved in its concrete implementation and annotation. Starting from the abstract notion of logical consequence, we show how this relation can be operationalized as a multi-step reasoning procedure (§3.2). Crucially, our work exposes the non-trivial adaptations required to make the framework work in practice. We then develop actionable annotation guidelines applicable across different practical domains. These guidelines are not only suitable for human annotators but also for LLMs, which could be promising because of the growing trend of employing LLMs as judges (Zheng et al., 2023; Tan et al., 2024; Bavaresco et al., 2024, *i.a.*).

The logical consequence relation between input and output can be difficult to assess in practice, with different NLG domains posing different challenges. We therefore decided to look at two very different data-to-text domains, namely: (i) descriptions of accommodation generated from database entries (henceforth, **hotel domain**), and (ii) mathematical statements generated from logical formulae (henceforth, **logic domain**). The hotel domain is characterized by simple inputs (i.e., conjunction of atomic facts), and lengthy outputs, which can be stylistically elaborate (Table 1). The logic domain uses short but potentially complex inputs, typically yielding outputs which are purely factual (Table 2).

Input
Name: Piscina Rei Star rating: 4 City: Muravera Country: Italy Accommodation type: Resort Hotel facilities: Hotel bar Room amenities: Balcony (upon inquiry)
Output
Indulge in coastal bliss at Piscina Rei Resort, a 4-star retreat in Muravera, Italy. This resort offers a tranquil haven with a hotel bar, while rooms may feature balconies (upon request).

Table 1: Input-output example from the hotel domain, categorized as "Well-matched with harmless information", following the decision tree in Figure 1.

Input
$\exists x \neg(\mathrm{Cube}(x) \rightarrow \forall y\ (\mathrm{Tet}(y) \rightarrow \mathrm{Smaller}(x, y)))$
Output
There is a cube that is not smaller than every tetrahedron.

Table 2: Input-output example from the logic domain, categorized as "Well-matched", following the decision tree in Figure 1.

We address two research questions about van Deemter's classification framework: (1) **Applicability by humans**: To what extent do human annotators agree among themselves and with reference annotations? (2) **Modeling with LLMs as judges**: Is there a realistic prospect of automating the annotation process using LLMs?

To investigate these questions, we adapted the framework to the hotel and logic domains and conducted annotation experiments. We analyzed annotator agreement and the alignment of annotators with reference annotations. We then presented the same experiment to four LLMs to assess whether the annotation process could be automated.[1]

2 Related Work

Hallucinations in NLG NLG systems can produce outputs that contain factual inaccuracies (Maynez et al., 2020; Raunak et al., 2021; Bouyamourn, 2023; Augenstein et al., 2024; Xu et al., 2025). Despite significant progress in detecting and mitigating such errors (Choi et al., 2023; Chen et al., 2023; Mishra et al., 2024; Agrawal et al., 2024; Tonmoy et al., 2024; Rawte et al., 2025), there is no consensus on how to categorize hallucinations (Guerreiro et al., 2023; Huidrom and Belz, 2023; Zhang et al., 2023).

Hallucination Annotation Annotation has been pivotal in studying hallucinations. The *SHROOM Shared Task Series has provided the community with high-quality manual annotations of hallucinations in multiple languages (Mickus et al., 2024; Vázquez et al., 2025). Other datasets have been developed for various domains and tasks (Chen et al., 2024; Niu et al., 2024), including machine translation (Zhou et al., 2021; Dale et al., 2023). We focus specifically on data-to-text NLG.

Hallucination Categorizations in Data-to-Text NLG Dušek and Kasner (2020) provided a logic-based analysis highlighting hallucination (i.e., when the output does not logically follow from the input) and omission (i.e., when the input does not logically follow from the output). A similar analysis was offered by Ji et al. (2023), who distinguished between intrinsic hallucination (i.e., output that contradicts the source) and extrinsic hallucination (i.e., output that can neither be supported nor contradicted by the source). Thomson and Reiter (2020) offered a heterogeneous analysis, categorizing errors into incorrect numbers, incorrect words, non-checkable information, and context errors.

3 Methodology

We implemented van Deemter's logic-based framework by applying a multi-step reasoning procedure. We adapted the framework to the hotel and logic domains through some modifications tailored to the challenges presented by these domains (pertaining to hallucination severity and ambiguity). We created data for annotation by retrieving inputs from two sources and generating outputs using LLMs in a data-to-text NLG setting. We also acquired reference annotations for these input-output pairs.

3.1 van Deemter's Framework

van Deemter (2024) argued that existing error classifications are in need of clarification, refinement, and extension. Suppose, for example, the input to a generator in the weather domain asserts that the temperature is above 20 degrees Celsius, whereas the text output says *The temperature is above 10 degrees*. Current classifications are unclear whether this constitutes an omission, because it would be

[1]All the codebase to reproduce the experiments described in the paper, along with the datasets, and the annotation material, will be publicly released upon acceptance.

difficult to pinpoint what part of the input fails to get expressed in the output. Furthermore, these classifications fail to make some important distinctions, such as between a case in which the input and output are *logically independent* of each other and where they *contradict* each other. Suppose, for example, the input asserts the temperature is between 10 and 20 degrees Celsius; then the output *The temperature is between 15 and 25* exemplifies the former (because it is possible for the input and output to both be true), whereas *The temperature is between 30 and 40 degrees* exemplifies the latter (because it is logically incompatible with the input). Standard classifications are also difficult to apply when outputs contain *internal inconsistencies*.

In view of these and other issues, van Deemter (2024) proposed a new categorization based on the logical consequence ($\models$) relation. By systematically examining the logical relations between the input (I) and output (O) of data-to-text NLG systems, the framework established different categories of hallucinations (Table 3).

The framework made some explicit assumptions. First, it is applicable in full only if the NLG system is tasked to express all and only the information in the input (in the classic NLG pipeline, this is the task of every step following Content Selection; Reiter 2007, 2025). Second, it only considers whether the output of the NLG system matches the input; the truth of the output in the real world is not considered.

3.2 Framework Adaptation

Question Structuring We structured the hallucination categories as a decision tree (cf. Ostyakova et al. 2023), where the path to each category is a series of binary questions (Figure 1). Their order is crucial. For instance, once an output is identified as contradictory, no further questions are necessary because, in classical logic, (a) anything follows from a contradictory statement, and (b) a contradictory statement can only follow from another statement if that other statement is itself contradictory. Logical considerations of this kind allow us to structure the annotation in such a way that only the minimal number of questions is asked. We decided to disregard the category "O tautologous" (fourth row in Table 3), because such outputs are exceedingly rare in both our domains.

Hallucination Severity In the hotel domain, some errors were far more serious than others. We therefore started by defining as *divergent* any information that is present in only one of the two information sources (i.e., it is present in either the input or the output, but not both). Hotel-related outputs often contain information that is divergent, but where this divergence is unlikely to lead to any complaints from customers. We call a piece of divergent information *factually wrong* if, despite everything the input says, the information *could* turn out to be manifestly wrong.

For instance, if the output asserts, without any basis in the input, that a hotel has a swimming pool, then this is both divergent and factually wrong. Divergent information that is not factually wrong can cover two kinds of cases. First, an output can contain subjective opinions. For example, an output can say that a hotel is cozy, without any basis in the input. This is a kind of commercial marketing that few customers would take seriously. Second, some information in the output may be inferable with high probability only (e.g., the output may describe a hotel as serving Mexican food, even though the only relevant information in the input is that the hotel is located in Mexico). In these cases, we ask annotators to mark these pieces of "harmless" hallucination as divergent but not factually wrong.

We strategically positioned the question of whether the output contains factually wrong information after determining whether the output follows from the input (Figure 1, first red node). First, we ask whether the output follows from the input in the strict sense (i.e., the output does not follow, because it contains divergent information of some kind). If it does not follow, we ask whether any of the divergent information is factually wrong.

Handling Ambiguity Ambiguity poses a challenge in hallucination annotation, as the classification of a hallucination type depends on the interpretation assigned to the output text. When an output permits multiple readings, distinct hallucination categories could be assigned to that output depending on the selected interpretation.

Logically rich outputs are prone to various types of ambiguity, leading different annotators to perceive distinct interpretations of the same text. These ambiguities include connective precedence (i.e., when it is unclear how logical connectives (e.g., *and*, *or*) bind in a sentence), quantifier scope (i.e., where it is unclear whether a given quantifier (*all*, *every*, *some*, etc.) is within the scope of another), and negation scope (i.e., where it is unclear

Case	Description	Category	Example Output
0	$I \models O$ and $O \models I$	Well-matched	*x is a 5-star hotel in Mexico.*
1	$I \models O$ and $O \not\models I$	O too weak	
1a	1 and $\not\models O$	Normal case	*x is a hotel in Mexico.*
1b	1 and $\models O$	O tautologous	*x has a star-rating of 5 or below.*
2	$I \not\models O$ and $O \models I$	O too strong	
2a	2 and $\not\models \neg O$	Normal case	*x is a child-friendly 5-star hotel in Mexico.*
2b	2 and $\models \neg O$	O contradictory	*x is a hotel in Mexico City, USA.*
3	$I \not\models O$ and $O \not\models I$	Neither follows	
3a	3 and $I \not\models \neg O$	I and O independent	*x is a child-friendly hotel.*
3b	3 and $I \models \neg O$	I and O contradictory	*x is a 5-star hotel in USA.*

Table 3: van Deemter (2024)'s classification, with examples from the hotel domain of the present study. Input is: Accom-Type(x) = Hotel $\wedge$ Country(x) = Mexico $\wedge$ Star-Rating$(x) = 5$. The output example for 1b is tautologous because 5 is the maximum quality rating. Example 2b is contradictory because Mexico City is not in the USA.

what part of a sentence is negated). Since ambiguities ended up playing a somewhat limited role in both domains, we opted not to encode ambiguity into the decision tree. In the logic domain (where outputs are more likely to contain ambiguity), annotators were instructed to first flag ambiguous outputs, as a preliminary separate step, and then proceed with their preferred interpretation.

3.3 Data Creation

Data creation followed a similar structure in both domains. We retrieved inputs from two sources: [COMPANY] database for the hotel domain and a corpus of existing formulae for the logic domain. We prompted multiple LLMs in a zero-shot setting to generate outputs, resulting in input-output pairs ready for annotation. We also acquired reference annotations to serve as a reference for evaluating annotators' responses. In our task, reference annotations are reasonable, because, in many cases, only one answer is justifiable. We allowed multiple reference labels for the limited number of cases where reference annotators agreed to disagree.

3.3.1 Hotel Domain

Input From [COMPANY] database, we retrieved five accommodations and their attributes (i.e., name, star rating, city, country, accommodation type, hotel facilities, room amenities, sport, childcare services, wellness, accessibility).

Output We used the prompt in Figure 3 (Appendix A) to generate the English descriptions of the input accommodation characteristics with five LLMs: Flan-T5-XXL (Chung et al., 2024), Mixtral 8x7B (Jiang et al., 2024), Falcon 180B (Almazrouei et al., 2023), ChatGPT (Brown et al., 2020), and Gemini 1.0 (Gemini Team et al., 2024a). Outputs had an average length of 130 words. For each input, we generated 5 descriptions (one per LLM), obtaining 25 descriptions (5 accommodations for 5 LLMs). See Appendix A for an input-output pair (Table 8).

Reference Annotations Three of the paper's authors annotated all 25 input-output pairs following the setup of §4. Each author annotated the 25 pairs individually, and subsequently, following a discussion between the three, a consensus annotation was reached for 17 out of 25 pairs. The authors agreed to disagree (2 votes vs. 1) on 8 pairs, in which the authors acknowledged that different answers were possible.

3.3.2 Logic Domain

Input We used the Grade Grinder Corpus (GGC; Barker-Plummer et al., 2011), a corpus of first-order logic formalizations of ~300 sentences made by 55k students answering exercises in Barwise et al. (2000). The students put their responses into a system and received feedback on whether the formalization was correct or incorrect. We restricted ourselves to the correct sentences in the geometrical shapes domain. From this pool, we randomly sampled in a stratified way 15 formulae, considering various aspects (i.e., length, structure, number of predicates, connectives, and quantifiers).

Output We used the prompt in Figure 2 (Appendix A) to generate the English translations of the input logical formulae with five LLMs: CodeLlama (Rozière et al., 2024), Mixtral 8x7B, Gemini 1.0, GPT-3.5 (Brown et al., 2020), and phi-3.5-mini (Abdin et al., 2024). Outputs had an average length of 35 words. For each input, we generated 5 translations (one per LLM), obtaining a total of 75 translations (15 formulae for 5 LLMs). See Appendix A for an input-output pair (Table 9).

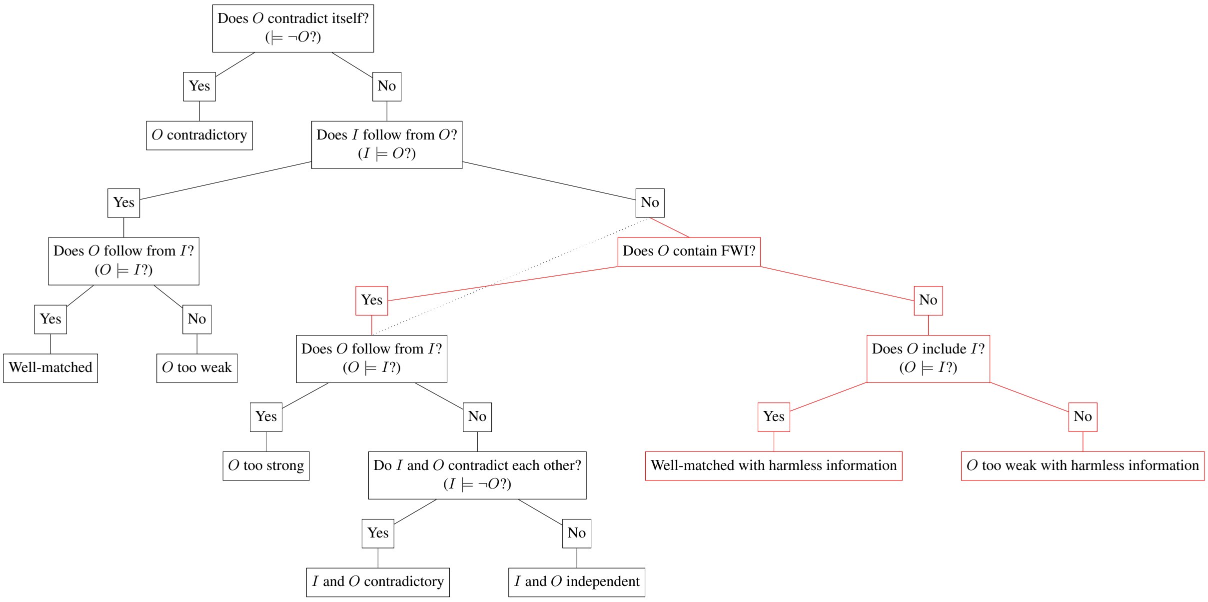

Figure 1: Framework adapted for the real-world data-to-text domains, used as the backbone for the annotation experiment. The black portion of the decision tree was used for the logic domain experiment, while the red portion was added for the hotel domain experiment. The tree illustrates the questions posed to annotators and the order in which they were presented. FWI stands for *factually wrong information*. In brackets, the original logic-based representation, which was not shown to the annotators. In the hotel domain, we used the term "include" instead of "follow from" (based on pilots with annotators). See Appendix B, E, and F for more details.

Reference Annotations Two of the paper's authors annotated all 75 input-output pairs independently, following the setup of §4. The two authors then discussed complex cases, reaching a consensus annotation on 68 out of 75 pairs. They agreed to disagree on three pairs, where ambiguity led to different hallucination categories. Four pairs were discarded, as it was impossible to determine the truth value of the outputs because the outputs were highly ungrammatical or incomplete (see §6).

3.4 Evaluation Metrics

Accuracy per Annotator (APA) To understand how often annotators agreed with the reference, we computed APA, defined as $\frac{m}{n}$, where m is the number of matches between the answers given by each annotator and the reference label(s),[2] and n is the number of pairs annotated by each annotator.

Inter-Annotator Agreement (IAA) While APA measures the agreement between annotators and reference labels, we also use Krippendorff's alpha (α; Krippendorff, 1980) to measure inter-annotator agreement. We adopt α because of its robustness in handling skewed label distributions and missing annotations (Artstein and Poesio, 2008).

[2]We consider a match to exist if the answers given by the annotators match any of the reference labels for that question.

F-Measures per Category (FPC) We computed F-measures (i.e., precision, recall, F1-score, and support) per category, to investigate annotators' performance by category. We did this by comparing, for each input-output pair, the most frequent label(s) on the annotators' side against the most frequent reference label(s).

4 Annotation with Humans

Setup We used Qualtrics to set up the annotation experiments with human annotators for both domains. We first gave annotators an interactive training session, designed to familiarize them with the concepts, terminology, and annotation interface, including definitions, guided examples, and practical exercises with feedback. To filter out annotators who had misunderstood the concepts explained in the training, we introduced a comprehension check in the hotel domain experiment (see Appendix B). During the annotation task, annotators were asked to answer a series of binary questions organized according to the decision tree in Figure 1.

In the hotel domain experiment, annotators were further asked to highlight divergent information in both the input and output and to indicate any parts in the output containing factually wrong information. In the logic domain experiment, we asked

Metric	Hotel				Logic			
	CAT	I $\models$ O	O $\models$ I	FWI	CAT	I $\models$ O	O $\models$ I	AMB
APA	0.62 (0.26)	0.86 (0.22)	0.89 (0.12)	0.84 (0.20)	0.74 (0.16)	0.85 (0.12)	0.85 (0.15)	0.80 (0.12)
IAA	0.30	0.26	0.63	0.34	0.19	0.18	0.41	0.18

Table 4: APA and IAA for human annotators for all dimensions, in the hotel and logic domains. APA scores are the means of single annotator scores; standard deviations are reported in brackets.

annotators to assess whether the output was ambiguous and, if so, to specify the type of ambiguity (i.e., connective precedence, quantifier scope, negation scope; see §3.2). We emphasized that they had to stick to a single interpretation throughout the annotation of a given input-output pair. See Appendix B for details on the annotation setup.

For the hotel domain experiment, we recruited 177 participants (from Prolific and [COMPANY]; median age = 35; male = 48.9%, female = 49.4%, non-binary = 1.7%). For the logic domain experiment, we recruited 16 experts with a strong knowledge of mathematical logic (median age = 35; male = 75.0%, female = 25.0%), who are professional contacts of the authors, unfamiliar with our research questions. Participants were randomly assigned to n groups (n = 5 for the hotel domain, and n = 15 for the logic domain) and rotated through a 5 (LLMs) $\times$ n (inputs) Latin square (Fisher, 1925). This ensured that each input-output pair was shown to approximately the same number of participants, that every participant saw all the inputs, and that each participant only encountered one LLM-generated output per original input.

Results In the hotel domain, 54 out of 177 participants passed the comprehension check. These 54 participants annotated 5 input-output pairs each, resulting in ~11 annotations per pair (270 in total). In the logic domain, the 16 experts in logic annotated 15 input-output pairs each, resulting in ~3 annotations per pair (240 in total). In both experiments, the average completion time was ~45 minutes.

For the hotel domain, we computed APA and IAA (i) on the final categories (**CAT**; resulting from traversing the decision tree; eight possible outcomes), (ii) on the question of whether **I** $\models$ **O** (two possible outcomes), (iii) on the question of whether **O** $\models$ **I** (two possible outcomes), and (iv) on the question of whether O contains factually wrong information (**FWI**; two possible outcomes: the output contains factually wrong information or not). For the logic domain, we computed the metrics (i) on the final categories (**CAT**; six possible outcomes), (ii) on the question of whether **I** $\models$ **O**, (iii) on the question of whether **O** $\models$ **I**, and (iv) on the question of whether O is ambiguous (**AMB**; two possible outcomes: the output is ambiguous or not). We interpret performance on the CAT dimension as indicative of how well annotators engage with the framework as a whole. The $I \models O$ and $O \models I$ dimensions represent the core inferential questions regarding logical consequence, and provide insight into how annotators perform on the higher-level reasoning tasks central to the framework. Refer to Table 4 for the figures.

Category	**P**	**R**	**F1**	**S**	**#**
		Hotel			
Well-matched	1.00	0.50	0.67	2	1
Well-matched (harmless)	0.80	1.00	0.89	8	10
O too weak	0.67	1.00	0.80	2	3
O too weak (harmless)	1.00	0.62	0.77	8	5
O too strong	0.50	0.25	0.33	4	2
I and O independent	0.86	0.75	0.80	8	7
I and O contradictory	0.00	0.00	0.00	1	0
Macro Average	0.69	0.59	0.61		
		Logic			
Well-matched	0.92	0.93	0.93	61	62
O too weak	0.25	0.50	0.33	4	8
O too strong	0.33	0.20	0.25	5	3
O contradictory	0.00	0.00	0.00	0	3
I and O independent	0.43	0.75	0.55	4	7
I and O contradictory	0.00	0.00	0.00	0	1
Macro Average	0.32	0.40	0.34		

Table 5: FPC for human annotators, in the hotel and logic domain: precision (P), recall (R), F1-score, support (S), annotators' count (#), and macro averages.

APA ranges from good to very good in all dimensions, indicating that annotators tended to agree with the reference annotations. APA for CAT is lower compared to APA for the other (binary) dimensions, which is expected for two reasons: (i) CAT involves a larger number of possible categories (8 or 6, vs. only 2), and (ii) the hierarchical nature of the decision tree presupposes answering the intermediate questions accurately to reach the correct final category. IAA ranges from low to moderate. Interestingly, IAA for $O \models I$ is consistently higher than for $I \models O$ in both domains.

To investigate annotators' performance across the different categories, we computed FPC for both domains (Table 5). Based on macro averages, annotators performed better in the hotel domain. Note that some categories are absent (e.g., "O contradictory" in the hotel domain), while others are un-

	Gemini 1.5				Llama 3.3				o1-mini				Grok-2				
Category	**P**	**R**	**F1**	**#**	**P**	**R**	**F1**	**#**	**P**	**R**	**F1**	**#**	**P**	**R**	**F1**	**#**	**S**
									Hotel								
Well-matched	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.17	0.50	0.25	6	0.00	0.00	0.00	0	2
Well-matched (harmless)	0.00	0.00	0.00	0	0.36	1.00	0.53	22	0.80	0.50	0.62	5	0.00	0.00	0.00	0	8
O too weak	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.25	0.50	0.33	4	0.00	0.00	0.00	0	2
O too weak (harmless)	0.32	1.00	0.48	25	0.50	0.12	0.20	2	0.30	0.38	0.33	10	0.36	1.00	0.53	22	8
O too strong	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	4
I and O independent	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	8
I and O contradictory	0.00	0.00	0.00	0	0.00	0.00	0.00	1	0.00	0.00	0.00	0	0.00	0.00	0.00	3	1
Macro Average	0.05	0.14	0.07		0.12	0.16	0.10		**0.22**	**0.27**	**0.22**		0.05	0.14	0.08		
									Logic								
Well-matched	0.94	0.95	0.94	62	0.98	0.89	0.93	55	1.00	0.56	0.72	34	0.96	0.84	0.89	53	61
O too weak	0.00	0.00	0.00	0	0.00	0.00	0.00	6	0.13	0.50	0.21	15	0.18	0.50	0.27	11	4
O too strong	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.27	0.60	0.37	11	0.00	0.00	0.00	0	5
O contradictory	0.00	0.00	0.00	1	0.00	0.00	0.00	1	-	-	-	0	-	-	-	0	0
I and O independent	0.00	0.00	0.00	0	0.00	0.00	0.00	1	0.27	0.75	0.40	11	0.00	0.00	0.00	0	4
I and O contradictory	0.00	0.00	0.00	8	0.00	0.00	0.00	8	-	-	-	0	0.00	0.00	0.00	7	0
Macro Average	0.16	0.16	0.16		0.16	0.15	0.16		**0.42**	**0.60**	**0.43**		0.23	0.27	0.23		

Table 6: FPC for LLMs, in the hotel and logic domain: precision (P), recall (R), F1-score, and LLM count (#), and macro averages, with support (S) as the last column shared by all LLMs. - indicates that # for that LLM and S were both 0. Boldfaced are the best macro averages per LLM per metric per domain.

	Hotel				Logic			
Model	**CAT**	**I $\models$ O**	**O $\models$ I**	**FWI**	**CAT**	**I $\models$ O**	**O $\models$ I**	**AMB**
Gemini 1.5	0.32	**0.88**	0.64	**0.76**	**0.82**	0.87	**0.90**	0.72
Llama 3.3	**0.36**	**0.88**	0.56	**0.76**	0.76	0.93	0.86	0.70
o1-mini	**0.36**	0.72	**0.72**	**0.76**	0.59	0.76	0.70	**0.75**
Grok-2	0.32	**0.88**	0.64	**0.76**	0.75	**0.96**	0.82	**0.75**

Table 7: APA for LLMs for all the dimensions, in the hotel and logic domain. Boldfaced are the higher values per LLM per dimension per domain.

derrepresented (e.g., "Well-matched" in the hotel domain). In the logic domain, the distributions are skewed toward the "Well-matched" category because many LLM-generated translations are near-literal renditions of the input formula, which tend to be faithful to the input, although they are often far from fluent. We follow up with additional domain-specific analyses in Appendix C.

5 Annotation with LLMs

Setup We conducted annotation experiments with LLMs for both domains in a way that deviated as little as possible from the human experiment. LLMs underwent the same interactive training with feedback. No LLM passed the comprehension check. We proceeded nonetheless with the annotation experiment, and made the LLMs annotate all input-output pairs (25 in the hotel domain and 75 in the logic domain) step by step, following the setup in §4. We chose the following four LLMs, which represent a variety of open-weight and proprietary models among the top-performing ones from the living benchmark proposed in White et al. (2025): Gemini 1.5 (Gemini Team et al., 2024b), Llama 3.3 70B (Grattafiori et al., 2024), o1-mini (OpenAI et al., 2024), and Grok-2. See Appendix D for further details on the experimental setup.

Results For all LLMs, we computed APA on the final categories (CAT), on the question of whether $I \models O$, and on the question of whether $O \models I$, for both domains, and on the question of whether O contains factually wrong information (FWI) in the hotel domain, and on the question of whether O is ambiguous (AMB) in the logic domain (see §4).

Based on APA, we observe that no LLM clearly stands out, with comparable figures both in-domain and across domains, except for CAT (Table 7). Note the consistently low scores for CAT in the hotel domain and the relatively poor performance of o1-mini in CAT in the logic domain. The figures become more nuanced when breaking down the analysis per category (see FPC in Table 6). Based on macro averages, o1-mini emerges as the top-performing model in both domains. However, generally, we found consistently low scores in most categories. Most LLMs (except for o1-mini) tended to provide highly repetitive answers, leading to skewed hallucination category distributions in both domains (columns #). In the logic domain, this behavior is expected, as many outputs are near-literal renditions of the input formulae (i.e., inputs and outputs are "Well-matched"), making them relatively straightforward for models to get right. In the hotel domain, where the reference labels distribution was more balanced, this pattern is less justified. Note also that, in the logic domain, the "O contradictory" category is empty for o1-mini and Grok-2, and the "I and O contradictory" category is empty in o1-mini, which is consistent with the reference label distributions (column S).

6 Discussion and Conclusion

Returning to the research questions of §1, the results of our experiments suggest that human annotation using the framework of van Deemter (2024) was feasible, but that the LLMs we investigated were not quite up to this annotation task (although, in the logic domain, o1-mini achieved better category-level performance than human annotators; Tables 5 and 6). For both human and LLM annotation, the reasonably high APA scores for $I \models O$ and FWI (Tables 4 and 7) suggest that the main adaptation that we made to the framework, in which divergent information was separated into yes/no *factually wrong* (§3.2) were effective. While interpreting these results, it is important to consider several factors that emerged throughout this study.

Hallucination annotation is challenging. Although annotation was feasible, it was far from easy. In the hotel domain, this was clear from the large proportion of would-be annotators who failed our comprehension check (§4). Annotators for the logic annotation task were not filtered in this way, because our recruitment process guaranteed a high level of expertise in judging logical consequence relations. Clearly, detailed annotation of different hallucination types requires close attention.

Performance on ambiguous output may look better than it is. An obstacle against hallucination annotation anticipated in van Deemter (2024) is the *ambiguity* of outputs. Ambiguous outputs were rare in the hotel domain but more frequent in the logic domain. For example, some LLMs used the word *otherwise* ambiguously. For instance, in *If c is larger than e, then b is larger than c. Otherwise, c is not larger than e*, the word *otherwise* can negate the antecedent, it can negate the consequent, or it can mean *or*. Ambiguities did not hurt the APA, IAA, and FPC metrics much because, as evidenced by the comments entered by annotators, when annotators encountered an ambiguous output, they tended to interpret the output "charitably", choosing a well-matched interpretation of the output whenever one was available (see Appendix C.2).

The distribution of categories was skewed. Our use of *bona fide* corpora (§3.3) suggests that the input-output pairs we studied have some real-world validity. However, it also had the unanticipated effect that (as noted in §4) the distribution of categories in the logic domain was skewed towards the "Well-matched" category. Since well-matched pairs were often relatively easy to judge, particularly when the output was highly formulaic (e.g., as in *For all z and for all y, if z is behind y, then z is larger than y*), this may make our results overly optimistic.

Inputs can be underspecified. Researchers in data-to-text NLG often assume that their inputs are well defined, but our evaluation of the hotel domain showed cases in which this assumption was not met. For example, if the input said *Room amenities: Sitting area*, it was unclear whether this pertained to *all* the rooms in the hotel, justifying the output *This hotel offers* [...] *amenities including a* [...] *sitting area in each room* (see Appendix C.1 for an example). Input ambiguities occurred in the logic domain as well, for instance when the input $\forall x \neg(\text{Adjoins}(a, x) \vee \text{Adjoins}(x, a))$ was rendered as *Nothing adjoins a*, which is a perfect match if and only if "Adjoins" is interpreted as a symmetric relation. Such variations in interpretation led to conflicting hallucination category assignments.

LLM outputs can be ill-formed. In the logic domain, LLMs sometimes produced English outputs that are so ungrammatical that it is impossible to say whether they follow from a given input (§3.3). This happened especially where inputs contained vacuous quantifiers (i.e., which do not bind any variables), e.g., *For all x and for all y, it is not true that for all y, x is larger than y*, where the double *for all* is hard to make sense of in natural language. We do not know how replicable this phenomenon will prove to be for better or more elaborately tuned models, but it appears to justify a new category "Output not well-formed", to be added to van Deemter (2024)'s framework (Table 3).

In conclusion, we have shown that, with some domain-specific adaptations, the relation of logical consequence between input and output in data-to-text NLG can be decomposed into manageable, human-annotatable reasoning steps. Our findings underscore the inherent complexity of the task: while human performance is reasonable, model performance remains poor. These results call for caution in the design of hallucination annotation studies and emphasize the importance of carefully calibrated annotation guidelines, alongside robust theoretical foundations and practical foundations, e.g., input data quality.

Limitations

We only focused on two data-to-text NLG domains. Obstacles to hallucination annotation other than the ones noted above may come to the fore in different domains. For example, domains in which numerical input plays an important role, such as weather forecasting (e.g., Reiter et al., 2005; González Corbelle et al., 2022), are likely to give rise to outputs that are *vague* (e.g., when a temperature of 25 degrees is described as *warm*, annotators may disagree whether this output does or does not follow from the input). We expect that vagueness will give rise to similar problems as ambiguity, and that these problems can be addressed along similar lines.

Our data creation process in the logic domain produced many outputs containing formulaic elements (see §4 and §6). Although we applied basic prompt engineering strategies to generate outputs (see Appendix A), alternative models or more extensive prompt engineering might yield higher-quality text. Nonetheless, it is important to note that our primary research questions focus on the applicability of an annotation framework in practical domains, and not on maximizing the quality of generated outputs.

In our annotation experiments, we relied on a limited number of LLMs to perform the annotation experiments. Although our results align with prior findings, that LLMs struggle with complex reasoning tasks (e.g., Huckle and Williams, 2025; Li et al., 2024), the models we tested may not be fully representative. Future work should expand the model pool to assess whether other LLMs can achieve better performance.

The relatively small number of input-output pairs on which our study was based (i.e., 25 for the hotel domain and 75 for the logic domain) may limit the generalizability of our findings. Future work should look at a larger number of input-output pairs.

Ethical Considerations

Ethical approval for the human experiments conducted in this study was obtained from the Ethics Board at [INSTITUTION]. All the annotators gave informed consent before participating in the experiment. The 12 [COMPANY] employees and the 16 experts in logic volunteered to participate without remuneration. The 165 crowdworkers recruited on Prolific were paid £3 for completing the training, and those who successfully passed the comprehension check were paid an additional £3 upon completion of the annotation experiment, which corresponds to £6 per hour, matching the minimum pay according to Prolific.

All the experiments involving LLMs, i.e., data creation (§3.3) and annotation (§5), cost us ~€90.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. ArXiv:2404.14219 [cs].

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can knowledge graphs reduce hallucinations in LLMs? : A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *arXiv preprint*. ArXiv:2311.16867 [cs].

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863. Publisher: Nature Publishing Group.

Dave Barker-Plummer, Richard Cox, and Robert Dale. 2011. Student translations of natural language into logic: the Grade Grinder Corpus release 1.0. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 51–60.

Jon Barwise, John Etchemendy, Gerard Allwein, Dave Barker-Plummer, and Albert Liu. 2000. *Language, Proof and Logic*. CSLI publications.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *arXiv preprint*. ArXiv:2406.18403 [cs].

Adam Bouyamourn. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Eduardo Calò, Lydia Penkert, and Saad Mahamood. 2025. Lessons from a user experience evaluation of NLP interfaces. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2915–2929, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lyu, Dan Zhang, and Huajun Chen. 2024. FactCHD: Benchmarking Fact-Conflicting Hallucination Detection. In *Thirty-Third International Joint Conference on Artificial Intelligence*, volume 7, pages 6216–6224. ISSN: 1045-0823.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 245–255, New York, NY, USA. Association for Computing Machinery.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. KCTS: Knowledge-constrained tree search decoding with token-level hallucination detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14035–14053, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653, Singapore. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ronald Aylmer Fisher. 1925. *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Loyd.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170.

Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. *AI Commun.*, 22(3):153–186.

Google Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024a. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint*. ArXiv:2312.11805 [cs].

Google Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and

1118 others. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*. ArXiv:2403.05530 [cs].

Javier González Corbelle, Alberto Bugarín-Diz, Jose Alonso-Moral, and Juan Taboada. 2022. Dealing with hallucination and omission in neural natural language generation: A use case on meteorology. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 121–130, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7. Publisher: Frontiers.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

James Huckle and Sean Williams. 2025. Easy Problems that LLMs Get Wrong. In *Advances in Information and Communication*, pages 313–332, Cham. Springer Nature Switzerland.

Rudali Huidrom and Anya Belz. 2023. Towards a consensus taxonomy for annotating errors in automatically generated text. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 527–540, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

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 Comput. Surv.*, 55(12):248:1–248:38.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of Experts. *arXiv preprint*. ArXiv:2401.04088 [cs].

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA. Section: 12.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. 2024. LLMs for Relational Reasoning: How Far are We? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, LLM4Code '24, pages 119–126, New York, NY, USA. Association for Computing Machinery.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.

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.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained Hallucination Detection and Editing for Language Models. In *First Conference on Language Modeling*.

Shashi Narayan and Claire Gardent. 2020. *Deep Learning Approaches to Text Production*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254, Prague, Czechia. Association for Computational Linguistics.

Chinonso Cynthia Osuji, Thiago Castro Ferreira, and Brian Davis. 2024. A Systematic Review of Data-to-Text NLG. *arXiv preprint*. ArXiv:2402.08496 [cs].

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Vipula Rawte, S.m Towhidul Islam Tonmoy, Shravani Nag, Aman Chadha, Amit Sheth, and Amitava Das. 2025. FACTOID: FACtual enTailment fOr hallucInation detection. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 599–617, Albuquerque, New Mexico. Association for Computational Linguistics.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104, Saarbrücken, Germany. DFKI GmbH.

Ehud Reiter. 2025. *Natural Language Generation*. Springer Nature Switzerland, Cham.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, and 7 others. 2024. Code Llama: Open Foundation Models for Code. *arXiv preprint*. ArXiv:2308.12950 [cs].

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2024. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Craig Thomson and Ehud Reiter. 2020. 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.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv preprint*. ArXiv:2401.01313 [cs].

Kees van Deemter. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Emiel van Miltenburg, Wei-Ting Lu, Emiel Krahmer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. 2020. Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 398–411, Dublin, Ireland. Association for Computational Linguistics.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent

Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona de Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared Task on Hallucinations and Related Observable Overgeneration Mistakes. *arXiv preprint*. ArXiv:2504.11975 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. LiveBench: A Challenging, Contamination-Free LLM Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint*. ArXiv:2401.11817 [cs].

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint*. ArXiv:2309.01219 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Details on Data Creation

Figure 3 shows the prompt we used to generate the textual descriptions of the input accommodation characteristics. Figure 2 shows the prompt we used to generate the textual translations of the input logical formulae.

In the hotel domain, for each attribute, we considered 1 to 3 values (chosen randomly), to avoid excessively complex inputs, as some attributes, e.g., hotel facilities or room amenities, present long lists of values.

We used the Hugging Face (Wolf et al., 2020) inference API for all models (with default parameters), except ChatGPT (for which we used the web interface (model GPT-3.5) accessed on February 20, 2024), GPT-3.5 (for which we used the dedicated API), and Gemini 1.0 (for which we used the dedicated API). Table 8 and Table 9 show input-output pair examples from the hotel and the logic domains, respectively.

Translate the following formula into English.
The following is the meaning of the predicates used in the formula:

SameSize(x, y) :	x and y are the same size.
Smaller(x, y) :	x is smaller than y.
SameCol(x, y) :	x and y are in the same column.
Larger(x, y) :	x is larger than y.
BackOf(x, y) :	x is behind y.
Medium(x) :	x is medium.
Large(x) :	x is large.
FrontOf(x, y) :	x is in front of y.
Adjoins(x, y) :	x adjoins y.
Small(x) :	x is small.
Between(x, y, z) :	x is between y and z.
LeftOf(x, y) :	x is to the left of y.
Cube(x) :	x is a cube.
Dodec(x) :	x is a dodecahedron.
RightOf(x, y) :	x is to the right of y.
SameRow(x, y) :	x and y are in the same row.
SameShape(x, y) :	x and y are the same shape.
Tet(x) :	x is a tetrahedron.

ONLY RETURN THE TRANSLATION. DO NOT USE LOGICAL SYMBOLS. DO NOT GIVE ANY EXPLANATION.

Formula: {input_formula}
Translation:

Figure 2: Prompt used for the generation of textual translations of input logical formulae.

Create a detailed description of an accommodation with the following characteristics:

{input_characteristics}

Figure 3: Prompt used for the generation of textual descriptions of the input accommodation characteristics.

B Details on Annotation with Humans

Piloting We designed the final annotation experiment through several rounds of piloting. In the initial phases, we learned several key lessons. Logic terminology proved challenging to convey, requiring multiple revisions (for example, in the hotel domain, we learnt to avoid terms like “logical consequence” or “follows from”, and instead use the term “inclusion”, which was better understood; see Appendix E). The design of the user interface played a fundamental role in the annotation process (cf. Calò et al. 2025), including choices such as positioning input-output pairs side-by-side, indenting input logical formulae, and selecting effective highlighting methods. The length of input-output pairs influenced annotation quality, leading us to limit the number of attribute-value pairs in the hotel domain and balance input formula lengths in the logic domain to mitigate annotator fatigue. The choice of annotation platform also impacted scalability; while we initially experimented with Label Studio, it proved difficult to scale beyond the piloting phase. Training the annotators turned out to be crucial due to the complexity of the task, making comprehension checks indispensable.

Comprehension Check The comprehension check for the hotel domain experiment[3] assessed participants’ understanding of (i) logical consequence and (ii) the distinction between divergent and factually wrong information. It was designed as follows. The input (I) was retrieved from [COMPANY] database, following the same methodology described in §3.3.1. The output (O) was generated using a simple rule-based generator, with the manual addition of some words: *alluring* (expected to be highlighted as divergent information) and *swimming pool* (expected to be highlighted as divergent information and considered factually wrong). See

[3] We did not implement this step for the logic domain, as the intended participants consisted of experts in logic. We assumed that a comprehension check would be unnecessary, especially after the training designed to align their understanding with the relevant concepts.

Input	Output
Name: Hotel The St Regis Punta Mita Resort Star rating: 5 City: Bahía de Banderas Country: Mexico Accommodation type: Hotel Hotel facilities: Laundry service Room amenities: Free WiFi (rooms) Sport: Horse riding, Diving Childcare services: Organised activities for kids, Playground Wellness: Beauty salon, Massage, Hydrotherapy Accessibility: Accessible bathroom, Accessible hotel	Indulge in luxury at the 5-star Hotel The St Regis Punta Mita Resort in Bahía de Banderas, Mexico. This hotel offers a serene escape with its accessible accommodations and soothing wellness amenities, including a beauty salon, massage treatments, and hydrotherapy. Keep entertained with thrilling sports like horse riding and diving. Families can enjoy organized activities and a playground for kids. The hotel also provides convenient laundry services and ensures a comfortable stay with free in-room WiFi.

Table 8: Input-output pair from the hotel domain. The input is retrieved from [COMPANY] database, and the output is generated by Gemini 1.0. Following the tree in Figure 1, this pair was annotated as follows. Reference label distribution (§3.3): "O too weak with harmless information": 0.67; "I and O independent": 0.33. Human annotators label distribution (§4): "Well-matched": 0.09; "Well-matched with harmless information": 0.27; "O too weak": 0.09; "O too weak with harmless information": 0.55. LLM annotators label distribution (§5): "Well-matched": 0.25; "Well-matched with harmless information": 0.25; "O too weak with harmless information": 0.50.

Input	Output
$\exists x \forall y \forall z\ ($ $\mathrm{Cube}(x) \wedge \mathrm{Large}(x) \wedge$ $($ $(\mathrm{Cube}(y) \wedge \mathrm{Large}(y) \wedge \mathrm{Dodec}(z)) \rightarrow$ $(x = y \wedge \neg \mathrm{BackOf}(z, y))$ $)$ $)$	There exists a cube that is large, and if there is another cube that is large and a dodecahedron, then the first cube is the same as the second cube and is not behind the dodecahedron.

Table 9: Input-output pair from the logic domain. The input is retrieved from the GGC, and the output is generated by GPT-3.5. Following the tree in Figure 1, this pair was annotated as follows. Reference label distribution (§3.3): "Well-matched": 0.50; "O too weak": 0.50. Human annotators label distribution (§4): "Well-matched": 0.33; "I and O independent": 0.33; "O contradictory: 0.33. LLM annotators label distribution (§5): "Well-matched": 0.75; "O too weak": 0.25.

Table 10 for the comprehension check itself. The input does not contain divergent information, while the output contains both divergent and factually wrong information. The output does not contradict itself, $I \not\models O$, and $O \models I$.

Annotators failed the comprehension check for several reasons. First, most did not highlight the term *alluring*, which was intended to test their ability to identify divergent information that is not factually wrong. Second, some annotators mistakenly judged the output as self-contradictory simply because of the presence of *swimming pool*, which was meant to be recognized as factually wrong information. Third, a few annotators struggled with the concept of logical consequence: they correctly marked *alluring* and/or *swimming pool* as extra content, yet still claimed that $I \models O$.

Training Material and Annotation Interface For both the hotel and logic domain experiments, annotators used the same interface for both training and annotation. The interactive training materials and annotation interface were slightly adapted between the two experiments to accommodate the specific characteristics of each domain.

The interactive training included exercises with adaptive feedback tailored to the annotators' responses. This allowed annotators to receive immediate clarification and guidance when their answers deviated from expectations. The content of the interactive training focused on the concepts of logical consequence, and on divergent and factually wrong information (for the hotel domain experiment). At the end of the training, annotators could download a document summarizing the key points (Appendix E for the document in the hotel domain, and Appendix F for the document in the logic domain).

As an example, Figure 4 presents the practical exercise on identifying divergent information in the hotel domain, while Figure 5 illustrates the corresponding feedback provided to the annotators. Figure 6 and Figure 7 show the annotation interfaces for the hotel and logic domain experiments, respectively.

PRACTICAL EXERCISE on *Divergent Information*:

Now it's your turn!

In the following example, highlight Divergent Information in Input and Divergent Information in Output.

You can select one word or multiple words at once. For single-word selection, hover over the word you want to highlight, then click. When a tag pops up above the word, click on it. For multiple-word selection, position yourself at the beginning of the span of text you want to highlight, double click and hover over the span of text. When you release the click, a tag will pop up. Click on it.

If you want to remove a tag, just do the same as when highlighting single or multiple words, and when the tag pops up, click on "Remove".

When you are done with highlighting, go to the next page, to receive feedback.

Input

Name: Four Seasons Resort Sharm El Sheikh
Star rating: 5
City: Sharm el-Sheikh
Country: Egypt
Accommodation type: Hotel
Hotel facilities: Pets allowed, Conference rooms, Luggage storage
Room amenities: Balcony (upon inquiry)
Sport: Table tennis, Tennis court
Childcare services: Organised activities for kids
Wellness: Steam room
Accessibility: Accessible parking

Output

The Four Seasons Resort Sharm El Sheikh is a luxurious 5-star hotel located in the heart of Sharm el-Sheikh, Egypt.
Many of the rooms offer stunning views, and some even have balconies that can be requested upon inquiry.
The hotel also has a range of dining options.
For those who like to stay active, the hotel has a range of sport facilities, including a tennis court and table tennis.
If you are traveling with children, the hotel has organized activities for kids.
In addition, the hotel has a range of conference rooms, making it an ideal choice for business travelers.
There is also luggage storage available for those who need to store their bags before or after check-in.
For those with accessibility needs, the hotel offers accessible parking.
And if you are traveling with pets, the hotel is pet-friendly!

Figure 4: Practical exercise on identifying divergent information in the hotel domain.

There are many ways of highlighting Divergent Information, and opinions may differ among people. However, the lower bound (i.e., the information that you should at least highlight) for this Input - Output pair is:

Wellness: Steam room
The hotel also has a range of dining options.

Remember: We are interested in Divergent Information in the broadest possible sense; therefore, ANY piece of information (even if its content might be taken for granted, or it is a subjective opinion, or contradicts another piece of information, etc.) that is present in one of the two information sources but not in the other may be highlighted.

The following is a way of highlighting Divergent Information in this Input - Output pair. Many other ways may be correct, and it is subject to your opinion.

Input

Name: Four Seasons Resort Sharm El Sheikh
Star rating: 5
City: Sharm el-Sheikh
Country: Egypt
Accommodation type: Hotel
Hotel facilities: Pets allowed, Conference rooms, Luggage storage
Room amenities: Balcony (upon inquiry)
Sport: Table tennis, Tennis court
Childcare services: Organised activities for kids
Wellness: Steam room
Accessibility: Accessible parking

Output

The Four Seasons Resort Sharm El Sheikh is a luxurious 5-star hotel located in the heart of Sharm el-Sheikh, Egypt.
Many of the rooms offer stunning views, and some even have balconies that can be requested upon inquiry.
The hotel also has a range of dining options.
For those who like to stay active, the hotel has a range of sport facilities, including a tennis court and table tennis.
If you are traveling with children, the hotel has organized activities for kids.
In addition, the hotel has a range of conference rooms,
making it an ideal choice for business travelers.
There is also luggage storage available for those who need to store their bags
before or after check-in.
For those with accessibility needs, the hotel offers accessible parking.
And if you are traveling with pets, the hotel is pet-friendly!

Figure 5: Feedback provided to the annotators on the exercise on identifying divergent information in the hotel domain.

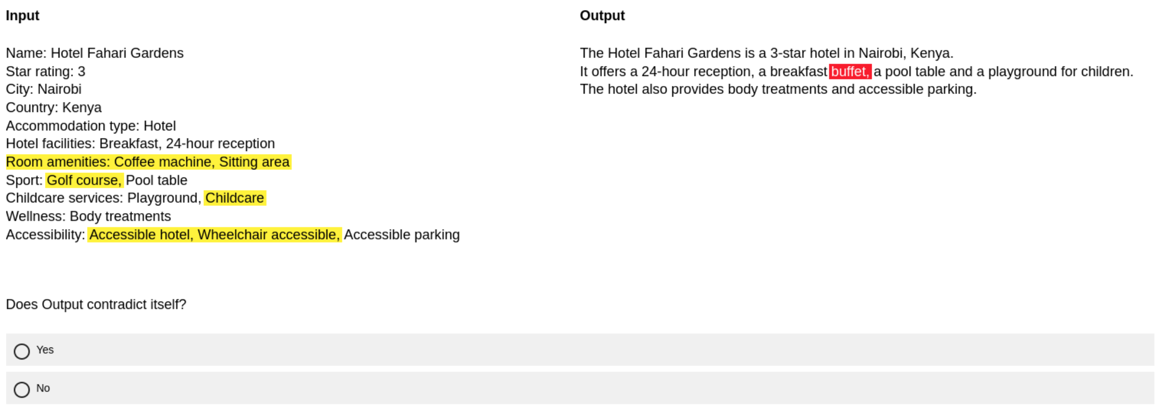

Figure 6: Annotation interface for the hotel domain with an experimental item. Divergent information in input and output is highlighted, and the first question is displayed. Subsequent questions would be revealed incrementally, based on the annotator's responses and the decision tree shown in Figure 1.

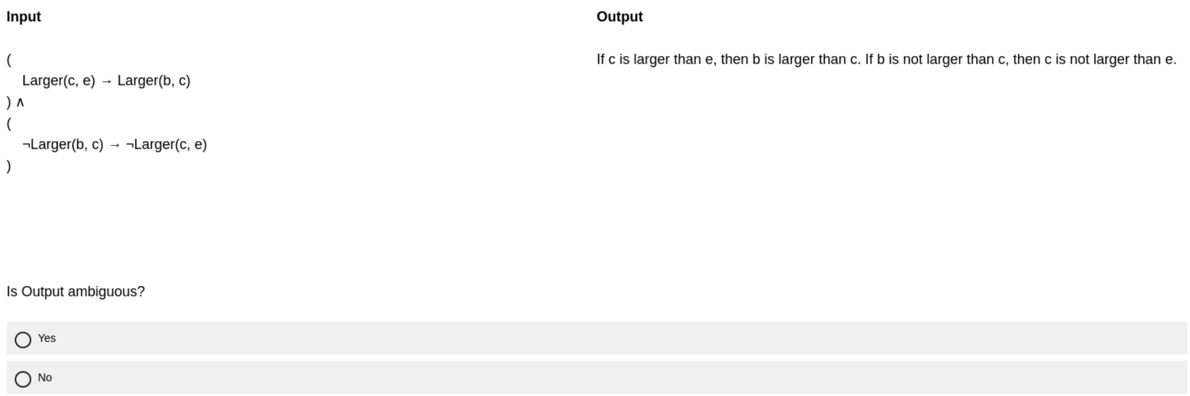

Figure 7: Annotation interface for the logic domain with an experimental item. The preliminary question on ambiguity is displayed. Subsequent questions would be revealed incrementally, based on the annotator's responses and the decision tree shown in Figure 1. The input formulae were presented in an indented form to improve readability.

Input	Output
Name: Aktiv Panoramahotel Daniel Star rating: 4 City: Sautens Country: Austria Accommodation type: Hotel Hotel facilities: Hotel safe Room amenities: Fridge, Cable TV Sport: Volleyball, Pool table Wellness: Beauty salon, Steam room, Body treatments Accessibility: Wheelchair accessible	The Aktiv Panoramahotel Daniel is an alluring 4-star hotel located in Sautens, Austria. Hotel facilities include hotel safe, swimming pool. Room amenities include fridge, cable TV. Sports facilities include volleyball, pool table. Wellness facilities include beauty salon, steam room, body treatments. Accessibility features include wheelchair accessible.

Table 10: Comprehension check for the hotel domain. The input is retrieved from [COMPANY] database, and the output is generated by a simple rule-based generator.

C Additional Human Results

This section includes detailed analyses of divergent and factually wrong information in the hotel domain, and details on ambiguity in the logic domain.

C.1 Hotel Domain

Divergent and Factually Wrong Information We analyzed the divergent information that was highlighted, and the factually wrong information reported by the annotators, to obtain more fine-grained insights.

To retrieve these spans of text, we followed two different procedures. Divergent information was highlighted using the Qualtrics interface, which returns the indices of the tokens highlighted in the original text. Factually wrong information was provided as free text (copied and pasted or written by the annotators, and might contain typos), so we could not straightforwardly retrieve the indices of the tokens in the original text. Thus, we aligned the factually wrong information provided by each annotator with the original text using CollateX, and then retrieved the indices of the tokens in the original text.

To study the extent to which output texts are "hallucinated" (i.e., they contain divergent or factually wrong information), we computed the ratio of hallucinated tokens over the total number of tokens for each item (i.e., input-output pair) and annotator. These ratios were then averaged across annotators and items to obtain overall proportions for divergent information in input and output and factually wrong information. See Table 11 for the figures. On average, items contain more divergent information in the output, often consisting of harmless additions such as *warming and inviting atmosphere*, which are expected in hotel descriptions. By contrast, factually wrong information is the least frequent, suggesting that models are generally relatively good at avoiding more severe hallucinations (e.g., falsely adding amenities).

Information Type	Ratio
Divergent information (Input)	0.08
Divergent information (Output)	0.19
Factually wrong information	0.02

Table 11: Ratios of divergent and factually wrong information over the original texts, normalized by length of inputs and outputs.

Intuition vs. Hallucinated Content During annotation, for each input-output pair, annotators also provided a judgment on a 7-point Likert scale (Likert, 1932), assessing their overall impression of the faithfulness of the output with respect to the input.

To study how the overall impression correlates with the presence of hallucinated information, we computed the Pearson correlation between the ratios of divergent and factually wrong information and mean slider ratings across all items. Refer to Table 12 for the figures.

We find a significant negative correlation between slider ratings and the presence of divergent information in the input, suggesting that such divergence lowers perceived faithfulness. The correlation for divergent information in the output is non-significant, possibly due to annotators having different perceptions of harmless added content. In contrast, factually wrong information shows the

Information Type	r	p
Divergent information (Input)	-0.51	0.009
Divergent information (Output)	-0.33	0.112
Factually wrong information	-0.58	0.002

Table 12: Pearson correlation (r) and p-values (p) between divergent and factually wrong information vs. slider ratings.

strongest significant negative correlation, which aligns with expectations; factual errors directly undermine the perceived faithfulness of an output.

Annotator Agreement over Hallucinated Content To study the extent to which annotators agree on which portions of text contain different types of hallucinations, we computed the pairwise Jaccard index (i.e., intersection over union), a common evaluation metric for annotator highlights (Herrewijnen et al., 2024). We calculated this metric for divergent information in input and output, and factually wrong information, considering all annotators who annotated a given item. For each item, we averaged the pairwise scores, and then computed the average across all items. As a reference, we computed the same metric among two reference annotators (see §3.3.1).[4] See Table 13 for the figures. The Jaccard index indicates that annotators tended to agree more on the spans of text identified as containing factually wrong information (FWI) compared to those identified as containing divergent information in the output. This strengthens the fact that it is easier to agree on more serious hallucinations (e.g., a falsely added room amenity not mentioned in the input) than on harmless additions (e.g., lofty content).

We manually analyzed the experimental item with the lowest overall Jaccard index (i.e., the item where annotators showed the greatest disagreement). Figure 8 presents a heatmap illustrating how many annotators identified each token as containing divergent information in the input, divergent information in the output, or factually wrong information. Several interesting patterns emerge from this analysis. First, although 10 annotators worked on this item, only 9 annotated divergent or factually wrong information.

Within these annotations, we observe instances of plausible disagreement, which appear to stem from genuine ambiguity or underspecification in the input (see §6). For example, the input mentions *Accessible hotel*, while the output refers to *accessible rooms*. Some annotators treated these as divergent, raising the question of whether accessibility at the hotel level entails accessibility of individual rooms. Similarly, while the input mentions *Accessible parking*, the output refers to *convenient parking options*. Here again, some annotators perceived a divergence, prompting interpretation-based disagreement (i.e., does *convenient* imply *accessible*?). Another case involves the part where the input mentions *Room amenities:* [...] *Sitting area*, whereas the output refers to *sitting area in each room*. This added specificity led some annotators to label *in each room* as just divergent information in output (a harmless addition), others also as factually wrong information, while others did not label the phrase at all. These contrasting views may be due to the nature of the input, whose underspecified phrasing does not clarify whether it necessarily applies to all rooms.

[4]One of the three authors did not highlight divergent information.

Annotator Group	Div. Info (Input)	Div. info (Output)	FWI
Reference	0.82	0.58	0.74
Crowd	0.65	0.42	0.63

Table 13: Annotator agreement (Jaccard index) on divergent and factually wrong information.

In contrast, there are also clear points of agreement. All annotators identified adjectives such as *charming*, *comfortable*, and *relaxing* as harmless additions (divergent information in output), and none considered them factually wrong.

However, other stylistic additions were more contentious. Phrases like *for convenience*, *for entertainment*, or *Families with children will appreciate* were highlighted by only a subset of annotators as divergent information in output, while others did not consider them divergent at all.

Finally, we can also observe a few clear annotation mistakes. For instance, one annotator mistakenly flagged *golf course* as factually wrong, likely due to an oversight, despite it being clearly mentioned in the input.

C.2 Logic Domain

Analysis of Ambiguity Types Ambiguity did not play a significant role, as originally expected. Most LLM-generated outputs were near-literal renditions of the input (see also §6). Annotators interpreted the outputs favorably, selecting interpretations that aligned with the corresponding input. Nonetheless, 36 items were flagged as containing some ambiguity type by at least one annotator. Table 14 shows the distribution of ambiguity types over all items.

The output considered most ambiguous (flagged by all annotators as containing multiple types of ambiguities) is the following: *For every x, x is a dodecahedron and there exists a y such that y is to the right of x and y is a cube, and x is large, or x*

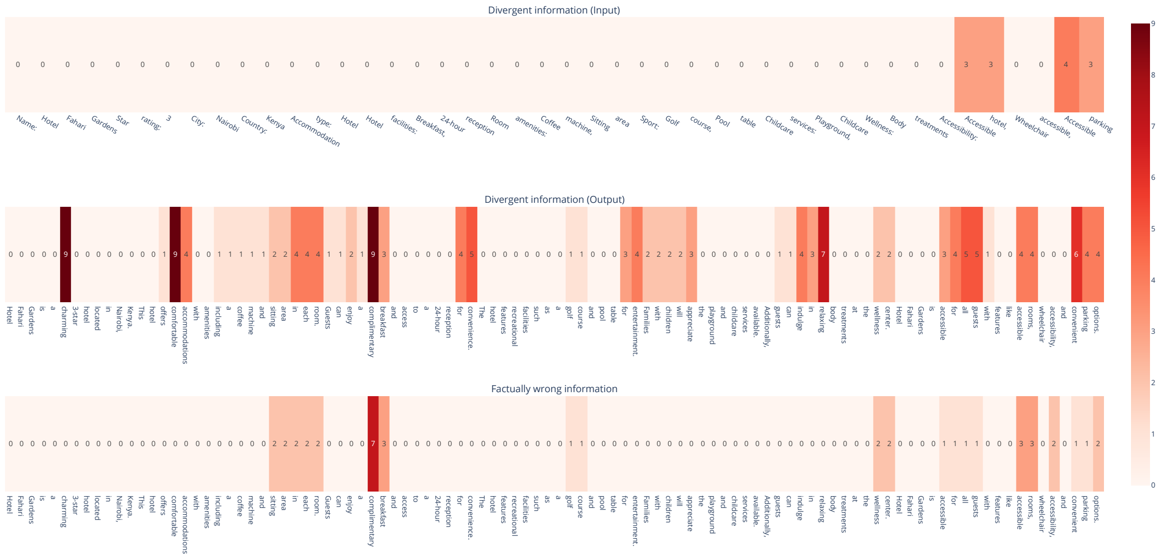

Figure 8: Heatmap of an input-output pair from the hotel domain experiment, illustrating the number of annotators who identified each token as containing divergent information in the input, divergent information in the output, or factually wrong information.

Ambiguity Type	Distribution
Connective precedence	23
Quantifier scope	23
Negation scope	11
Other	10

Table 14: Distribution of ambiguity types.

is not a dodecahedron or there does not exist a y such that y is to the right of x and y is a cube.

D Details on Annotation with LLMs

We structured the annotation experiments with LLMs in both the hotel and logic domains to closely mirror the human annotation setup. Technically, the format aligns with a "guided" chain-of-thought approach (Wei et al., 2022): at each step requiring an LLM response, we retrieved the answer, appended it to the ongoing conversation, and proceeded to the next step of the experiment. Following the setup in §4, LLMs underwent the interactive training with adaptive feedback, completed the comprehension check in the hotel domain described in §4 and Appendix B, and annotated all input-output pairs (25 in the hotel domain and 75 in the logic domain) step by step, following the questions in the decision tree in Figure 1.

We did not modify the textual instructions from the human experiment, apart from adding some constraints such as "DO NOT PROVIDE ANY FURTHER INFORMATION NOR EXPLANATION." after each question, to prevent excessive verbosity.

We used the wrappers provided by LangChain to perform calls to the dedicated APIs for all models (with default parameters),[5] except Llama 3.3 (for which we used the together.ai inference API).

As a final remark, since LLMs could not properly highlight divergent information, we excluded this aspect from the analysis.

[5]For Grok-2, we used the version `grok-2-1212`.

E Hotel Domain Annotation Experiment Guidelines

Guidelines Summary

We will show you pairs of information sources, which we call Input (I) and Output (O).

Input (I)	Output (O)
Name: Hotel Torre Azul City: El Arenal Star rating: 4	Hotel Torre Azul is a 4-star hotel located in El Arenal.

O is automatically generated from I by an artificial intelligence system and can contain *problematic content*. I is our point of reference: **We ALWAYS assume that I includes all the relevant facts and they are CORRECT**.

When performing the annotation disregard any grammatical mistakes or typos.

a. Divergent Information

We define as ***divergent*** any information that is present in one of the two information sources (e.g., in I) but not in the other (e.g., in O).

Divergent Information in Input Divergent Information in Output

I	O
Name: Hotel Torre Azul City: El Arenal Star rating: 4 Hotel facilities: Lockers, Wi-Fi	Hotel Torre Azul is an incredibly charming 3-star hotel located in El Arenal, Spain. The hotel facilities include a lobby and free Wi-Fi access, making the hotel ideal for working remotely.

b. Inclusion

If O does **NOT** contain any Divergent Information, then I ***includes*** O.
If I does **NOT** contain any Divergent Information, then O ***includes*** I.

I	O
Name: Hotel Torre Azul City: El Arenal Star rating: 4	Hotel Torre Azul, a 4-star hotel, is located in El Arenal, Spain.

O includes I, because I does not contain any *Divergent Information*.
I does not include O, because O contains some *Divergent Information* (i.e., Spain).

I	O
Name: Hotel Torre Azul City: El Arenal Star rating: 4 Hotel facilities: Lockers	Hotel Torre Azul, a 4-star hotel, is located in El Arenal.

O does not include I, because I contains some *Divergent Information* (i.e., Hotel facilities: Lockers).
I includes O, because O does not contain any *Divergent Information*.

I	O
Name: Hotel Torre Azul City: El Arenal Room amenities: Balcony	Hotel Torre Azul is a charming hotel located in El Arenal.

O does not include I, because I contains some *Divergent Information* (i.e., Room amenities: Balcony).
I does not include O, because O contains some *Divergent Information* (i.e., charming).

c. Factually Wrong Information

When I **DOES NOT** include O, O contains information that is not in I. This Divergent Information in Output may or may not be ***factually wrong***.

> We define as ***factually wrong*** any information in O that conveys facts that could well turn out to be wrong, given the information in I.

Factually wrong information **DOES NOT** comprise: (i) subjective opinions, (ii) vague or ambiguous statements, (iii) information that is unverifiable, (iv) information that can be inferred from I to hold in most (but not all) cases.

Divergent Information in Input Divergent Information in Output

I	O
Name: Hotel Torre Azul City: El Arenal Country: Spain Star rating: 4 Hotel facilities: Housekeeping Room amenities: Balcony	Hotel Torre Azul is an incredibly charming 3-star hotel located in the center of El Arenal, Mexico. The hotel facilities include free housekeeping and lockers. All rooms have a balcony.

3, Mexico, lockers **ARE** pieces of factually wrong information.
incredibly charming (i), the center of (ii), All (iii), free (iv) **ARE NOT** pieces of factually wrong information.

d. Contradiction

> I and O ***contradict each other***, if I and O contain information that cannot be true simultaneously.

Divergent Information in Input Divergent Information in Output

I	O
Name: Hotel Torre Azul City: El Arenal Star rating: 4	Hotel Torre Azul, a 3-star hotel, is located in El Arenal.

I and O contradict each other, since I states that the hotel has 4 stars, while O states that the hotel has 3 stars.

e. Self-Contradiction

> O ***contradicts itself***, if O contains pieces of information that cannot be true simultaneously.

Divergent Information in Input Divergent Information in Output

I	O
Name: Hotel Torre Azul City: El Arenal Star rating: 4 Hotel facilities: Lockers	Hotel Torre Azul, a 4-star hotel, is located in El Arenal. The hotel facilities include lockers. The hotel does not provide lockers.

O contradicts itself, since O states that the hotel provides lockers and, at the same time, that it does not provide them.

F Logic Domain Annotation Experiment Guidelines

Guidelines Summary

We will show you pairs of information sources, which we call Input (I) and Output (O).

Input (I)	Output (O)
$\forall x(Cube(x) \rightarrow Large(x))$	All cubes are large.

O is automatically generated from I by an artificial intelligence system and can contain *problematic content*. I is our point of reference: **We ALWAYS assume that I is CORRECT**.

When performing the annotation disregard any grammatical mistakes or typos.

a. Logical Consequence

> I ***entails*** O, if by reading I, the annotator infers that O is true.
> O ***entails*** I, if by reading O, the annotator infers that I is true.

Example 1:

I	O
$\forall x(Cube(x) \rightarrow Large(x))$	All red cubes are large.

I entails O, because if all cubes are large, then all red cubes are large too.
O does not entail I, because even if all red cubes are large, it may well be the case that there are other cubes that are not large.

Example 2:

I	O
$\forall x((Cube(x) \land Green(x)) \rightarrow Large(x))$	All cubes are large.

I does not entail O, because even if all green cubes are large, it may well be the case that there are other cubes that are not large.
O entails I, because if all cubes are large, then all green cubes are large too.

Example 3:

I	O
$\forall x((Cube(x) \land Green(x)) \rightarrow Large(x))$	All red cubes are large.

I does not entail O, because even if all green cubes are large, that does not say anything about red cubes.
O does not entail I, because even if all red cubes are large, that does not say anything about green cubes.

b. Contradiction

I and O ***contradict each other***, if I and O contain information that cannot be true simultaneously.

I	O
$\forall x(Cube(x) \rightarrow Large(x))$	All cubes are not large.

I and O contradict each other, since I states that all cubes are large, while O states that all cubes are not large.

c. Self-Contradiction

O ***contradicts itself***, if O contains pieces of information that cannot be true simultaneously.

I	O
$\forall x(Cube(x) \rightarrow Large(x))$	All large cubes are not large.

O contradicts itself, since O states that, at the same time, all cubes are large and not large.

d. Ambiguity

O is ***ambiguous***, if by reading O, the annotator perceives distinct interpretations for O.

I	O
$\forall x(Cube(x) \rightarrow \exists y(Tet(y) \land Behind(x, y)))$	Every cube is behind a tetrahedron.

O is ambiguous, because an annotator could perceive (at least) two distinct interpretations for O: (i) each cube is behind a (possibly different) tetrahedron, or (ii) there is some tetrahedron that is in front of all cubes.

Appendix: Predicates

The following are the predicates you may encounter and how you are supposed to read them:

Predicate	Description
$SameSize(x, y)$	x and y are the same size.
$Smaller(x, y)$	x is smaller than y.
$SameCol(x, y)$	x and y are in the same column.
$Larger(x, y)$	x is larger than y.
$BackOf(x, y)$	x is behind y.
$Medium(x)$	x is medium.
$Large(x)$	x is large.
$FrontOf(x, y)$	x is in front of y.
$Adjoins(x, y)$	x adjoins y.
$Small(x)$	x is small.
$Between(x, y, z)$	x is between y and z.
$LeftOf(x, y)$	x is to the left of y.
$Cube(x)$	x is a cube.
$Dodec(x)$	x is a dodecahedron.
$RightOf(x, y)$	x is to the right of y.
$SameRow(x, y)$	x and y are in the same row.
$SameShape(x, y)$	x and y are the same shape.
$Tet(x)$	x is a tetrahedron.
$Behind(x, y)$	x is behind y.
$Green(x)$	x is green.