A Logic-Based Approach to Hallucinations in Data-to-Text NLG: Experiments with Human and LLM Annotators

Eduardo Calò[α] **Saad Mahamood**[ω] **Albert Gatt**[α] **Kees van Deemter**[α]
[α]Utrecht University [ω]Shopware
{e.calo, a.gatt, c.j.vandeemter}@uu.nl saad@saad.me.uk

Abstract

Hallucinations are a persistent challenge in natural language generation, including data-to-text. van Deemter (2024) introduced a framework based on the relation of logical consequence ("follows from"), which divides all data-to-text hallucinations into seven disjoint categories. 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, although there are significant domain-dependent variations, as well as discrepancies between human and model judgments. We also uncover several issues that should inform future work on hallucination.

https://github.com/Eduardo-Calo/hallucination-framework

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 generate natural language text from structured input data, e.g., from sensors (Gatt et al., 2009), knowledge bases (Colin et al., 2016), or tables (Parikh et al., 2020) (see §2).

van Deemter (2024) offered a critique of these analyses and proposed a categorization of all the logical relationships that can be obtained between the input and output of a data-to-text system (§3.1 for details). The idea is to compare the factual contents of the input and output with each other, asking whether the two are "well-matched" (i.e., whether the input follows from the output and the output follows from the input), and if not, then why not. This analysis claims to cover all types of hallucination, and all types of input data. It promises to enhance our understanding of factual inaccuracies committed by large language models (LLMs) and to offer a starting point for addressing questions of error severity (van Miltenburg et al., 2020) and hallucination mitigation (Ji et al., 2024), including in real-world scenarios (e.g., Hegselmann et al., 2024; Schmidtová et al., 2025).

This paper presents the first implementation of this logic-based framework for hallucination analysis in real-world data-to-text domains. Starting from the notion of logical consequence (i.e., "follows logically from"), we show how the framework can be operationalized as a multi-step reasoning procedure (§3.1), detailing the adaptations required to make the framework usable in practice. We develop annotation guidelines applicable across different domains, suitable for both human annotators and LLMs, because (i) there is a growing trend of using LLMs as judges in evaluation tasks (Zheng et al., 2023; Tan et al., 2024; Bavaresco et al., 2025, *i.a.*); (ii) given that large-scale manual annotation is expensive and time-consuming, low-cost LLM-based alternatives might be used instead of (or in addition to) human annotation (Calderon et al., 2025); and (iii) LLMs have been shown to perform well in logic-based inference tasks akin to ours (e.g., the FOLIO benchmark for first-order logic reasoning; Han et al. 2024).

It can sometimes be difficult to say whether a given output follows from a given input, and vice versa, with different domains posing different challenges (van Deemter, 2024). We thus consider two very different data-to-text domains: (i) tourist accommodation descriptions generated from database

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 28–62
July 3, 2026 ©2026 Association for Computational Linguistics

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
$\exists x\, \neg($ $\text{Cube}(x) \rightarrow$ $\forall y\, (\text{Tet}(y) \rightarrow \text{Smaller}(x, y))$ $)$
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" (Figure 1).

entries (henceforth, **hotel domain**), and (ii) descriptions of simple geometrical scenes generated from logical formulae (henceforth, **logic domain**). The hotel domain is characterized by simple inputs (i.e., conjunctions of atomic facts), and lengthy outputs, which can be stylistically elaborate (Table 1). The logic domain uses short but potentially complex inputs, yielding purely factual outputs (Table 2). Both domains have real-world relevance: hotel NLG has been deployed commercially (e.g., Mahamood and Zembrzuski, 2019), while logic NLG has been applied in classroom settings (e.g., Mpagouli and Hatzilygeroudis, 2009). We address two **research questions**: First, to what extent do human annotators agree among themselves and with reference annotations? Second, can the annotation process be automated by using LLMs?

To investigate these questions, we adapted the framework to the hotel and logic domains and conducted annotation experiments with both human annotators and several state-of-the-art LLMs under four prompting strategies. Through this study, we: (i) provide the first cross-domain implementation of van Deemter's framework, breaking it down into manageable reasoning steps; (ii) conduct annotation experiments with both humans and LLMs, analyzing annotator agreement and alignment with reference annotations; (iii) present ample comparisons between human and model results, showing that while annotation was challenging for people, some LLMs performed the task remarkably well.

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., 2023b).

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 and tasks (Mickus et al., 2024; Vazquez 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 coarse-grained analysis highlighting hallucination (i.e., the output does not logically follow from the input) and omission (i.e., the input does not logically follow from the output). A more informal 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, and context errors, among others.

3 Methodology

We implemented van Deemter's framework by applying a multi-step reasoning procedure and adapted it to the hotel and logic domains. We created data for annotation by retrieving inputs from two data sources and generating outputs using LLMs in a data-to-text setting. We acquired reference annotations for these input-output pairs.

3.1 Framework Implementation

Original Framework van Deemter (2024) argued that existing error classifications require clarification, refinement, and extension. Suppose the input says the Dutch soccer team *lost all its matches*. Outputs can get this wrong in different ways, for example, by saying "Holland lost *some/most/none* of its matches"; yet, presumably, Thomson and Reiter (2020)'s scheme would classify each of these three as "incorrect word" errors, failing to distinguish between errors that end up misinforming users in very different ways. Furthermore, all these classifications fail to make some subtler distinctions. Suppose, for instance, the input and output are such that neither follows from the other. For Dušek and Kasner (2020)'s scheme, this is the end of the story; hence, it fails to distinguish between cases in which the input and output are *logically independent* of each other and cases in which they *contradict* each other. Suppose the input says the temperature is between 10 and 20 degrees Celsius; then the output *The temperature is above 15 degrees* exemplifies the former (because the input and output can both be true or false, independently of each other), whereas the output *The temperature is above 30 degrees* would exemplify the latter (because input and output cannot both be true).

In view of these and other issues, van Deemter (2024) proposed a new categorization. Like Dušek and Kasner (2020)'s, it is based on the logical consequence relation. Unlike that, however, it systematically examines all the ways in which the input (I) and output (O) of data-to-text NLG systems can be logically related to each other. As a result, the framework proposed seven categories (Table 3).

The framework made some assumptions. 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 includes every step following Content Selection, see e.g., Reiter 2007, 2025). Second, it only asks whether the output of the NLG system matches the input, without considering the truth of the output in the real world.

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 disregarded the category "O tautologous" (case 1b in Table 3), because such outputs are exceedingly rare in both our domains.

Hallucination Severity We found that, in the hotel domain, some errors were more serious than others. We 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). Not all divergent information is necessarily factually wrong. Hotel-related outputs can contain information that is divergent, but where the divergence is unlikely to lead to any complaints from customers. On the other hand, a piece of divergent information is *factually wrong* if, despite everything the input says, the information *could* turn out to be manifestly incorrect.

For instance, if the output asserts, without any basis in the input, that a hotel has a swimming pool, then this is factually wrong. Divergent information that is not factually wrong can cover different 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 commercial phrasing 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., whether the output contains divergent information). 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, because when an output permits multiple interpretations, then different hallucination categories could be assigned to it depending on the interpretation selected.

Logically rich outputs can contain various types of ambiguity, potentially leading different annotators to perceive distinct interpretations of the same text. These ambiguities include connective prece-

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 the USA.*

Table 3: van Deemter (2024)'s classification, with examples from the hotel domain. Input: 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.

dence (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 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.

3.2 Data Creation

3.2.1 Hotel Domain

Input: From trivago 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., 2024). 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 7).

Reference Annotations: Three of the paper's authors independently annotated all 25 input-output pairs following the setup of §4. Each annotated the 25 pairs individually. Following a discussion between the three, a consensus annotation was reached for 17 out of 25 pairs. The authors agreed to disagree on 8 pairs, in which they acknowledged that different answers were possible (2 votes vs. 1), allowing these cases to have multiple labels (IAA: Krippendorff's α = 0.73).

3.2.2 Logic Domain

Input: We used the Grade Grinder Corpus (GGC; Barker-Plummer et al., 2011), a corpus of first-order logic formalizations made by students answering exercises in Barwise et al. (2000). We restricted ourselves to the correct formalizations 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 8).

Reference Annotations: Two of the paper's authors (proficient in logic) 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 (IAA: Krippendorff's α = 0.85). Four pairs were discarded, as it was impossible to determine the truth value of the outputs because they were highly ungrammatical or incomplete (see §6).

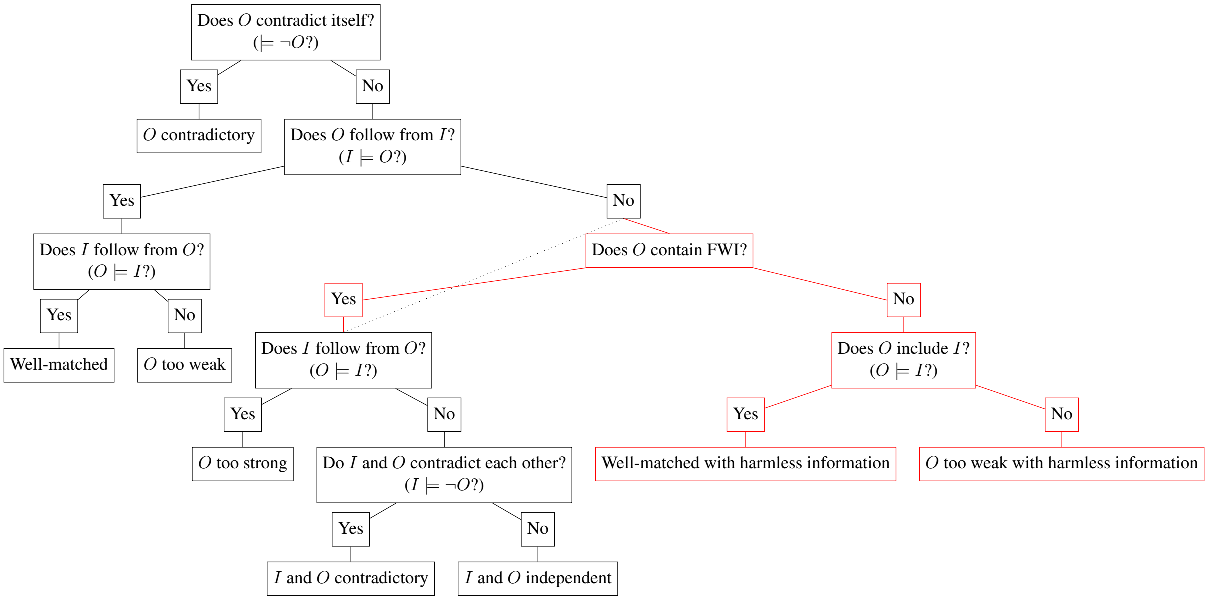

Figure 1: Framework adapted for our data-to-text domains. The black portion of the decision tree was used for the logic domain experiment, while the red portion was added for the hotel domain experiment. The tree illustrates the questions posed to annotators and the order in which they were presented. FWI: *factually wrong information*. In brackets, the logic-based representation, not shown to the annotators. In the hotel domain, we used the term "include" instead of "follow from". See Appendices B, F, and G for more details.

3.3 Evaluation Metrics

Accuracy per Annotator (APA): To understand how often annotators aligned 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),[1] and n is the number of pairs annotated by each annotator.

Inter-Annotator Agreement (IAA): While APA measures the alignment of annotators with reference labels, we also used Krippendorff's alpha (α; Krippendorff, 1980) to measure inter-annotator agreement. We adopted α since it is robust in handling skewed label distributions and missing annotations (Artstein and Poesio, 2008; James, 2026).

F1-Score per Category (FPC): We computed 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 Setup

Human 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 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 §3.1). 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 trivago; 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 logic (median age = 35; male = 75.0%, female = 25.0%), professional contacts of the authors, unfamiliar with our research questions.

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

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 saw one LLM output per original input.

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.

LLMs We chose the following six LLMs, which represent a variety of open-weight and proprietary models among the top-performing ones[2] on the benchmark proposed in White et al. (2025) under the reasoning subcategory: DeepSeek-R1-0528 (DeepSeek-AI et al., 2025), Gemini 2.5 Pro (Comanici et al., 2025), GPT-5, Grok-4, Claude Opus 4.1, and Claude Sonnet 4. We used OpenRouter to perform API calls for all models (with default parameters, and temperature set to 0).

We experimented with four prompting strategies: zero-shot (0-shot), few-shot (F-shot), chain-of-thought (CoT; Wei et al., 2022), and tree-of-thought (ToT; Yao et al., 2023). We chose these strategies (Calò et al., 2026), because they cover both widely used basic approaches (0-shot and F-shot) and more elaborate reasoning-based ones (CoT and ToT), well-suited for the task at hand (Schulhoff et al., 2025). In 0-shot, LLMs were given only the leaf categories from Figure 1 and asked to assign a category to each input-output pair. In F-shot, they were provided with the categories plus one example per category. In CoT, models received additional step-by-step reasoning for each example. The ToT strategy closely resembles the human annotation setup: LLMs responded step by step following the decision tree in Figure 1. Using these four setups, LLMs annotated all input-output pairs (25 in the hotel domain and 75 in the logic domain). See Appendix D for the prompts.

5 Results

For both human and LLM annotators, we computed APA, IAA, and FPC across domains (Table 6 and Table 4). In the hotel domain, we computed APA and IAA (i) on the leaf categories (**CAT**; eight possible outcomes), (ii) on **I** $\models$ **O** (two possible outcomes), (iii) on **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). In the logic domain, we computed APA and IAA (i) on **CAT** (six possible outcomes), (ii) on **I** $\models$ **O**, (iii) on **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 CAT as indicative of how well annotators engage with the framework as a whole. $I \models O$ and $O \models I$ 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. For LLM annotators, we also computed APA for CAT across prompting strategies (Table 5).

APA scores are generally good across prompting strategies for most models (Table 5). In the hotel domain, GPT-5 achieves the best performance overall under CoT and ToT, while in the logic domain, DeepSeek-R1 consistently leads across most strategies. Strong performance was also obtained by Gemini-2.5 and Grok in both domains. Opus-4.1 and Sonnet-4 yield the lowest scores overall. Prompting effects vary by domain: in the hotel domain, performance sharply improves from F-shot to CoT/ToT (e.g., GPT-5, Opus-4.1), whereas in the logic domain, 0-shot/F-shot often perform as well as or better than CoT/ToT. For three models in each domain, ToT results are lower than CoT.

Comparing ToT and human annotation results (Table 6),[3] APA ranges from good to very good across all dimensions for both humans and LLMs, with most scores above 0.70. APA for CAT is consistently lower than for the other (binary) dimensions, which is expected given that the hierarchical structure of the decision tree requires intermediate questions to be answered correctly in order to reach the final category. Among models, GPT-5 is the best one in the hotel domain and DeepSeek-R1 in the logic domain. IAA among human annotators ranges from low to moderate, whereas model IAA is consistently higher across all dimensions than human IAA (e.g., in the logic domain, $I \models O$ shows

[2] At the time we ran the experiments (September 2025).

[3] We compare the human setup and the ToT strategy for models, because ToT is the strategy that resembles the human setup (§4), and so the only one that gives information for the higher levels of the tree.

	Human	DeepSeek-R1				Gemini-2.5				GPT-5				Grok-4				Opus-4.1				Sonnet-4				S
		0-shot	F-shot	CoT	ToT	0-shot	F-shot	CoT	ToT	0-shot	F-shot	CoT	ToT	0-shot	F-shot	CoT	ToT	0-shot	F-shot	CoT	ToT	0-shot	F-shot	CoT	ToT	
Hotel																										
Well-matched	0.67	0.00	0.67	1.00	0.50	1.00	1.00	1.00	1.00	0.40	0.40	0.67	0.80	0.67	0.67	0.67	0.67	0.24	0.29	1.00	0.57	0.25	0.29	1.00	0.57	2
Well-matched (harmless)	0.89	0.70	0.78	0.62	0.67	0.74	0.78	0.71	0.59	0.61	0.64	0.86	0.67	0.76	0.78	0.67	0.75	0.14	0.46	0.50	0.46	0.33	0.40	0.59	0.71	8
O too weak	0.80	1.00	0.80	0.67	0.67	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.80	1.00	1.00	0.67	0.50	0.40	0.57	1.00	0.57	0.00	0.67	0.40	0.44	2
O too weak (harmless)	0.77	0.43	0.71	0.40	0.40	0.71	0.62	0.50	0.50	0.22	0.00	0.55	0.80	0.59	0.62	0.55	0.71	0.00	0.00	0.55	0.33	0.00	0.00	0.00	0.40	8
O too strong	0.33	0.00	0.44	0.57	0.40	0.25	0.22	0.57	0.25	0.40	0.33	0.67	0.86	0.00	0.50	0.57	0.67	0.00	0.29	0.60	0.40	0.00	0.33	0.57	0.33	4
I and O independent	0.80	0.00	0.00	0.62	0.62	0.00	0.22	0.62	0.36	0.00	0.00	0.62	0.77	0.00	0.36	0.71	0.77	0.00	0.00	0.55	0.62	0.00	0.00	0.40	0.55	8
Macro Average	0.71	0.36	0.57	**0.65**	0.54	0.62	0.64	**0.73**	0.62	0.27	0.23	0.73	**0.78**	0.50	0.66	0.64	**0.68**	0.13	0.27	**0.70**	0.49	0.10	0.28	0.49	**0.50**	
Logic																										
Well-matched	0.93	0.92	0.92	0.89	0.92	0.91	0.90	0.95	0.83	0.89	0.91	0.92	0.88	0.91	0.91	0.93	0.88	0.89	0.90	0.68	0.90	0.90	0.90	0.92	0.88	61
O too weak	0.33	0.62	0.62	0.62	0.53	0.73	0.62	0.60	0.50	0.43	0.62	0.55	0.29	0.53	0.62	0.67	0.47	0.33	0.50	0.33	0.44	0.44	0.62	0.22	0.20	4
O too strong	0.25	0.57	0.57	0.57	0.50	0.57	0.57	0.75	0.50	0.50	0.57	0.57	0.50	0.57	0.57	0.89	0.67	0.00	0.00	0.00	0.29	0.25	0.57	0.29	0.29	5
O contradictory	0.00	-	-	-	-	-	-	-	0.00	-	-	-	-	-	-	0.00	0.00	-	-	0.00	0.00	-	-	-	0.00	0
I and O independent	0.55	0.89	0.86	0.86	0.50	0.86	0.75	0.73	0.00	0.75	0.75	0.73	0.62	1.00	0.80	0.75	0.50	0.40	0.00	0.57	0.43	0.00	0.75	0.22	0.25	4
I and O contradictory	0.00	-	0.00	0.00	-	0.00	0.00	-	-	-	0.00	-	-	-	-	-	-	0.00	0.00	0.00	-	0.00	0.00	0.00	0.00	0
Macro Average	0.34	**0.75**	0.59	0.59	0.61	0.61	0.57	**0.76**	0.37	0.64	0.57	**0.69**	0.57	**0.75**	0.72	0.65	0.50	0.32	0.28	0.26	**0.41**	0.32	**0.57**	0.33	0.27	

Table 4: FPC for LLMs per prompting strategy, and human performance in the leftmost column, in the hotel and logic domain. Reference label support (S) in the rightmost column is shared by all. - indicates that S and annotator count were both 0 for that category. Boldfaced are the best macro averages per LLM per strategy per domain.

	Hotel				Logic			
Model	**0-shot**	**F-shot**	**CoT**	**ToT**	**0-shot**	**F-shot**	**CoT**	**ToT**
DeepSeek-R1	0.52	0.68	0.68	0.64	**0.87**	**0.85**	0.84	**0.83**
Gemini-2.5	**0.68**	0.68	0.76	0.60	**0.87**	0.84	**0.89**	0.67
GPT-5	0.44	0.40	**0.80**	**0.88**	0.81	**0.85**	0.85	0.77
Grok-4	0.64	**0.72**	0.72	0.80	0.85	0.83	0.88	0.79
Opus-4.1	0.16	0.32	0.68	0.56	0.75	0.77	0.49	0.79
Sonnet-4	0.20	0.32	0.52	0.60	0.79	0.71	0.77	0.71

Table 5: APA for CAT for LLMs across prompting strategies, in the hotel and logic domains. Boldfaced are the higher values per strategy per domain.

	Hotel				Logic			
	CAT	**I $\models$ O**	**O $\models$ I**	**FWI**	**CAT**	**I $\models$ O**	**O $\models$ I**	**AMB**
Human								
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
Model								
DeepSeek-R1	0.64	0.76	0.88	0.84	**0.83**	0.89	**0.88**	**0.75**
Gemini-2.5	0.60	**1.00**	0.76	0.84	0.67	0.80	0.76	0.64
GPT-5	**0.88**	0.92	**1.00**	**0.96**	0.77	0.87	0.84	0.67
Grok-4	0.80	0.96	0.92	0.92	0.79	**0.91**	0.83	0.68
Opus-4.1	0.56	0.76	0.88	0.88	0.79	0.87	0.87	0.71
Sonnet-4	0.60	0.68	0.92	0.88	0.71	0.85	0.80	0.61
IAA	0.45	0.43	0.69	0.61	0.51	0.53	0.64	0.39

Table 6: APA and IAA for human and LLM annotators for all dimensions, in the hotel and logic domain. APA scores for human annotators are the means of single annotator scores; standard deviations are reported in brackets. Boldfaced are the higher values per LLM per dimension per domain.

nearly a 0.40 increase from humans to models). Interestingly, IAA for $O \models I$ is systematically higher than for $I \models O$ in both domains.

When analyzing results per category (Table 4), macro averages show that human annotators performed better in the hotel domain than in the logic domain. In the hotel domain, several models (e.g., Gemini-2.5 in CoT, GPT-5 in ToT) performed on par or even better than humans. Others performed poorly, failing to classify even a single item correctly in several categories (e.g., Sonnet-4 in 0-shot). In the logic domain, models consistently outperform humans, sometimes by large margins (e.g., DeepSeek-R1 and Grok-4 in 0-shot, Gemini-2.5 in CoT). Note that the reference label distribution (column S) in the logic domain is 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, yet far from fluent. Finally, note that the "O contradictory" and "I and O contradictory" categories are empty (–) for many models in the logic domain, which aligns with the reference distribution (column S). We follow up with additional domain-specific analyses in Appendix C.

6 Discussion

Applicability of the framework. Returning to the research questions of §1, the high APA scores for $I \models O$ and FWI for both humans and models (Table 6) suggest that our main adaptation to the framework of van Deemter (2024), in which divergent information was separated into yes/no *factually wrong* (§3.1), was effective, demonstrating the framework's applicability in real-world scenarios. These results also suggest that hallucination detection can be approached similarly to other reasoning benchmarks, such as those in Tian et al. (2021) and Han et al. (2024), which likewise operationalize reasoning tasks in terms of logical consequence.

Models > humans. Human annotation is feasible, yet challenging. In the hotel domain, this was clear from the large proportion of would-be annotators who failed our comprehension check (§4). Annotators in the logic domain were not filtered in this way, because our recruitment process guaranteed a high level of expertise. This reflects a broader issue: high-quality annotation demands strict criteria that many annotators may not meet (Zhang et al., 2023a). Conversely, model annotation is promising, with some models (including open-weight ones) consistently outperforming humans in alignment with reference labels, especially in the logic domain (Table 6, Table 4). This indicates strong potential for automation, which can help scale up expensive and time-consuming large-scale annotation.

APA and IAA scores. Across the board, APA scores are higher than IAA scores. This happens because they are fundamentally different metrics. IAA measures the extent to which annotators agree with each other, without considering reference annotations; APA measures the extent to which annotators align with any of the reference annotations.[4]

Model APA in the higher-level binary dimensions $I \models O$ and $O \models I$ reached high and sometimes perfect scores (Table 6). This mirrors findings from Mickus et al. (2024), who also operationalize hallucination detection in terms of logical consequence within a binary classification setting, noting accuracies close to 0.85 for the best models.

Human IAA was low to moderate (Table 6), echoing results from other hallucination annotation studies in real-world domains (e.g., consultation note generation; Moramarco et al., 2022; Hegselmann et al., 2024), where IAA was moderate at best. This is consistent with evidence of genuine disagreement on other difficult and subjective annotation tasks, reflecting the point emphasized by Plank (2022) that human label variation is ubiquitous and that high IAA is often achievable only under artificial conditions.

Prompting strategies. There is no definitive formula for selecting the most effective prompting strategy, as model performance varies across strategies (Table 5). Importantly, the need for complex prompting appears to be domain-dependent. In the logic domain, basic strategies (0-shot and F-shot) often outperform more elaborate ones, while in the hotel domain, the presence of commercial content (e.g., phrases such as *incredibly charming*) makes it harder for models to succeed with only minimal examples. In such cases, reasoning-based strategies (CoT and ToT) provide models with the necessary scaffolding to effectively break down the task.

Reference annotations. Acquiring reference annotations proved to be laborious and time-consuming. The authors devoted substantial individual effort and held multiple meetings to discuss ambiguous or complex cases in order to reach consensus (echoing the agreement process mentioned in Hegselmann et al. (2024), where expert doctors discussed to reach consensus). Even if our reference IAA was high (0.73 in the hotel domain and 0.85 in the logic domain), in some instances, consensus was not reached, resulting in items being assigned multiple hallucination categories (§3.2). For example, consider the case (Table 7) where the input specifies *Childcare services: Organised activities for kids, Playground* and the output states *organised activities and a playground for kids*. One author argued that the output fails to clearly indicate that the organized activities are for children. In the same pair, the input includes *Accessibility: [...] Accessible hotel*, while the output reads *This hotel [...] with its accessible accommodations*. Here, the authors disagreed on whether *accessible hotel* necessarily entails *accessible accommodations*.

Skewed label distribution. Our use of *bona fide* corpora (§3.2) ensures that the input-output pairs studied retain real-world validity. However, this also introduced skewness in the logic domain data, where most outputs fell into the "Well-matched" category (Table 4). These cases were often relatively easy to judge, especially when outputs were highly formulaic (Calò et al., 2025a), e.g., *For all z and for all y, if z is behind y, then z is larger than y*. Importantly, this skewness is not an artifact of our framework, but rather a natural consequence of using real-world data. Similar imbalances have been observed elsewhere. For instance, Thomson and Reiter (2020) found that in another data-to-text domain (i.e., basketball summary generation from box scores data), numerical errors were heavily represented, while other error types (e.g., context errors) were rare.

Ambiguity. Ambiguous outputs were rare in the hotel domain but more frequent in the logic domain.[5] 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

[4]We verified that APA scores in Table 6 would not change much if we considered only the majority reference label (i.e., a single correct label per input-output pair; §3.2 and §3.3). See Table 14 in Appendix E for the detailed results.

[5]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*.

to interpret the output "charitably", choosing a well-matched interpretation of the output whenever available (see Appendix C.2).

Finally, our study highlights two more issues with wider significance within the field of NLG.

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 some 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, the LLMs we used for generation sometimes did little more than "read out" the input formula. Sometimes, this led to English sentences that are so ungrammatical that it is impossible to say whether they follow from a given input (§3.2). 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. We do not know how replicable this phenomenon will prove to be for better, or differently prompted, models, but it appears to justify a new category "Output not well-formed", to be added to van Deemter (2024)'s and other hallucination frameworks.

7 Conclusion

Based on a domain-dependent modification of the classification scheme of van Deemter (2024), we showed how hallucination annotation in data-to-text NLG can be decomposed into manageable annotation steps. Human performance highlights the task's inherent complexity. Model performance points to the potential for automation and scalable annotation. The results call for caution in the design of hallucination annotation studies and emphasize the importance of calibrated annotation guidelines, robust theoretical foundations and practical considerations, e.g., on input data quality.

Limitations

We only focused on two data-to-text NLG domains. Obstacles to hallucination annotation other than those 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 vagueness to give rise to similar problems as ambiguity, and that these problems can be addressed along similar lines.

Both van Deemter (2024)'s original framework and our adaptation assume that all and only input content should appear in the output (§3.1). While appropriate for many data-to-text tasks, this does not hold for other cases like summarization or image captioning, in which expressing the entire input in the output is neither feasible nor desirable. Future work could focus on relaxing the $O \models I$ constraint in cases where only the most salient aspects of the input need to be conveyed in the output.

Since the focus of our work was to test whether hallucinations in LLM-generated texts could be annotated, by humans and LLMs, it was important to use state-of-the-art LLMs for annotation (§4). The choice of LLMs for the underlying data-to-text generation task was less crucial, and indeed, the LLMs we used for this (§3.2) are no longer state-of-the-art. In future research, it will be interesting to see how well (future) LLM annotators will be able to categorize hallucinations committed by (future) LLM generators, and hallucinations produced by human authors as well.

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) might limit the generalizability of our findings. Future work should look at a larger number of pairs.

Ethical Considerations

Ethical approval for the human experiments conducted in this study was obtained from the Ethics Board at Utrecht University. All the annotators gave informed consent before participating in the experiment. The 12 trivago 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 (§3.2) and annotation (§4), cost us ~€80.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement № 860621.

The work was partially carried out while EC was an intern at trivago, and SM was employed there.

We thank Lasha Abzianidze, Michele Cafagna, and Massimo Poesio for constructive feedback on this work.

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.

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, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

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.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Eduardo Calò, Guanyi Chen, Elias Stengel-Eskin, Albert Gatt, and Kees van Deemter. 2025a. Incorporating formulaicness in the automatic evaluation of naturalness: A case study in logic-to-text generation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 352–365, Hanoi, Vietnam. Association for Computational Linguistics.

Eduardo Calò, Lydia Penkert, and Saad Mahamood. 2025b. 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.

Eduardo Calò, David M. Howcroft, Leo Leppänen, Saad Mahamood, Simon Mille, Patrícia Schmidtová, and Emiel van Miltenburg. 2026. Justify Your Prompts! *Computational Linguistics*, pages 1–12.

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.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint*. ArXiv:2507.06261 [cs].

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. 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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

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. 2024. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint*. ArXiv:2312.11805 [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.

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.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Stefan Hegselmann, Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2024. A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models. In *Proceedings of the fifth Conference on Health, Inference, and Learning*, volume 248 of *Proceedings of Machine Learning Research*, pages 339–379. PMLR.

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.

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.

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.

Joseph James. 2026. Counting on Consensus: Selecting the Right Inter-annotator Agreement Metric for NLP Annotation and Evaluation. *arXiv preprint*. ArXiv:2603.06865 [cs].

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.

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

Saad Mahamood and Maciej Zembrzuski. 2019. Hotel scribe: Generating high variation hotel descriptions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 391–396, Tokyo, Japan. Association for Computational Linguistics.

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*.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Aikaterini Mpagouli and Ioannis Hatzilygeroudis. 2009. A Knowledge-based System for Translating FOL Formulas into NL Sentences. In Iliadis, Maglogiann, Tsoumakasis, Vlahavas, and Bramer, editors, *Artificial Intelligence Applications and Innovations III*, volume 296, pages 157–163. Springer US, Boston, MA. Series Title: IFIP Advances in Information and Communication Technology.

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.

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.

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.

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].

Patrícia Schmidtová, Ondřej Dušek, and Saad Mahamood. 2025. Real-World Summarization: When Evaluation Reaches Its Limits. *arXiv preprint*. ArXiv:2507.11508 [cs].

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. *arXiv preprint*. ArXiv:2406.06608 [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.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. 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.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez 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. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

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].

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, 36:11809–11822.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023a. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.

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. 2023b. 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 7 and Table 8 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 F). The design of the user interface played a fundamental role in the annotation process (cf. Calò et al. 2025b), 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[6] 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 trivago database, following the same methodology described in §3.2.1. 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

[6] 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 7: Input-output pair from the hotel domain. The input is retrieved from trivago 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 (§3.2): "O too weak with harmless information": 0.67; "I and O independent": 0.33. Human annotators label distribution (§5): "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 in ToT (§5): "Well-matched": 0.50; "Well-matched with harmless information": 0.33; "O too weak with harmless information": 0.17.

Input	Output
$\exists x \forall y \forall z$ ($\text{Cube}(x) \land \text{Large}(x) \land$ ($(\text{Cube}(y) \land \text{Large}(y) \land \text{Dodec}(z)) \rightarrow$ $(x = y \land \neg\text{BackOf}(z, y))$))	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 8: 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 (§3.2): "Well-matched": 0.50; "O too weak": 0.50. Human annotators label distribution (§5): "Well-matched": 0.33; "O contradictory: 0.33; "I and O independent": 0.33. LLM annotators label distribution in ToT (§5): "O too weak": 0.33; "O contradictory: 0.33; "I and O independent": 0.33.

Table 9 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 F for the document in the hotel domain, and Appendix G 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.

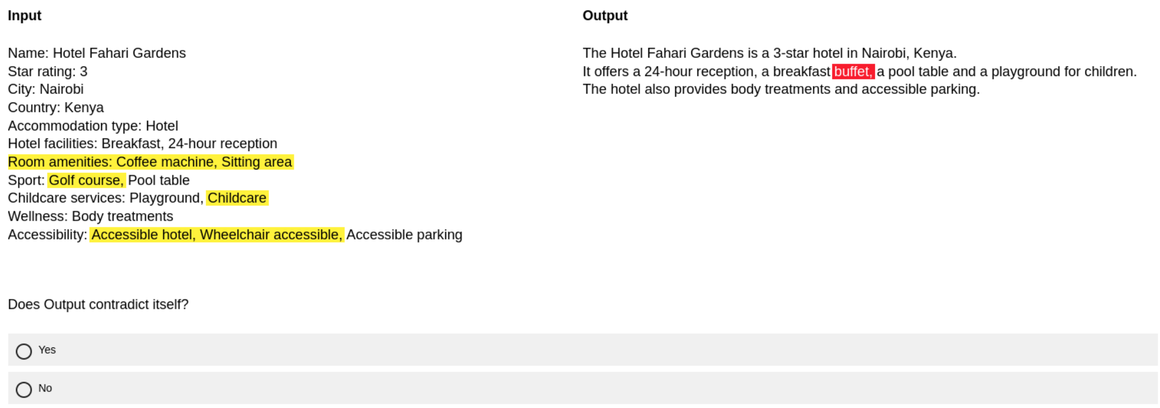

Figure 6: Annotation interface for the hotel domain with an experimental item. Divergent information in input and output is highlighted, and the first question is displayed. Subsequent questions would be revealed incrementally, based on the annotator's responses and the decision tree shown in Figure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

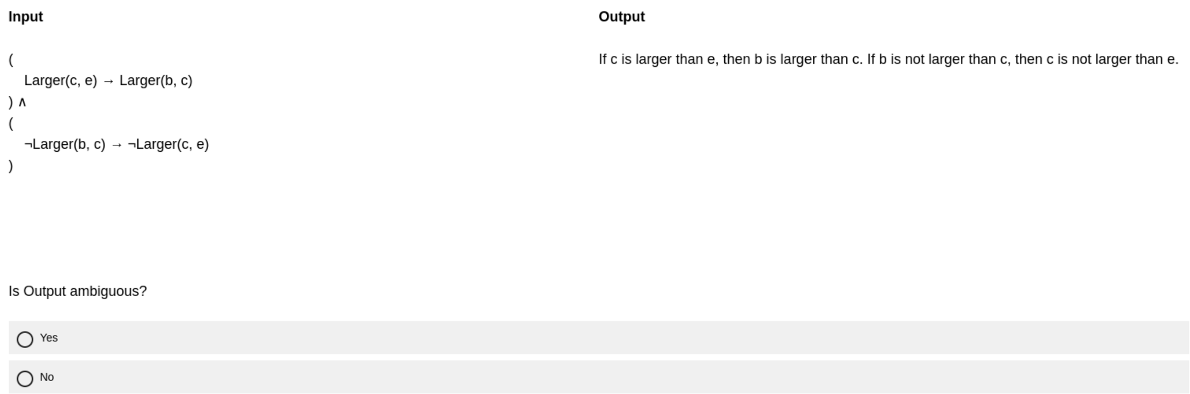

Figure 7: Annotation interface for the logic domain with an experimental item. The preliminary question on ambiguity is displayed. Subsequent questions would be revealed incrementally, based on the annotator's responses and the decision tree shown in Figure 1. The input formulae were presented in an 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 9: Comprehension check for the hotel domain. The input is retrieved from trivago 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 10 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 10: 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 11 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 11: 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 §3.2.1).[7] See Table 12 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.

[7] 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 12: 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 13 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*

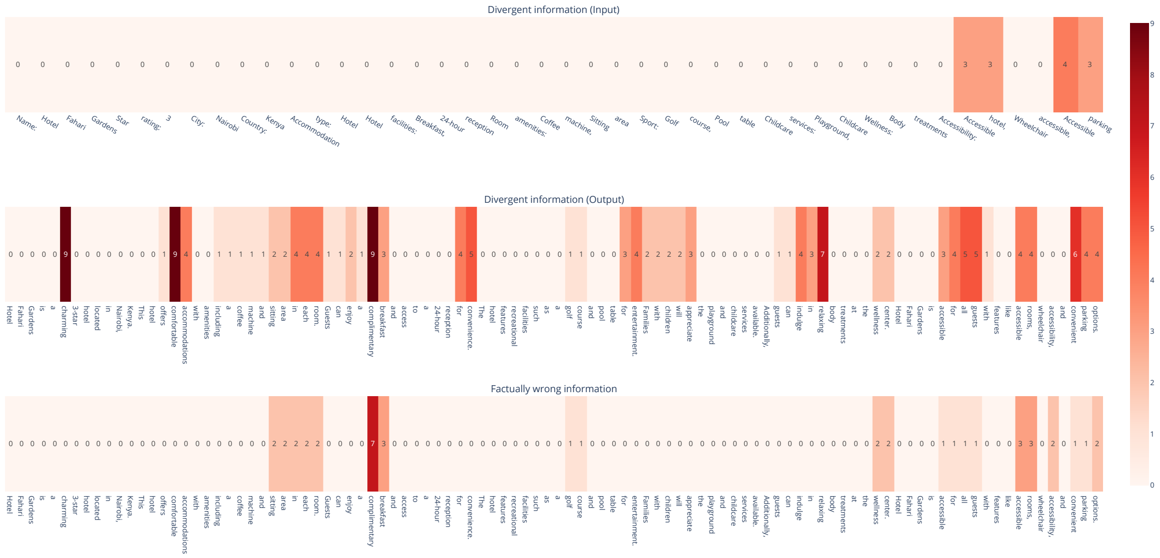

Figure 8: Heatmap of an input-output pair from the hotel domain experiment, illustrating the number of annotators who identified each token as containing divergent information in the input, divergent information in the output, or factually wrong information.

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

Table 13: 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

Figures 10, 11, 12, and 13 contain the prompts used in the LLM annotation experiment for the hotel domain, while Figures 14, 15, 16, and 17 those used for the logic domain. Figure 9 contains the *system prompt* used in all LLM annotation experiments across all prompting strategies and domains.

E Additional APA and IAA Scores

Table 14 presents the figures when APA is computed in a "harsh" way, i.e., considering only the majority reference label (i.e., a single correct label per input-output pair) from §3.2, instead of considering *any* of the reference labels for that question, as defined in §3.3. There are no important differences with respect to Table 6.

	Hotel				Logic			
	CAT	I $\models$ O	O $\models$ I	FWI	CAT	I $\models$ O	O $\models$ I	AMB
Human								
APA	0.54 (0.25)	0.84 (0.23)	0.86 (0.15)	0.77 (0.23)	0.73 (0.16)	0.85 (0.13)	0.84 (0.15)	0.80 (0.12)
IAA	0.30	0.26	0.63	0.34	0.19	0.18	0.41	0.18
Model								
DeepSeek-R1	0.56	0.72	0.88	0.80	**0.80**	0.89	**0.85**	**0.75**
Gemini-2.5	0.52	**0.96**	0.76	0.76	0.65	0.80	0.73	0.64
GPT-5	0.72	0.88	**0.96**	0.88	0.77	0.87	0.81	0.67
Grok-4	**0.76**	0.92	0.92	**0.92**	0.76	**0.91**	0.80	0.68
Opus-4.1	0.52	0.72	0.84	0.88	0.77	0.85	0.84	0.71
Sonnet-4	0.52	0.64	0.92	0.84	0.69	0.84	0.77	0.61
IAA	0.45	0.43	0.69	0.61	0.51	0.53	0.64	0.39

Table 14: "Harsh" APA and IAA for human and LLM annotators for all dimensions, in the hotel and logic domain. APA scores for human annotators are the means of single annotator scores; standard deviations are reported in brackets. Boldfaced are the higher values per LLM per dimension per domain.

```
You are an expert logician and evaluator of data-to-text/logic-to-text systems.
```

Figure 9: *System prompt* used in all annotation experiments with LLMs across all prompting strategies and domains.

```
You are participating in an experiment.
You will see an Input that is a set of structured facts and an Output that is its natural language description.
Your task is to CLASSIFY the nature of the logical relationship between Input and Output.

Ignore typos or grammar mistakes in Output.

Only return the capital letter corresponding to the correct category. Do not add explanations.

Possible Categories:
A: Input and Output are well-matched
B: Input and Output are well-matched, with Output containing harmless additions
C: Output is too weak with respect to Input
D: Output is too weak with respect to Input, with Output containing harmless additions
E: Output is too strong with respect to Input
F: Output is self-contradictory
G: Input and Output are logically independent of each other
H: Input and Output contradict each other

____________________

Now CLASSIFY the following:

Input: {input}
Output: {output}

CATEGORY (write ONLY the letter):
```

Figure 10: Zero-shot prompt for the hotel domain.

You are participating in an experiment.
You will see an Input that is a set of structured facts and an Output that is its natural language description.
Your task is to CLASSIFY the nature of the logical relationship between Input and Output.

Ignore typos or grammar mistakes in Output.

Only return the capital letter corresponding to the correct category. Do not add explanations.

Possible Categories:
A: Input and Output are well-matched
B: Input and Output are well-matched, with Output containing harmless additions
C: Output is too weak with respect to Input
D: Output is too weak with respect to Input, with Output containing harmless additions
E: Output is too strong with respect to Input
F: Output is self-contradictory
G: Input and Output are logically independent of each other
H: Input and Output contradict each other

Examples:

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a 5-star hotel in Mexico.

CATEGORY: A

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is an incredibly charming 5-star hotel in Mexico.

CATEGORY: B

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a hotel in Mexico.

CATEGORY: C

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is an incredibly charming hotel in Mexico.

CATEGORY: D

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel
Hotel facilities: Lockers

Output:
Luna Blanca is a 5-star hotel in Mexico that provides lockers and a swimming pool.

CATEGORY: E

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a 5-star hotel in New York, Mexico.

CATEGORY: F

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel
Hotel facilities: Lockers

Output:
Luna Blanca is a 5-star hotel in Mexico that provides a swimming pool.

CATEGORY: G

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a 5-star hotel in the USA.

CATEGORY: H

Now CLASSIFY the following:

Input: {input}
Output: {output}

CATEGORY (write ONLY the letter):

Figure 11: Few-shot prompt for the hotel domain.

You are participating in an experiment.
You will see an Input that is a set of structured facts and an Output that is its natural language description.
Your task is to analyze Output in light of Input based on the relation of LOGICAL CONSEQUENCE, and to CLASSIFY the nature of the logical relationship between them.

DEFINITIONS:

LOGICAL CONSEQUENCE:
Input entails Output if and only if by reading Input you can infer that Output is true. Output entails Input if and only if by reading Output you can infer that Input is true.

FACTUALLY WRONG INFORMATION:
When Input does not entail Output, Output contains information that is not in Input. This information in Output may or may not be factually wrong. We define as factually wrong any information in Output that conveys facts that could well turn out to be wrong, given the information in Input. 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 Input to hold in most (but not all) cases.

Following are EXAMPLES for each CATEGORY along with the REASONING:

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a 5-star hotel in Mexico.

REASONING: Input entails Output, and Output entails Input.
CATEGORY A: Input and Output are well-matched

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is an incredibly charming 5-star hotel in Mexico.

REASONING: Input does not entail Output, and Output entails Input, but the additional information in Output is not factually wrong.
CATEGORY B: Input and Output are well-matched, with Output containing harmless additions

---

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a hotel in Mexico.

REASONING: Input entails Output, and Output does not entail Input.
CATEGORY C: Output is too weak with respect to Input

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is an incredibly charming hotel in Mexico.

REASONING: Input does not entail Output, and Output does not entail Input, and the additional information in Output is not factually wrong.
CATEGORY D: Output is too weak with respect to Input, with Output containing harmless additions

---

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel
Hotel facilities: Lockers

Output:
Luna Blanca is a 5-star hotel in Mexico that provides lockers and a swimming pool.

REASONING: Input does not entail Output, Output entails Input, and the additional information in Output is factually wrong.
CATEGORY E: Output is too strong with respect to Input

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a 5-star hotel in New York, Mexico.

REASONING: Output contradicts itself, that is, it contains statements that cannot be true at the same time.
CATEGORY F: Output is self-contradictory

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel
Hotel facilities: Lockers

Output:
Luna Blanca is a 5-star hotel in Mexico that provides a swimming pool.

REASONING: Input does not entail Output, and Output does not entail Input, and the additional information in Output is factually wrong, but Output does not contradict Input.
CATEGORY G: Input and Output are logically independent of each other

---

Input:
Name: Luna Blanca
Star rating: 5
Country: Mexico
Accommodation type: Hotel

Output:
Luna Blanca is a 5-star hotel in the USA.

REASONING: Input does not entail Output, Output does not entail Input, and the additional information in Output is factually wrong, and Output contradicts Input.
CATEGORY H: Input and Output contradict each other

Now analyze the following new Input - Output pair. Think step-by-step with the REASONING as shown in the examples, then provide your final classification. The final line of your response must contain ONLY the CATEGORY letter (A, B, C, ...).

Input: {input}
Output: {output}

REASONING:
CATEGORY (write ONLY the letter):

Figure 12: Chain-of-thought prompt for the hotel domain.

You are participating in an experiment. You will see an Input that is a set of structured facts and an Output that is its natural language description. Your task is to analyze the pair by answering a series of specific questions.
Here are definitions and examples for the concepts you will be asked about:

LOGICAL CONSEQUENCE: Input entails Output if and only if by reading Input, you can infer that Output is true. Output entails Input if and only if by reading Output, you can infer that Input is true.

Examples:

Input:
Name: Hotel Torre Azul
City: El Arenal
Star rating: 4
Output:
Hotel Torre Azul, a 4-star hotel, is located in El Arenal, Spain.
Explanation: Output entails Input, because Output contains all the information in Input. Input does not entail Output, because Output contains some extra information (Spain).

Input:
Name: Hotel Torre Azul
City: El Arenal
Star rating: 4
Hotel facilities: Lockers
Output:
Hotel Torre Azul, a 4-star hotel, is located in El Arenal.
Explanation: Output does not entail Input, because Input contains some extra information (Hotel facilities: Lockers). Input entails Output, because Output does not contain any extra information.

Input:
Name: Hotel Torre Azul
City: El Arenal
Room amenities: Balcony
Output:
Hotel Torre Azul is a charming hotel located in El Arenal.
Explanation: Output does not entail Input, because Input contains some extra information (Room amenities: Balcony). Input does not entail Output, because Output contains some extra information (charming).

FACTUALLY WRONG INFORMATION: When Input does not entail Output, Output contains information that is not in Input. This information may or may not be factually wrong. We define as factually wrong any information in Output that conveys facts that could well turn out to be wrong, given the information in Input. 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 Input to hold in most but not all cases.

Example:

Input:
Name: Hotel Torre Azul
City: El Arenal
Country: Spain
Star rating: 4
Hotel facilities: Housekeeping
Room amenities: Balcony
Output:
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.
Explanation: The values 3, Mexico, and lockers are pieces of factually wrong information. The expressions incredibly charming (i), the center of (ii), All (iii), free (iv) are not pieces of factually wrong information.

CONTRADICTION: Input and Output contradict each other if they contain information that cannot be true simultaneously.

Example:

Input:
Name: Hotel Torre Azul
City: El Arenal
Star rating: 4
Output:
Hotel Torre Azul, a 3-star hotel, is located in El Arenal.
Explanation: Input and Output contradict each other, since Input states that the hotel has 4 stars, while Output states that the hotel has 3 stars.

SELF-CONTRADICTION: Output contradicts itself if it contains pieces of information that cannot be true simultaneously.

Example:

Input:
Name: Hotel Torre Azul
City: El Arenal
Star rating: 4
Hotel facilities: Lockers
Output:
Hotel Torre Azul, a 4-star hotel, is located in El Arenal. The hotel facilities include lockers. The hotel does not provide lockers.
Explanation: Output contradicts itself, since Output states that the hotel provides lockers and at the same time states that it does not provide them.

This is a new Input - Output pair. Please answer the question based on the definitions and examples above.

Input: {input}
Output: {output}

Based on all the above, answer the following question. The final line of your response must contain ONLY YES or NO.

Question: Does Output contradict itself?

Figure 13: Tree-of-thought prompt for the hotel domain, with the first question displayed. Subsequent questions would be revealed incrementally, based on the annotator's responses and the decision tree shown in Figure 1.

You are participating in an experiment.
You will see an Input that is a first-order logic formula, and an Output that is its natural language translation.
Your task is to CLASSIFY the nature of the logical relationship between Input and Output.

Ignore typos or grammar mistakes in Output.

Only return the capital letter corresponding to the correct category. Do not add explanations.

Possible Categories:
A: Input and Output are well-matched
B: Output is too weak with respect to Input
C: Output is too strong with respect to Input
D: Output is self-contradictory
E: Input and Output are logically independent of each other
F: Input and Output contradict each other

Now CLASSIFY the following:

Input: {input}
Output: {output}

CATEGORY (write ONLY the letter):

Figure 14: Zero-shot prompt for the logic domain.

You are participating in an experiment.
You will see an Input that is a first-order logic formula, and an Output that is its natural language translation.
Your task is to CLASSIFY the nature of the logical relationship between Input and Output.

Ignore typos or grammar mistakes in Output.

Only return the capital letter corresponding to the correct category. Do not add explanations.

Possible Categories:
A: Input and Output are well-matched
B: Output is too weak with respect to Input
C: Output is too strong with respect to Input
D: Output is self-contradictory
E: Input and Output are logically independent of each other
F: Input and Output contradict each other

Examples:

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All cubes are large.

CATEGORY: A

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All red cubes are large.

CATEGORY: B

Input:
$\forall x((\text{Cube}(x) \land \text{Green}(x)) \rightarrow \text{Large}(x))$

Output:
All cubes are large.

CATEGORY: C

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All large cubes are not large.

CATEGORY: D

Input:
$\forall x((\text{Cube}(x) \land \text{Green}(x)) \rightarrow \text{Large}(x))$

Output:
All red cubes are large.

CATEGORY: E

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All cubes are not large.

CATEGORY: F

Now CLASSIFY the following:

Input: {input}
Output: {output}

CATEGORY (write ONLY the letter):

Figure 15: Few-shot prompt for the logic domain.

You are participating in an experiment.
You will see an Input that is a first-order logic formula, and an Output that is its natural language translation.
Your task is to analyze Output in light of Input based on the relation of LOGICAL CONSEQUENCE, and to CLASSIFY the nature of the logical relationship between them.

DEFINITIONS:

LOGICAL CONSEQUENCE:
Input entails Output if and only if by reading Input you can infer that Output is true. Output entails Input if and only if by reading Output you can infer that Input is true.

Following are EXAMPLES for each CATEGORY along with the REASONING:

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All cubes are large.

REASONING: Input entails Output, and Output entails Input.
CATEGORY A: Input and Output are well-matched

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All red cubes are large.

REASONING: Input entails Output, and Output does not entail Input.
CATEGORY B: Output is too weak with respect to Input

Input:
$\forall x((\text{Cube}(x) \wedge \text{Green}(x)) \rightarrow \text{Large}(x))$

Output:
All cubes are large.

REASONING: Input does not entail Output, and Output entails Input.
CATEGORY C: Output is too strong with respect to Input

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All large cubes are not large.

REASONING: Output contradicts itself (it contains statements that cannot be true at the same time).
CATEGORY D: Output is self-contradictory

Input:
$\forall x((\text{Cube}(x) \wedge \text{Green}(x)) \rightarrow \text{Large}(x))$

Output:
All red cubes are large.

REASONING: Input does not entail Output, and Output does not entail Input, and Output does not contradict Input.
CATEGORY E: Input and Output are logically independent of each other

---

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$

Output:
All cubes are not large.

REASONING: Input does not entail Output, Output does not entail Input, and Output contradicts Input.
CATEGORY F: Input and Output contradict each other

Now analyze the following new Input - Output pair. Think step-by-step with the REASONING as shown in the examples, then provide your final classification. The final line of your response must contain ONLY the CATEGORY letter (A, B, C, ...).

Input: {input}
Output: {output}

REASONING:
CATEGORY (write ONLY the letter):

Figure 16: Chain-of-thought prompt for the logic domain.

You are participating in an experiment.
You will see an Input that is a first-order logic formula, and an Output that is its natural language translation.
Your task is to analyze the pair by answering a series of specific questions.

Here are definitions and examples for the concepts you will be asked about:

LOGICAL CONSEQUENCE:
Input entails Output if and only if by reading Input, you can infer that Output is true.
Output entails Input if and only if by reading Output, you can infer that Input is true.

Examples:

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$
Output:
All red cubes are large.
Explanation: Input entails Output, because if all cubes are large, then all red cubes are large too. Output does not entail Input, because even if all red cubes are large, it may well be the case that there are other cubes that are not large.

Input:
$\forall x((\text{Cube}(x) \wedge \text{Green}(x)) \rightarrow \text{Large}(x))$
Output:
All cubes are large.
Explanation: Input does not entail Output, because even if all green cubes are large, it may well be the case that there are other cubes that are not large. Output entails Input, because if all cubes are large, then all green cubes are large too.

Input:
$\forall x((\text{Cube}(x) \wedge \text{Green}(x)) \rightarrow \text{Large}(x))$
Output:
All red cubes are large.
Explanation: Input does not entail Output, because even if all green cubes are large, that does not say anything about red cubes. Output does not entail Input, because even if all red cubes are large, that does not say anything about green cubes.

CONTRADICTION:
Input and Output contradict each other if they contain information that cannot be true simultaneously.

Example:

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$
Output:
All cubes are not large.
Explanation: Input and Output contradict each other, since Input states that all cubes are large, while Output states that all cubes are not large.

SELF-CONTRADICTION:
Output contradicts itself if it contains pieces of information that cannot be true simultaneously.

Example:

Input:
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$
Output:
All large cubes are not large.
Explanation: Output contradicts itself, since Output states that, at the same time, all cubes are large and not large.

AMBIGUITY:
Output is ambiguous if by reading Output, you perceive distinct interpretations for Output.

Example:

Input:
$\forall x(\text{Cube}(x) \rightarrow \exists y(\text{Tetrahedron}(y) \wedge \text{Behind}(x, y)))$
Output:
Every cube is behind a tetrahedron.
Explanation: Output is ambiguous, because you could perceive at least two distinct interpretations: (i) each cube is behind a (possibly different) tetrahedron, or (ii) there is some tetrahedron that is in front of all cubes.

This is a new Input - Output pair. Please answer the question based on the definitions and examples above.

Input: {input}
Output: {output}

Based on all the above, answer the following question. The final line of your response must contain ONLY YES or NO.

Question: Is Output ambiguous?

Figure 17: Tree-of-thought prompt for the logic domain. 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.

F 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.

G 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)
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$	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
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$	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
$\forall x((\text{Cube}(x) \land \text{Green}(x)) \rightarrow \text{Large}(x))$	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
$\forall x((\text{Cube}(x) \land \text{Green}(x)) \rightarrow \text{Large}(x))$	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
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$	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
$\forall x(\text{Cube}(x) \rightarrow \text{Large}(x))$	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
$\forall x(\text{Cube}(x) \rightarrow \exists y(\text{Tet}(y) \wedge \text{Behind}(x, y)))$	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.
Large$r(x, y)$	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.