SemEval-2024 Task 6 : SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes

Mickus, Timothee; Zosa, Elaine; Vazquez, Raul; Vahtola, Teemu; Tiedemann, Jörg ...

Ojha, Atul Kr.; Doruöz,, A. Seza; Tayyar Madabushi, Harish; Da San Martino, Giovanni; Rosenthal, Sara ...

2024-06-01

http://hdl.handle.net/10138/579095

Mickus, T, Zosa, E, Vazquez, R, Vahtola, T, Tiedemann, J, Segonne, V, Raganato, A & Apidianaki, M 2024, SemEval-2024 Task 6 : SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes. in A K Ojha, A S Doruöz, H Tayyar Madabushi, G Da San Martino, S Rosenthal & A Rosá (eds), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). The Association for Computational Linguistics, Stroudsburg, pp. 1979-1993, International Workshop on Semantic Evaluation, Mexico City, Mexico, 20/06/2024. https://doi.org/10.18653/v1/2024.semeval-1.273

SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes

Timothee Mickus **Elaine Zosa** **Raúl Vázquez** **Teemu Vahtola**
Jörg Tiedemann **Vincent Segonne** **Alessandro Raganato** **Marianna Apidianaki**
University of Helsinki Silo AI, Finland Université Bretagne Sud
University of Milano-Bicocca University of Pennsylvania
{firstname.lastname}@{helsinki.fi, silo.ai, univ-ubs.fr, unimib.it}
marapi@seas.upenn.edu

Abstract

This paper presents the results of the SHROOM, a shared task focused on detecting hallucinations: outputs from natural language generation (NLG) systems that are fluent, yet inaccurate. Such cases of overgeneration put in jeopardy many NLG applications, where correctness is often mission-critical. The shared task was conducted with a newly constructed dataset of 4000 model outputs labeled by 5 annotators each, spanning 3 NLP tasks: machine translation, paraphrase generation and definition modeling.

The shared task was tackled by a total of 58 different users grouped in 42 teams, out of which 26 elected to write a system description paper; collectively, they submitted over 300 prediction sets on both tracks of the shared task. We observe a number of key trends in how this approach was tackled—many participants rely on a handful of model, and often rely either on synthetic data for fine-tuning or zero-shot prompting strategies. While a majority of the teams did outperform our proposed baseline system, the performances of top-scoring systems are still consistent with a random handling of the more challenging items.

1 Introduction

The modern NLG landscape is plagued by two interlinked problems: On the one hand, our current neural models have a propensity to produce inaccurate but fluent outputs; on the other hand, our metrics are most apt at describing fluency, rather than correctness. This leads neural networks to "hallucinate", i.e., produce fluent but incorrect outputs that we currently struggle to detect automatically. For instance, Dopierre et al. (2021) report that when trying to produce a paraphrase for the input "*I am not sure where my phone is*", they obtain the following 'hallucination' behavior: "*How can I find the location of any Android mobile*". For many NLG applications, the correctness of an output is however mission critical. For instance, producing a plausible-sounding translation that is inconsistent with the source text puts in jeopardy the usefulness of a machine translation pipeline.

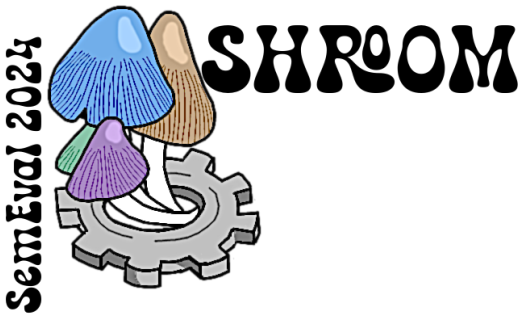


Figure 1: The SHROOM logo.

This motivates us to organize a **S**hared-task on **H**allucinations and **R**elated **O**bservable **O**vergeneration **M**istakes, or SHROOM. With our shared task, we hope to foster the growing interest in this topic in the community (e.g., Ji et al., 2023; Raunak et al., 2021; Guerreiro et al., 2023; Xiao and Wang, 2021; Guo et al., 2022). In particular, in the SHROOM we adopt a *post hoc* setting, where models have already been trained and outputs already produced. Participants were asked to perform binary classification to identify cases of **fluent overgeneration hallucinations** in two different setups: **model-aware** and **model-agnostic** tracks. That is, participants had to detect grammatically sound outputs which contain incorrect or unsupported semantic information, inconsistent

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1979–1993

with the reference input, with or without having access to the model that produced the output.

To that end, we constructed a dataset comprising a collection of checkpoints, inputs, references and outputs of systems covering three different NLG tasks: definition modeling (DM, Noraset et al., 2017), machine translation (MT) and paraphrase generation (PG) trained with varying degrees of accuracy. Datapoints were all annotated by 5 human annotators each resulting in 1000 validation items and 3000 test items.

Beyond simply detecting factually unsupported outputs, one of the goals of this shared task was to establish whether hallucinations are best construed as a categorical phenomenon or a gradient one. Similar remarks have been made with respect to textual entailment (Bowman et al., 2015). As such, participants' submission were scored both for accuracy (whether classifiers correctly identify hallucinations) and calibration (whether classifiers are confident about their prediction when they ought to be).

The shared task attracted a total of 58 different users grouped in 42 teams, out of which 26 elected to write a system description paper. Collectively, over the three weeks of the evaluation phase, participants submitted 300 valid sets of predictions on the model-aware track, and 320 on the model-agnostic track. We take this participation rate, along with the breadth of methodological approaches developed by participants, as clear signs of success for our shared task: This large pool of participants allows us to identify and discuss some key trends in how the task was tackled. Crucially, many participants rely on a handful of model, and often rely either on synthetic data for fine-tuning or zero-shot prompting strategies. In terms of raw performance, we note that while a majority of the teams (64 to 71%) did outperform our proposed baseline system, the performances of top-scoring systems are still consistent with a random handling of the more challenging items. In sum, this first iteration of the SHROOM underscores both an interest of the research community as well as the current limitations in our approaches.

The remainder of this article is structured as follows: Iin Section 2, we provide an overview of the current research landscape. Section 3 defines our theoretical framework, and Section 4 summarizes our data collection process. We then present and discuss shared task results in Sections 5 and 6 before concluding with a few thoughts on further research in Section 7.

2 Connecting with the past: related works and state of the art

It is now widely accepted that NLG models often generate outputs that are not faithful to the given input, commonly referred to in the community as hallucinations (Vinyals and Le, 2015; Raunak et al., 2021; Maynez et al., 2020). Yet there is minimal consensus on the optimal framework for its application. This lack of agreement is due in part to the diversity of tasks that NLG encompasses (Ji et al., 2023).

Guerreiro et al. (2023) propose a taxonomy of hallucinations that includes oscillatory productions, and fluent but strongly or fully "detached" outputs. While this taxonomy is well constructed, we find it inadequate for the needs of the community at large for four reasons: *(i)* It conflates some issues of fluency with semantic correctness (oscillatory productions are cases of non-fluent overgeneration where no extraneous semantic material is introduced); *(ii)* It only considers the most extreme cases of hallucinations (strongly or fully detached productions), whereas diagnosis of intermediary cases is bound to be more challenging and useful to the community; *(iii)* It focuses only on MT, although other tasks are also known to suffer from fluent overgeneration (e.g., Rohrbach et al., 2018), including the ones we propose to address; *(iv)* It uses only lowest scoring outputs, whereas any tool built to verify system outputs ought not to flag non-pathological outputs.

Alternative studies have built benchmarks for hallucination detection, with a predominant emphasis on dialogue systems. Li et al. (2023) propose the HaluEval benchmark using an annotation framework that does not necessarily center on the input given to the model and requires the annotators to search the internet for facts. Moreover, they opted to annotate the outputs of a popular LLM, with the major downsides that it is closed, not-transparent and commercial; rendering the research outputs that may stem from future studies less interesting. Other benchmarks include the works of Liu et al. (2022) and Zhou et al. (2021), which automatically insert hallucinations into training instances to generate syntactic data for token-level hallucination detection; Lin et al. (2022), which work with factual claims supported by reliable, publicly available

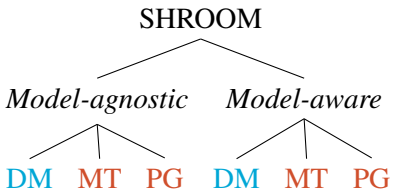


Figure 2: Shared task overview. Both tracks feature all three NLG tasks. Datapoints from systems in blue correspond to target-referential datapoints and in red the ones that are either target- or source-referential; which we refer to as *dual-referential*.

evidence; and Dziri et al. (2022), which focus on knowledge-based dialogue systems and base their annotation on NLI, relying only on the system's input, just as we do.

3 Tripping over hallucinations: task definition and annotation

In contrast with previous works (e.g. Guerreiro et al., 2023; Li et al., 2023), we focus on cases of fluent overgeneration since judgments pertaining to the over-generative nature of a production can be elicited by means of **inferential semantics**: if an output cannot be inferred from its semantic reference, then it contains some information that is not present in the reference—i.e., the model has generated more than we expected.[1] This approach connects with the theoretical framework sketched by van Deemter (2024), who likewise relies on inferential semantics but also considers undergeneration issues in NLG outputs. We provide multiple annotations and a gold majority label, given the low consensus on semantic annotations (Nie et al., 2020).

In Figure 2 we provide an overview of the task. The SHROOM is framed around two key distinctions: (i) model-aware vs. model-agnostic approaches, and (ii) source-referential vs. dual-referential datapoints. The former corresponds to whether participants have access to the model that generated the item: **Model-agnostic** approaches are practical, as models may not be accessible to end users; **Model-aware** approaches can lead to richer and more accurate diagnoses. The latter is a consequence of our inferential take on over-

[1]Note that if the output can be inferred from the reference but the information is not explicitly present in the reference, then the model is actually making a correct semantic inference: it is generating a semantically sound output. E.g., if the the model produces "my tie is blue" for the reference "my tie is the color of the sky", the model output is semantically sound.

```
{ "hyp":"A cigarette .",
  "ref":"tgt",
  "src":"I stepped outside to smoke myself a j .
         What is the meaning of J ?",
  "tgt":"( plural Js or J 's ) A marijuana
          cigarette .",
  "model":"ltg\/flan-t5-definition-en-base",
  "task":"DM",
  "labels":["Hallucination","Not Hallucination",
            "Not Hallucination", "Hallucination",
            "Hallucination"],
  "label":"Hallucination",
  "p(Hallucination)":0.6  }
```

Figure 3: Target-referential datapoint example from the validation set for the model-aware track.

generation: what can effectively serve as a semantic reference varies across NLP systems. For DM, where we fine-tune a language model to produce a definition for a given example of usage the datapoints are **target-referential**, i.e. the target is the sole usable semantic reference. In this context, the target serves as the sole usable semantic reference. Conversely, the target is expected to be semantically implied from the source in source-referential tasks, such as summarization. Note that we do not annotate source-referential tasks due to annotation challenges that make them unreliable for our purposes. In dual-referential tasks like PG & MT, this distinction bears no weight.

In Figure 3, we present an example datapoint displaying how we plan to encode all relevant information in a JSON format is provided. The datapoint keeps track of the source provided to the model as input (`src`), the intended target (`tgt`), the model production (`hyp`), the task this production was derived from (`task`), can correspond to DM, MT or PG), whether this datapoint is target-referential (`ref`), the annotations, the gold label and the proportion of annotators that labeled the utterance as a hallucination (`labels`, `label`, and `p(Hallucination)`). In the model-aware track, we will also provide a HuggingFace model name (`model`).

4 Foraging and harvesting season: Collected data

All SHROOM data (models, outputs and annotations) are available under a CC-BY license.[2]

[2]See helsinki-nlp.github.io/shroom

4.1 Data & model provenance

Participants have access to generated outputs from multiple systems trained to generate English output at various stages of their training, stemming from three sequence-to-sequence NLG tasks: DM, MT and PG. The SHROOM dataset consists of annotated *test* and *dev* sets, as well as a *unlabeled training split* of 30k datapoints per track and the full set of possible target references to allow corpus-wide approaches. To ensure effective annotation of the development and test sets, and to be able to guarantee a gradient in quality as measured by automated metrics, we pre-selected fluent outputs for the annotators, which we describe in the following.[3]

MT: For the model agnostic track we use the models from Mickus and Vázquez (2023). We compute perplexity for the all MT outputs and BERTScores with regards to the outputs and corresponding targets. We filter outputs with perplexity scores above the 2% quantile. From the filtered outputs, we randomly select 200 samples with BERTscores in the 1/7, 2/7, 3/7, 4/7, and 5/7 quantiles. For the model-aware track, we use the NLLB model (NLLB Team et al., 2022) and produce translations on the Flores-200 dataset from languages marked as low-resource to English. Next, we manually select a sample that is sufficiently fluent.

DM: We use the model of Segonne and Mickus (2023) for the model-agnostic track, and for the model-aware track we used the `flan-t5-definition-en-base` (Giulianelli et al., 2023). We generate outputs on the English portion of the CoDWoE dataset (Mickus et al., 2022), and manually select a sample that is reasonably fluent and contains no profanities.

PG: We used a pretrained and fine-tuned paraphrasing model[4] based on Pegasus (Zhang et al., 2020) for the model-aware track, and the controlled paraphrase generation model of Vahtola et al. (2023) for the model-agnostic track.

We generated paraphrase hypotheses using Europarl (Koehn, 2005) and Opusparcus (Creutz, 2018) for the model-aware and -agnostic tracks, respectively. For the model-aware setup, we generated 50 hypotheses for each source sentence using diverse beam search (Vijayakumar et al., 2016) using BLEU scores (Papineni et al., 2002) to select the least similar hypothesis for each source sentence to serve as its paraphrase. For the model-agnostic setup, we calculated control tokens for each source sentence as in Vahtola et al. (2023), scaled the length-controlling value in range (1, 1.5) with a uniform probability distribution to provoke hallucination in the generated sequences, and used beam search with a beam size of 5 to produce the paraphrases. We manually curated the final validation and test examples.

4.2 Annotation

We annotate a total of 4,000 items, which are split 25%–75% between development and test sets: 1000 datapoints come from PG, 1500 from DM and 1500 from MT. Each item is annotated by five annotators on whether the reference entails the output. Annotations are binary, for ease of dataset construction. Gold labels are defined with respect to the annotators' majority vote.

The annotators were enlisted via Prolific,[5] a paid platform specialized in gathering human data for research studies and AI dataset creation, among other purposes. We did not target any particular group of participants; the only screening prerequisites were that (i) participants had to be fluent in English and (ii) they should not have taken part in an initial pilot study.

We used Potato (Pei et al., 2022), an open-source annotation tool specifically designed to seamlessly integrate with Prolific. Annotators were first presented with a pre-annotation screen outlining the annotation guidelines, after which they commenced the annotation of items individually. Each item consisted of the Reference, the AI-generated output, and relevant context regarding the NLG task (DM, MT, or PG). The annotators were asked to answer the question *"Does the following AI output only contain information supported by the Reference?"* responding with either "yes" or "no," and were also given the opportunity to provide comments if necessary. Additionally, they could navigate back and forth through their assigned items. We set up a timer that notified the participants every 60 seconds of the time spent on an item. In Appendix B, we present a copy of the instructions we used.

To control for annotation quality, we manually reviewed annotations from two sets of selected an-

[3] Note that we do not warranty that the training split contains fluent outputs.

[4] https://huggingface.co/tuner007/pegasus_paraphrase

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

notators: (i) five randomly selected annotators; and (ii) the five annotators who completed the task the fastest (under 3.5 minutes). All 10 annotators completed 20 annotations each. We judged all 200 annotations to be sound, in that a reasoning could be reconstructed to explain the provided annotation.

Label distribution. Figure 4 provides an overview of the distribution of labels in the SHROOM dataset splits (validation and test), broken down per NLG task (MT, DM and PG) and track (model-aware vs. model-agnostic). In this figure, we consider the empirical probability that a given item is judged to be a hallucination, i.e., the proportion of annotators judging the NLG output is not supported by the intended semantic reference.

We can highlight two trends in this figure. The first one, and perhaps most important, is that hallucinations are not consensual among our annotators. If intuitions regarding hallucinations were clear-cut, we would strongly expect a bi-modal distribution of empirical label distributions being consistently judged as hallucinations or not hallucinations. Instead, we find a number of intermediate cases, where annotators are split: These account for 29–32% of the data, depending on the split (validation or test) and track (model-aware or model-agnostic). Given the small number of annotators per datapoint, we cannot confidently rule out the possibility of a sampling bias—it is plausible that a larger pool of annotator would yield more bimodal empirical distributions. On the other hand, this tentative evidence is also in line with what has been argued elsewhere for natural language inference (Nie et al., 2020; Zhou et al., 2022). This is in fact well exemplified by the datapoint provided in Figure 3: Whether the term *cigarette* is underspecified and can apply to any smokable substance, or whether it is to be understood as prototypically referring to tobacco cigarettes by default is, in fact, up for discussion—and it stands to reason that different speakers may form different opinions.

Second, it is difficult to find hallucinations: The higher the empirical probability, the fewer the datapoints. This is especially true in the PG task: these outputs rarely yields consensual hallucinations, whereas we can find such items in DM and MT much more frequently. Looking at the expected value of the empirical probability per task, we find that DM consistently ranks higher than MT, which in turns ranks higher than PG. Both of these differences are significant under a one-sided Mann-Whitney U-test in the two test tracks ($p < 0.0003$); in the model-aware validation dataset, only the difference between MT and PG is significant ($p < 2 \cdot 10^{-8}$), in the model-agnostic validation dataset, only the difference between DM and MT is ($p < 0.04$). We note that DM requires a more complex processing of its input, as it has to rely on facts captured by the underlying LLM during its pre-training phase; for MT and PG, the input of the NLG task contains the semantic information necessary to produce a valid output. As such, we conjecture that the difficulty of an NLG task fosters hallucinatory behavior.[6]

5 They got so high: shared task results

The competition was held via Codalab (Pavao et al., 2023). The leaderboard was left hidden during the evaluation phase (i.e., participants were not notified of their submissions' scores until the end of the evaluation phase) but users were allowed to make a high number of submissions (50).

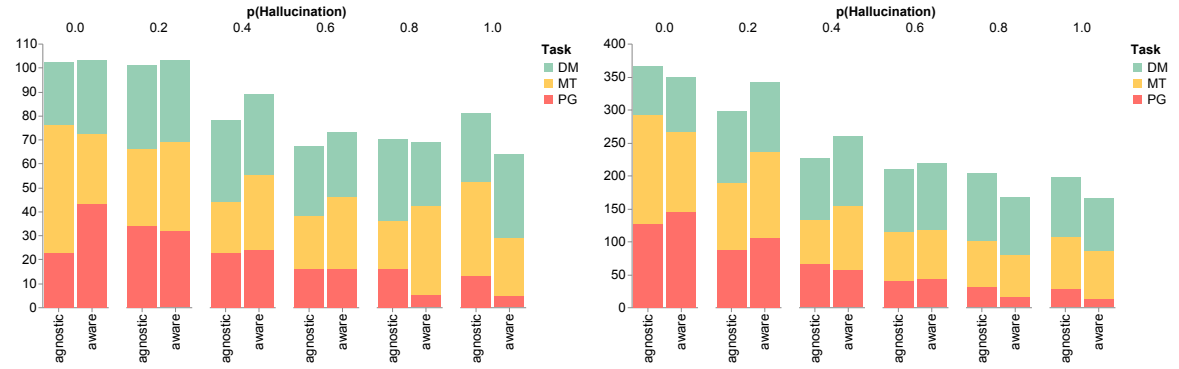
Systems are evaluated according to two criteria: the **accuracy** that the system reached on the binary classification, and their **calibration**, measured as the Spearman correlation of the systems' output probabilities with the proportion of the annotators marking the item as overgenerating. We rank systems by accuracy and break possible ties using calibration.

5.1 Baseline system

As a baseline for the task, we use an LLM[7] to evaluate whether the generated hypotheses are coherent with the provided context. Drawing upon Manakul et al. (2023), we use the prompt template listed in Figure 5. The system of Manakul et al. (2023), which has gathered some attention from the community, constitutes a straightforward approach based on a modern LLM, and is therefore well-suited to serve as a baseline in our shared-task:

[6] We also remark that the two tracks are broadly comparable in terms of hallucinatory content. Two-samples Kolmogorov-Smirnov tests for either split (test or validation) do not provide sufficient grounds to suggest a difference of distribution in labels between model-aware and model-agnostic tracks—which again suggests that the relevant difference is at the task level, rather than at the model level.

[7] We use quantized `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), from the `Hugging Face hub` `huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF` or the `llamacpp` project `github.com/ggerganov/llama.cpp`.



(a) Validation split

(b) Test split

Figure 4: Distributions of annotations

```
Context: {}
Sentence: {}
Is the Sentence supported by the Context above?
Answer using ONLY yes or no:
```

Figure 5: Prompt template used in the baseline system, adapted from Manakul et al. (2023).

it corresponds to a reasonable default approach to tackle the problem we challenge participants with.

The specific context varies depending on the task addressed, i.e. the source sentence for the paraphrase generation task, and the target sentence for machine translation and definition modeling tasks. As for the probability of hallucination, we rely on the probability assigned by the model to the first output word.[8] In cases where the output does not clearly indicate *yes* or *no*, we randomly select one, attributing a hallucination probability of 0.5.

On the model-agnostic track, our baseline system achieves an accuracy of 0.697 (with a calibration of $\rho = 0.403$), on the model-aware track, we observe an accuracy of 0.745 (with $\rho = 0.488$). We can also indicate some other simple heuristics, such as picking the most frequent label (viz., `Not Hallucination`): In this case, one would expect an accuracy of 0.593 on the model-agnostic track, and 0.633 on the model-aware track. A purely random guess between the two possible labels would result in an accuracy of 0.5. In short, our baseline systems systematically outperforms these crude heuristics.

5.2 Participating teams

A total of 59 individual users grouped in 42 teams participated in the shared task, out of which 26 elected to write a system description paper. During the evaluation phase, we received a total of 512 submissions, out of which 368 were successful. 264 of these submissions targeted both tracks, while 68 only targeted the model-agnostic track, and 36 only targeted the model-aware track. That is, we received 332 model-agnostic submissions and 300 model-aware submissions.

We present the model-agnostic track rankings in Table 1a and the model-aware track in Table 1b. As one might expect, there is a high correlation between the accuracy and calibration scores of each team's top ranking submission, which translates into a Spearman's ρ correlation of 0.909 on the model-agnostic track and 0.949 on the model-aware track. Most of the top submissions per team rank above our baseline ($30/42 \approx 71.4\%$ in the model-agnostic track, $25/39 \approx 64.1\%$ in the model-aware track). This appears roughly in line with all submissions globally: 69.9% of all model-agnostic submissions and 57.0% of all model-aware submissions score higher than our baseline.

Another point worth stressing is that teams that fare well on one track usually fare equally well on the other: For the 38 teams participating in both tracks, we find that the rank they obtain on the model-aware track correlates with the rank they obtain on the model-agnostic track (Spearman's $\rho = 0.884$). This would tentatively suggest that participants could not effectively leverage the supplementary data available in the model-aware track.[9]

Lastly, we note that there is a ceiling in terms

[8] We note that this simple heuristic may not accurately represent the true hallucination probability.

[9] An alternative account would be that all teams that participated in both tracks equally benefited from the access to the model weights, which we deem much less likely.

	team	Acc	ρ
1	Halu-NLP (Mehta et al., 2024)	0.847	0.770
2	OPDAI (Chen et al., 2024)	0.836	0.732
3	HIT-MI&T Lab (Liu et al., 2024)	0.831	0.768
4	SHROOM-INDElab (Allen et al., 2024)	0.829	0.721
5	Alejandro Mosquera	0.826	0.709
6	DeepPavlov (Belikova and Kosenko, 2024)	0.821	0.752
7	BruceW	0.821	0.735
8	TU Wien (Arzt et al., 2024)	0.817	0.737
9	SmurfCat (Rykov et al., 2024)	0.814	0.723
10	HaRMoNEE (Obiso et al., 2024)	0.814	0.626
11	AMEX AI LABS	0.813	0.728
12	Pollice Verso (Kobs et al., 2024)	0.803	0.676
13	MALTO (Borra et al., 2024)	0.801	0.681
14	UCC-NLP	0.795	0.664
15	Team CentreBack	0.792	0.623
16	Atresa	0.788	0.646
17	ustc_xsong	0.785	0.695
18	IRIT-Berger-Levrault (Bendahman et al., 2024)	0.783	0.636
19	silk_road	0.781	0.672
20	AILS-NTUA (Grigoriadou et al., 2024)	0.778	0.668
21	zhuming	0.773	0.481
22	SibNN	0.770	0.613
23	UMUTeam (Pan et al., 2024)	0.769	0.561
24	Noot Noot (Bahad et al., 2024)	0.765	0.584
25	HalluSafe (Rahimi et al., 2024)	0.763	0.629
26	Maha Bhaashya (Bhamidipati et al., 2024)	0.749	0.605
27	DUTh (Iordanidou et al., 2024)	0.744	0.475
28	Compos Mentis (Das and Srihari, 2024)	0.738	0.595
29	daixiang	0.737	0.583
30	NU-RU (Markchom et al., 2024)	0.728	0.595
	baseline system	0.697	0.403
31	SLPL SHROOM (Fallah et al., 2024)	0.694	0.423
32	Skoltech	0.684	0.674
33	CAISA	0.677	−0.430
34	AlphaIntellect (Choudhury et al., 2024)	0.654	0.295
35	deema	0.646	0.566
36	BrainLlama (Siino, 2024)	0.625	0.204
37	Byun (Byun, 2024)	0.617	0.239
38	Bolaca (Rösener et al., 2024)	0.613	0.217
	most frequent guess	0.593	
39	AI Blues	0.587	0.025
	random guess	0.500	
40	MARiA (Sanayei et al., 2024)	0.498	0.025
41	0x.Yuan	0.461	0.134

(a) Model-agnostic track rankings

	team	Acc	ρ
1	HaRMoNEE (Obiso et al., 2024)	0.813	0.699
2	Halu-NLP (Mehta et al., 2024)	0.806	0.715
3	TU Wien (Arzt et al., 2024)	0.806	0.707
4	OPDAI (Chen et al., 2024)	0.805	0.680
5	HIT-MI&T Lab (Liu et al., 2024)	0.805	0.712
6	SHROOM-INDElab (Allen et al., 2024)	0.802	0.656
7	AMEX AI LABS	0.801	0.696
8	DeepPavlov (Belikova and Kosenko, 2024)	0.799	0.713
9	silk_road	0.798	0.687
10	AILS-NTUA (Grigoriadou et al., 2024)	0.795	0.685
11	BruceW	0.794	0.660
12	Team CentreBack	0.789	0.606
13	UCC-NLP	0.789	0.644
14	ustc_xsong	0.787	0.658
15	UMUTeam (Pan et al., 2024)	0.784	0.507
16	HalluSafe (Rahimi et al., 2024)	0.783	0.537
17	SmurfCat (Rykov et al., 2024)	0.783	0.671
18	Atresa	0.783	0.624
19	IRIT-Berger-Levrault (Bendahman et al., 2024)	0.781	0.601
20	Pollice Verso (Kobs et al., 2024)	0.777	0.601
21	NU-RU (Markchom et al., 2024)	0.768	0.582
22	zhuming	0.768	0.472
23	SibNN	0.763	0.587
24	Compos Mentis (Das and Srihari, 2024)	0.756	0.566
25	DUTh (Iordanidou et al., 2024)	0.755	0.528
	baseline system	0.745	0.488
26	AlphaIntellect (Choudhury et al., 2024)	0.711	0.426
27	SLPL SHROOM (Fallah et al., 2024)	0.706	0.426
28	deema	0.688	0.519
29	BrainLlama (Siino, 2024)	0.671	0.244
30	daixiang	0.649	0.218
	most frequent guess	0.633	
31	Bolaca (Rösener et al., 2024)	0.626	0.283
32	Noot Noot (Bahad et al., 2024)	0.613	0.355
33	Byun (Byun, 2024)	0.610	0.234
34	Maha Bhaashya (Bhamidipati et al., 2024)	0.606	0.209
35	CAISA	0.567	−0.100
36	Skoltech	0.557	−0.011
37	MARiA (Sanayei et al., 2024)	0.505	0.009
	random guess	0.500	
38	octavianB (Brodoceanu, 2024)	0.483	−0.064

(b) Model-aware track rankings

Table 1: SHROOM team rankings. Codalab usernames are used to define teams when no other information was provided.

of performances: The most effective systems misclassify between 15 to 19% of all items, or almost one in every six or five datapoints. We have discussed above that, as hallucinations are a graded phenomenon, a large segment of our data (30%) corresponds to ambiguous cases where annotators are split 2 vs. 3. As such, it is worth stressing that top scores are consistent with models that classify consensual items well (where at most one annotator disagree), but perform at random chance on the more challenging ambiguous datapoints.

6 A bunch of fun guys: qualitative analysis of participants systems

We derive our analyses from system description papers as well as self-reports from a handful of participants who elected to not provide a full description of their systems. This corresponds to 33 systems

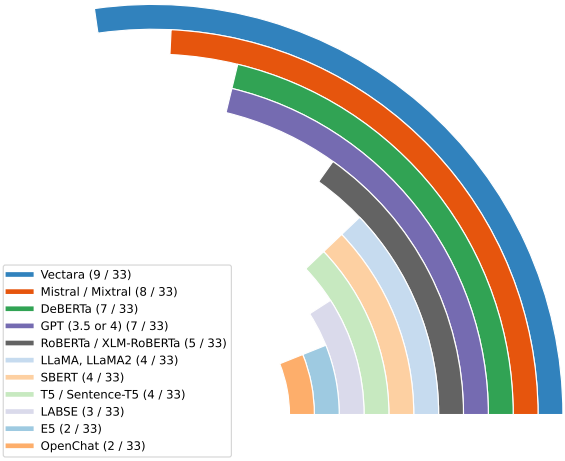


Figure 6: Known models used by more than one team. A full circle would correspond to a given model used by all of respondents, half a circle to 50% of respondents using said model. Best viewed in color.

out of the 42 identified teams that participated to the shared task, out of which 7 did not provide a full description. See also Table 2 in Appendix A for further details.

How the task was approached. The teams used a variety of methods to address the problem, ranging from ensemble techniques to fine-tuning pre-trained language models (LLMs) and prompt engineering. As expected, most teams used popular pre-trained LLMs such as GPT, LLaMA, DeBERTa, RoBERTa, and XLM-RoBERTa; Figure 6 provides a summary of which models were most popular among our teams. The Vectara hallucination evaluation model[10] turned out to be extremely popular, as more than 1 in 4 teams that provided information about their systems report having used it in their experiments. If we add other DeBERTa-based models, this number climbs to $16/33$, i.e. almost every other team used DeBERTa or a variant thereof.

Yet, the ways in which these LLMs were used cover a wide range of approaches: Some either fine-tuned on hallucination data or optimized with prompts; others employed in-context learning with role-playing, automatic prompt generation, and ensemble methods. Furthermore, some teams focused on zero-shot and few-shot approaches, while others focused on synthetic data generation and semi-supervised learning techniques to construct a labeled training set. Especially noteworthy, Rahimi et al. (2024) report constructing a manual dataset of 3000 datapoints for training their systems.

Teams predominantly relied on the data constructed for the SHROOM, although some teams added datasets such as QQP and PAWS. Interestingly, we also note five teams relying on NLI/entailment data or models, including some that achieved high results (Obiso et al., 2024; Sanayei et al., 2024; Borra et al., 2024; Liu et al., 2024 and Team Centre-Back)—and this matches the theoretical framework adopted in this shared task.

What worked well. We now turn to what distinguishes top scorers from other submissions. We note that systems based on the closed-source models GPT-3.5 and GPT-4 tend to fare well: 4 out of the 6 highest scoring systems on either track—Mehta et al. (2024); Obiso et al. (2024); Liu et al. (2024); Allen et al. (2024) and Alejandro Mosquera—all report using these models. This is however not a strict requisite as OPDAI (Chen et al., 2024) manages to rank high (2^{nd} on the model-agnostic track and 4^{th} on the model-aware track) without it. Neither does using closed-source models guarantee a high result: UCC-NLP and Markchom et al. (2024) also use GPT-3.5, and while the former is ranked 14^{th} on the model-agnostic track and 13^{th} on the model-aware track, the latter is ranked 30^{th} on the model-agnostic track and 21^{st} on the model-aware track, and only outperforms the baseline model in accuracy by 0.02 to 0.03 points.

Remarkably, many of the top-scoring approaches rely on fine-tuning (Liu et al., 2024; Obiso et al., 2024; Arzt et al., 2024; Chen et al., 2024) or ensembling (Mehta et al., 2024; Belikova and Kosenko, 2024, Alejandro Mosquera), suggesting that high performances do not come out of the box from off-the-shelf LLMs and systems. It is necessary to adapt existing models or establish to what extent their predictions are useful to the task at hand.

Another important trend we identify is that the number of submissions per team anti-correlates with the rank they obtain: The more participants submitted, the higher their best scores went. This is visualized in Figure 7: On both tracks, we find reasonable anti-correlations ($-0.58 < \rho < -0.44$) indicating that top-scorers tended to submit more. This might provide an alternative explanation for what distinguishes top-scorers from other participants: If we were to model participants' submissions as a random process, we would expect that sampling more often (i.e., submitting more) would mechanically yield a better rank.

[10] https://huggingface.co/vectara/hallucination_evaluation_model

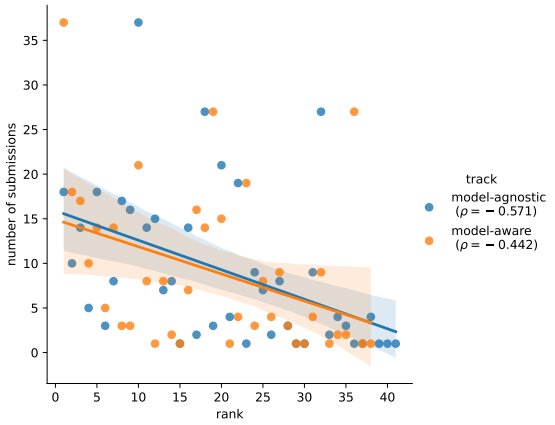


Figure 7: Rank obtained vs. number of submissions made on both tracks.

Overall, the high methodological diversity highlights the complexity of hallucination detection, even when contained the simple inferential semantics framework of our shared task: While a focus on NLI or using high-performance closed source models may help, the highest scores are obtained through thorough involvement—both in terms of model training and prediction set submissions.

7 Much room to grow: conclusions and future perspectives

This first iteration of the SHROOM shared task on detecting hallucinations has allowed us to make significant headway into understanding the confabulatory behavior of modern NLG systems. The data collected demonstrate that *hallucinations correspond to a gradient phenomenon*, and that different speakers form different opinions as to what counts as a hallucination. We were also able to showcase that *ambiguous items remain challenging*, and that the current state of the art on the dataset we provided is compatible with simple random guesses whenever the data is more ambiguous. This results underscore the massive gap that NLP research urgently needs to address: one out of every six items is still misclassified by the most effective systems showcased during this shared task.

The diversity of methodologies employed by participants underscores how *out-of-the-box solutions are not sufficient*: Highest scoring teams had to rely on fine-tuning or ensembling and made a high number of submissions. Relatedly, *access to the model parameters was of limited help*: Few approaches attempted to perform model-specific investigations, and performances on the model-aware track are in fact lower than what we observed on the model-agnostic track. Properly leveraging the parameter space for finer-grained hallucination detection remains a point for future research to investigate.

This shared task has not broached some crucial aspects and questions: How do these results translate insofar as modern LLMs—often much larger and better trained than the systems we studied here—are concerned? Can we leverage sentence-level predictions to pinpoint token-level issues with the output of our NLG systems? And will the difficulties that we underscored in this purely English be exacerbated when studying other languages—especially those that are less well-resourced and typologically different? Answering these questions and more will require further research—and perhaps future iterations of this shared task.

Overall, the success of this shared task is owed to its committed participants. We received over 350 submissions in the span of three weeks from across the world. The width of approaches studied and reported upon provides a useful snapshot of where the field is at, what approaches are favored, and what gaps still need to be overcome. We expect that the results of the SHROOM will provide a useful starting point for future work on hallucinations.

Doing SHROOM responsibly: ethical considerations

We strive to adhere to the ACL Code of Ethics.

Broader Impact. Hallucinated outputs from large language models can be used to further spread disinformation and advance misleading narratives. Detecting hallucinated outputs is an important step in elucidating the factors of this phenomena and contribute to ongoing efforts to mitigate hallucination. This leads to the development of more trustworthy generative language models.

Data and Annotators. Our annotators were suitably compensated for their work in excess of minimum wage. Due to the nature of the proposed task, the data we release might contain false or misleading statements. In the case of annotated data, these statements are labeled as such, but this does not for the unannotated portions of the data. We manually pre-filtered the data to remove profanities before providing them to annotators. Such precautions were not taken for the unannotated portion of the dataset, which might therefore contain offensive, obscene or otherwise unconscionable items.

Acknowledgments

The construction of the SHROOM dataset was made possible by a grant from the Oskar Öfflund Foundation. This work is also supported by the ICT 2023 project "Uncertainty-aware neural language models" funded by the Academy of Finland (grant agreement № 345999). We also thank the CSC-IT Center for Science Ltd., for computational resources.

The shared task logo (cf. Figure 1) uses the "Retro Cool" font from Nirmana Visual (https://nirmanavisual.com/), made available for personal / non-commercial uses.

References

Bradley Allen, Fina Polat, and Paul Groth. 2024. Shroom-indelab at semeval-2024 task 6: Zero- and few-shot llm-based classification for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 826–831, Mexico City, Mexico. Association for Computational Linguistics.

Varvara Arzt, Mohammad Mahdi Azarbeik, Ilya Lasy, Tilman Kerl, and Gábor Recski. 2024. Tu wien at semeval-2024 task 6: Unifying model-agnostic and model-aware techniques for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1172–1186, Mexico City, Mexico. Association for Computational Linguistics.

Sankalp Sanjay Bahad, Yash Bhaskar, and Parameswari Krishnamurthy. 2024. Nootnoot at semeval-2024 task 6: Hallucinations and related observable overgeneration mistakes detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 964–968, Mexico City, Mexico. Association for Computational Linguistics.

Julia Belikova and Dmitrii Kosenko. 2024. Deeppavlov at semeval-2024 task 3: Multimodal large language models in emotion reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1757–1767, Mexico City, Mexico. Association for Computational Linguistics.

Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, and Mokhtar BILLAMI. 2024. Irit-berger-levrault at semeval-2024: How sensitive sentence embeddings are to hallucinations? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 560–565, Mexico City, Mexico. Association for Computational Linguistics.

Patanjali Bhamidipati, Advaith Malladi, Manish Shrivastava, and Radhika Mamidi. 2024. Maha bhaashya at semeval-2024 task 6: Zero-shot multi-task hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1709–1713, Mexico City, Mexico. Association for Computational Linguistics.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. Malto at semeval-2024 task 6: Leveraging synthetic data for llm hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1688–1694, Mexico City, Mexico. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Octavian Brodoceanu. 2024. octavianb at semeval-2024 task 6: An exploration of humanlike qualities of hallucinated llm texts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1149–1154, Mexico City, Mexico. Association for Computational Linguistics.

Cheolyeon Byun. 2024. Semeval2024 task6 group byun. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 269–272, Mexico City, Mexico. Association for Computational Linguistics.

Ze Chen, Chengcheng Wei, Songtan Fang, Jiarong He, and Max Gao. 2024. Opdai at semeval-2024 task 6: Small llms can accelerate hallucination detection with weakly supervised data. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 707–715, Mexico City, Mexico. Association for Computational Linguistics.

Sohan Choudhury, Priyam Saha, Subharthi Ray, Shankha Shubhra Das, and Dipankar Das. 2024. Alphaintellect at semeval-2024 task 6: Detection of hallucinations in generated text. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 939–945, Mexico City, Mexico. Association for Computational Linguistics.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Souvik Das and Rohini Srihari. 2024. Compos mentis at semeval2024 task6: A multi-faceted role-based large language model ensemble to detect hallucination. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1459–1464, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. PROTAUGMENT: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2454–2466, Online. Association for Computational Linguistics.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Pouya Fallah, Soroush Gooran, Mohammad Jafarinasab, Pouya Sadeghi, Reza Farnia, Amirreza Tarabkhah, Zeinab Sadat Taghavi, and Hossein Sameti. 2024. Slpl shroom at semeval2024 task 06 : A comprehensive study on models ability to detect hallucination. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1137–1143, Mexico City, Mexico. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Natalia Grigoriadou, Maria Lymperaiou, George Filandrianos, and Giorgos Stamou. 2024. Ails-ntua at semeval-2024 task 6: Efficient model tuning for hallucination detection and analysis. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1559–1570, Mexico City, Mexico. 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.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ioanna Iordanidou, Ioannis Maslaris, and Avi Arampatzis. 2024. Duth at semeval-2024 task 6: Comparing pre-trained models on sentence similarity evaluation for detecting of hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1053–1059, Mexico City, Mexico. 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 Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Konstantin Kobs, Jan Pfister, and Andreas Hotho. 2024. Pollice verso at semeval-2024 task 6: The roman empire strikes back. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1539–1546, Mexico City, Mexico. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Wei Liu, Wanyao Shi, Zijian Zhang, and Hui Huang. 2024. Hit-mi&t lab at semeval-2024 task 6: Deberta-based entailment model is a reliable hallucination detector. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1798–1808, Mexico City, Mexico. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Thanet Markchom, Subin Jung, and Huizhi Liang. 2024. Nu-ru at semeval-2024 task 6: Hallucination and related observable overgeneration mistake detection using hypothesis-target similarity and self-checkgpt. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 253–260, Mexico City, Mexico. 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.

Rahul Mehta, Andrew Hoblitzell, Jack O'Keefe, Hyeju Jang, and Vasudeva Varma. 2024. Halu-nlp at semeval-2024 task 6: Metacheckgpt - a multi-task hallucination detection using llm uncertainty and meta-models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 335–341, Mexico City, Mexico. Association for Computational Linguistics.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Timothee Mickus and Raúl Vázquez. 2023. Why bother with geometry? on the relevance of linear decompositions of transformer embeddings. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 127–141, Singapore. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3259–3266. AAAI Press.

Timothy Obiso, Jingxuan Tu, and James Pustejovsky. 2024. Harmonee at semeval-2024 task 6: Tuning-based approaches to hallucination recognition. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1311–1320, Mexico City, Mexico. Association for Computational Linguistics.

Ronghao Pan, José Antonio García-Díaz, Tomás Bernal-Beltrán, and Rafael Valencia-García. 2024. Umuteam at semeval-2024 task 6: Leveraging zero-shot learning for detecting hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 661–667, Mexico City, Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Zahra Rahimi, Hamidreza Amirzadeh, Alireza Sohrabi, Zeinab Taghavi, and Hossein Sameti. 2024. Hallusafe at semeval-2024 task 6: An nli-based approach to make llms safer by better detecting hallucinations and overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 139–147, Mexico City, Mexico. 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.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Elisei Sergeevich Rykov, Yana Shishkina, Ksenia Petrushina, Ksenia Titova, Sergey Petrakov, and Alexander Panchenko. 2024. Smurfcat at semeval-2024 task 6: Leveraging synthetic data for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 869–880, Mexico City, Mexico. Association for Computational Linguistics.

Béla Linus Rösener, Hong-Bo Wei, and Ilinca Vandici. 2024. Team bolaca at semeval-2024 task 6: Sentence-transformers are all you need. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1687–1690, Mexico City, Mexico. Association for Computational Linguistics.

Reza Sanayei, Abhyuday Singh, MohammadHossein Rezaei, and Steven Bethard. 2024. Maria at semeval 2024 task-6: Hallucination detection through llms, mnli, and cosine similarity. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1594–1598, Mexico City, Mexico. Association for Computational Linguistics.

Vincent Segonne and Timothee Mickus. 2023. Definition modeling : To model definitions. generating definitions with little to no semantics. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 258–266, Nancy, France. Association for Computational Linguistics.

Marco Siino. 2024. Brainllama at semeval-2024 task 6: Prompting llama to detect hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 82–87, Mexico City, Mexico. Association for Computational Linguistics.

Teemu Vahtola, Mathias Creutz, and Jrg Tiedemann. 2023. Guiding zero-shot paraphrase generation with fine-grained control tokens. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 323–337, Toronto, Canada. Association for Computational Linguistics.

Kees van Deemter. 2024. The Pitfalls of Defining Hallucination. *Computational Linguistics*, pages 1–10.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

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.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.

A Shared consciousnesses: Overview of approaches used by SHROOM teams

In Table 2, we provide a short overview of the various teams, the resources they utilized (models & datasets), as well as a short description of their approach.

B What SHROOM makes you do: Annotation guidelines

In Figure 8, we provide an exact copy of the annotation guidelines given to the annotators. These guidelines are based on five of the organizers' experience of annotating the trial set, and were provided to annotators recruited for the validation and test splits.

Annotation guidelines for SHROOM

Thank you for agreeing to participate in the SHROOM survey! The present document will provide you some general guidelines as to what is expected of you.

What are we looking for?

Our goal is to assess the **truthfulness** of sentences and documents written with artificial intelligence. In particular, we are interested in cases when the **AI outputs** are not supported by the facts. Such unsupported outputs are colloquially referred to as **'hallucinations'**.
You will be presented with a series of items. Each item contains an **AI output**, as well as one or more **References**, i.e., examples of what the model should have produced in an ideal scenario. Outputs and references can correspond to different types of sentences and paragraphs: news headlines, dictionary definitions, movie subtitles…
We ask that **you mark for every item**, whether the output contains or describes facts that are **not supported by the provided reference**. Such items are instances of hallucinations.

In other words: **hallucinations are cases where the AI output is more specific than it should be, given the available reference**.

Some Examples and counterexamples Hallucinations

Hallucinations

Below are some examples of items our research team unanimously considered as hallucinations:

Example 1

```
Reference: The worship of trees.

AI output: (uncountable) The study of trees.
```

Example 2

```
    Reference: Why is everyone laughing?
    AI output: Why is everyone okay?
```

Example 3

```
    Reference: You’re a scam artist.
    AI output: You’re not a good scam artist.
```

Not Hallucinations

And next, here are a few items that we unanimously did not consider as hallucinations:

Example 1

```
    Reference: Capable of being deployed by parachute.
    AI output: Capable of being parachuted.
```

Example 2

```
    Reference: When did you see him?
    AI output:When was the last time you saw him?
```

One last thing: Hallucinations vs. undergeneration

In some instances, the AI output can contain **less** information than the reference. We refer to such items as cases of **“undergeneration”**. These are not necessarily cases of hallucinations: as long as what is stated in the AI output is supported by the reference, such items should not be marked as hallucinations.

Here is one straightforward example:

```
Reference: I can't do it alone. You have to help me.
AI output:I can't do it alone.
```

Given that all the information present in the AI output is also found in the reference, **this should not be marked as a hallucination**.

That's it!

Figure 8: Annotation guidelines.

Team & Paper	Resources	Overview
AI Blues		*(No report)*
AILS-NTUA Grigoriadou et al. (2024)	SHROOM datasets; Vectara model.	Fine-tuned models and voting classifier.
Alejandro Mosquera	SHROOM datasets; COMET, Vectara, LaBSE, GPT35 and GPT4 models.	Ensemble of publicly available models. Logistic Regression was used as final scoring model.
AlphaIntellect Choudhury et al. (2024)	SHROOM dataset, SBERT	Fully-connected neural network classifiers with SBERT embeddings as input.
AMEX AI LABS	SHROOM datasets; Vectara and OpenChat models.	Ensemble of LLM (using Openchat) zero shot and few shot with Vectara cross encoder based scores.
Atresa		*(No report)*
Bolaca Rösener et al. (2024)	SHROOM dataset, SBERT	Logistic regression and feed-forward classifier trained on SBERT embeddings
BrainLlama Siino (2024)	LLaMA model.	Prompt-based approach with LLaMA.
BruceW		*(No report)*
Byun Byun (2024)	SHROOM dataset, data augmentation, RoBERTa	Finetuned a BERT or RoBERTa model with a softmax layer to output the probability of hallucinated text. Finetuning data is the labelled SHROOM data augmented with data points constructed by replacing words with synonyms.
CAISA		*(No report)*
Compos Mentis Das and Srihari (2024)	HalluEval dataset; Mistral 7B instruct model.	Ensemble of several role-based LLMs, which were either fine-tuned on hallucination data or role-based prompting.
daixiang		*(No report)*
deema		*(No report)*
DeepPavlov Belikova and Kosenko (2024)	SHROOM dataset; OpenChat, DeBERTa, RoBERTa and T5 models.	Ensemble of several pretrained Transformer-based models to get features for validation and test data of SHROOM dataset and trained a boosting-based meta-model on top.
DUTh Iordanidou et al. (2024)	SHROOM, LaBSE, T5, DistilUSE	Using pre-trained LLMs and classifiers
HalluSafe Rahimi et al. (2024)	SHROOM, labeled 3000 samples of the training data	Fine-tuned a DeBERTa-v3-large
Halu-NLP Mehta et al. (2024)	SHROOM datasets; GPT, SelfCheckGPT and Vectara models.	Prompts and GroupCheckGPT. NB: due to a team name change, this team is also referred to as GroupCheckGPT by some participants.
HaRMoNEE Obiso et al. (2024)	SHROOM, SNLI, MNLI and PAWS datasets; Vectara and GPT4 models.	Highest results obtained with zero-shot prompting in the model-aware track; pretraining on NLI and PAWS followed by finetuning on the model-agnostic track.
HIT-MI&T Lab Liu et al. (2024)	SHROOM with training dataset labeled using GPT-4; DeBERTaV3, InternLM2, SBERT, and UniEval.	Fine-tune the DeBERTaV3 and InternLM2 models, and call the SBERT and UniEval models to select the optimal threshold usinf SHROOM & syntheticaly labeled data. The system obtians the final results by combining the prediction results of each model.
IRIT-Berger-Levrault Bendahman et al. (2024)	SHROOM datasets; Sentence-t5, BGE, e5 models.	Computes the cosine similarity of sentence embeddings and classify based on an empirical threshold value.
Maha Bhaashya Bhamidipati et al. (2024)	DeBERTa models.	Zero shot inference, pretrained cross encoder model
MALTO Borra et al. (2024)	SHROOM model-agnostic dataset, DeBERTa pretrained and finetuned on MNLI, SOLAR-10.7B quantized from TheBloke (for synthetic data generation)	Encoder and classifier, fine-tuned in various ways (including with synthetic data)
MARiA Sanayei et al. (2024)	SHROOM dataset, SBERT, bart-large-mnli, Mixtral	Three approaches: (1) Cosine similarity of SBERT embeddings between source-hypothesis and source-target pairs; (2) NLI classification using bart-large-mnli model; and (3) Mixtral prompting. Only the Mixtral results were submitted.
Noot Noot Bahad et al. (2024)	SHROOM dataset; Mixtral and RoBERTa models.	Mixtral prompting and RoBERTa finetuning.
NU-RU Markchom et al. (2024)	SHROOM, GPT-3.5, Sentence Transformers	Tried two approaches: (1) hypothesis-target cosine similarity, using a threshold value to determine whether the hypothesis is a hallucation. (2) SelfCheckGPT with a customized prompt for each NLG task, designed to assess its coherence with the provided source and target. Each prompt is iterated through the GPT-3.5 model five times, and the final label is determined by the majority response.
octavianB Brodoceanu (2024)	RoBERTa	Used a pretrained model (roberta-large-openai-detector) that has been trained to distinguish between text generated by LLMs and text written by humans.
OPDAI Chen et al. (2024)	SHROOM, Mistral-7B-Instruct-v0.2, self constructed training data	Supervised fine-tuning over synthetically constructed weakly supervised training data.
Pollice Verso Kobs et al. (2024)	Mistral2, LLaMa2, Phi2 and Zephyr models; uses SHROOM train set for prompt optimization.	Ensembling over the output logits of prompt-based LLMs (mistral, llama etc) after automatically optimizing their prompts ("OPRO").
SHROOM-INDElab Allen et al. (2024)	SHROOM dataset; GPT 3.5 and GPT 4 models.	In-context learning with role-play and automatic prompt generation in a few-shot classfier, using a closed-source LLM.
SibNN	SHROOM datasets; XLM-RoBERTa model.	Fine-tunes a self-adaptive hierarchical variant of XLM-RoBERTa-XL twice: first as an embedder (in a few-shot mode), then as a binary classifier. More details at https://huggingface.co/bond005/xlm-roberta-xl-hallucination-detector.
silk_road	SHROOM datasets; Vectara model.	Fine-tunes an off-the-shelf Cross-Encoder hallucination evaluation model.
Skoltech		*(No report)*
SLPL SHROOM Fallah et al. (2024)	SHROOM datasets; LaBSE, DeBERTa, Zephyr, Mistral and Llama2 models.	Using two LLMs to classify and explain their decision and another LLM to judge and decide based on those explanations.
SmurfCat Rykov et al. (2024)	SHROOM (synthetically augmented), QQP and PAWS datasets; E5, T5, Vectara models.	Fine-tuning of e5-mistral-7b-instruct using synthetic data collected with LLaMA2-7B adapters trained to produce data with and without hallucinations. However, there are two other systems: one works as a voting ensemble of multiple LLMs, and another uses the Mutual Implication Score architecture.
Team CentreBack	SHROOM dataset; DeBERTa model.	Uses an off-the-shelf library (SelfCheckGPT's SelfCheckNLI function) to calculate contradiction scores on a small labeled test set and then defined a threshold for hallucination.
TU Wien Arzt et al. (2024)	SHROOM dataset; Vectara model.	Model-aware track best submissions uses a Vectara hallucination detection model finetuned on the validation set. The best model-agnostic track submission is a meta-model that utilizes linear regression and is trained on features that correspond to probabilities predicted by individual systems we implemented.
UCC-NLP	SHROOM dataset; GPT-3.5 and Vectara models.	Uses BertScore and GPT-3.5 to create synthetic labels and fine-tune a Vectara LLM.
UMUTeam Pan et al. (2024)	SHROOM dataset; TULU-DPO model.	Zero-shot approach
ustc_xsong		*(No report)*
zhuming		*(No report)*
0x.Yuan	Mistral, Mixtral, LLaMA, Falcon, WizardLM and Capybara models.	Zero-shot prompt engineering. Expects most LLMs will have different hallucination patterns, and tests whether ensembling can mitigate this.

Table 2: Participating teams and their respective works.