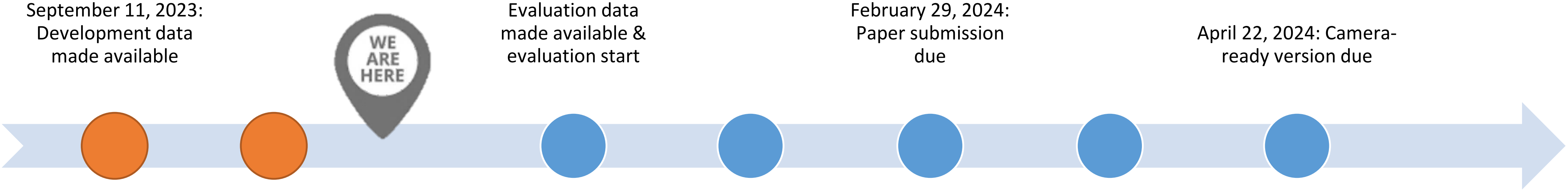
Hallucination challenge

María Navas Loro

mnavas@fi.upm.es

Links de referencia / schedule

- https://codalab.lisn.upsaclay.fr/competitions/15726
- https://semeval.github.io/SemEval2024/



September 11, 2023: Development data made available

September 22, 2023: Unlabelled training data made available

January 10, 2024: Evaluation data made available & evaluation start

January 31, 2024: Evaluation end

February 29, 2024: Paper submission due

April 1, 2024: Notification to authors

April 22, 2024: Camera-ready version due

GRAN OPORTUNIDAD

SemEval workshop: June 16–21, 2024 (co-located with NAACL 2024 in Mexico City, Mexico).

Esto es lo que pasa cuando le dices a DALL-E 3

“Una conferencia en México”

Enfoque “humano”

- Mi propuesta: que vaya quien más trabaje/mejores resultados obtenga, sea post/predoc, becario…
- Si hay varias líneas interesantes, puede que incluso puedan ir varios subgrupos, uno por paper ☺
- Obviamente, los postdoc tenemos menos tiempo, así que será principalmente “apoyo”: ayudar con la escritura, estado del arte, esbozar las líneas de acción… el cacharreo real (lo diver ☹) será más becarios/predocs, pero obviamente esto cuenta de cara a orden, etc…
- … porque aquí el que no trabaje no figura. Nada de aparecer el ultimo día, o decir que se es supervisor. Aquí firma el que trabaja.

Sobre el challenge: evaluación

- Submissions will be divided into two tracks:
  - a model-aware track, where we provide a checkpoint to a model publically available on HuggingFace for every datapoint considered
  - a model-agnostic track where we do not. We highly encourage participants to make use of model checkpoints in creative ways.
- For both tracks, all participants' submissions will be evaluated using two criteria:
  - the **accuracy** that the system reached on the binary classification
  - the **Spearman correlation** of the systems' output probabilities with the proportion of the annotators marking the item as overgenerating

Instructions

Anyone wishing to participate in the task is welcome! Participants will have to:

- Submit at least once during the evaluation phase next January;
- Write a system description paper;
- **Review other system description papers (max. 2).**

About the data

The task consists in a binary classification, where participants are asked to determine whether a given production from an NLP model constitutes a hallucination

https://helsinki-nlp.github.io/shroom/

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

About the data

- Ya descargado, descomprimido y en csv (más legible) en:

https://delicias.dia.fi.upm.es/nextcloud/index.php/s/y9wqgcexPDSDdcM

```
|   JSON2CSV.txt
|
├──SHROOM_dev-v1
|      README-v1.txt
|      val.model-agnostic.json
|      val.model-aware.json
|
├──SHROOM_trial-v1.1
|      README-v1.1.txt
|      trial-v1.json
|
└──SHROOM_unlabeled-training-data-v1
      train.model-agnostic.json
      train.model-aware.json
```

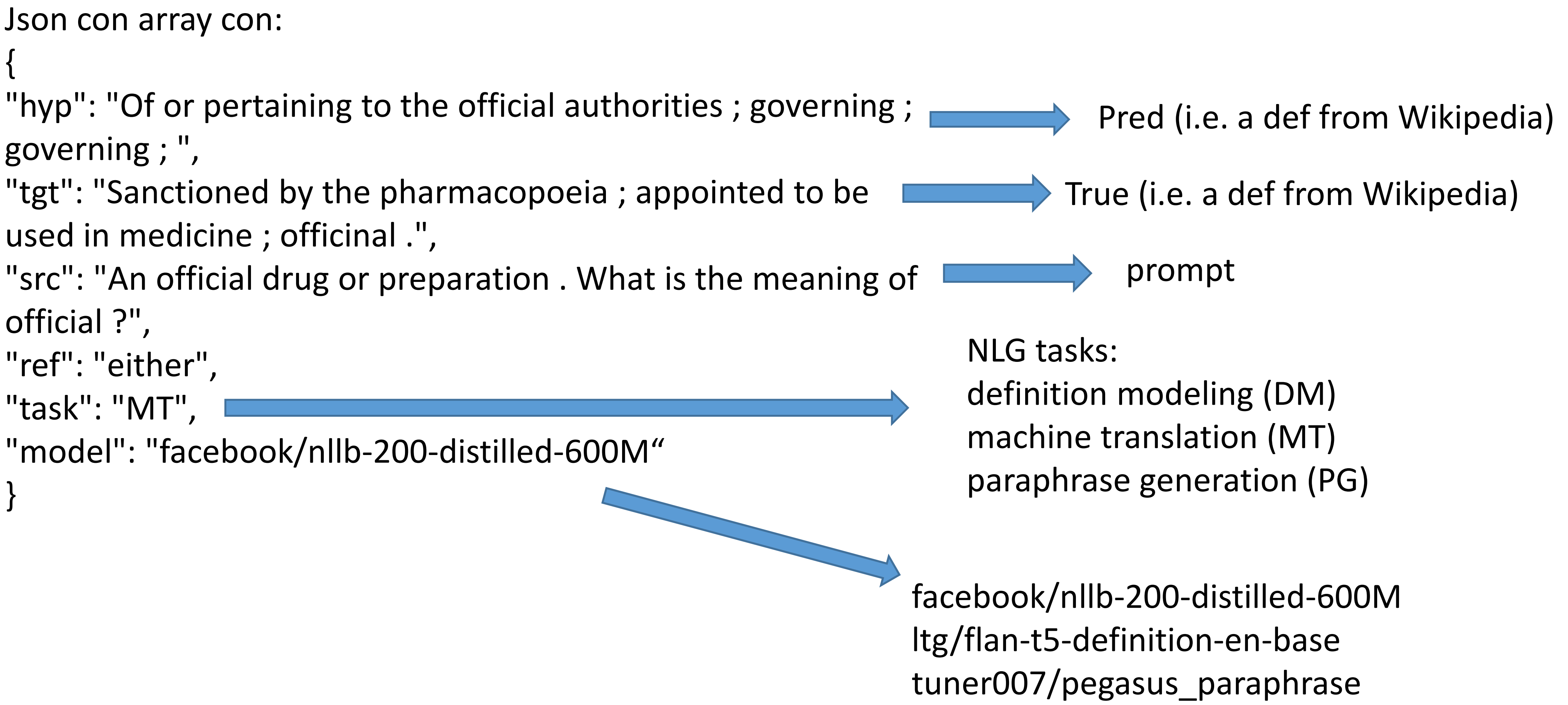
Info available

- a task (`task`), indicating what objective the model was optimized for;
- a source (`src`), the input passed to the models for generation;
- a target (`tgt`), the intended reference "gold" text that the model ought to generate;
- a hypothesis (`hyp`), the actual model production;
- a set of per annotator labels (`labels`), indicating whether each individual annotator thought this datapoint constituted a hallucination or not;
- a majority-based gold-label (`label`), based on the previous per-annotator labels;
- a probability assigned to this datapoint being a hallucination (`p(Hallucination)`), corresponding to the proportion of annotators who considered this specific datapoint to be a hallucination.

unlabeled-training-data-v1

Train dataset (model-aware)

30k



Json con array con:

```
{
"hyp": "Of or pertaining to the official authorities ; governing ; governing ; ",
"tgt": "Sanctioned by the pharmacopoeia ; appointed to be used in medicine ; officinal .",
"src": "An official drug or preparation . What is the meaning of official ?",
"ref": "either",
"task": "MT",
"model": "facebook/nllb-200-distilled-600M“
}
```

Pred (i.e. a def from Wikipedia)

True (i.e. a def from Wikipedia)

prompt

NLG tasks:
definition modeling (DM)
machine translation (MT)
paraphrase generation (PG)

facebook/nllb-200-distilled-600M
ltg/flan-t5-definition-en-base
tuner007/pegasus_paraphrase

unlabeled-training-data-v1

Train dataset (model-agnostic)

30k

Json con array con:

```
{"hyp": "Don't worry, it's only temporary.",
"tgt": "Don't worry. It's only temporary.",
 "src": "\u041d\u0435
\u0432\u043e\u043b\u043d\u0443\u0439\u0441\u044f.
\u042d\u0442\u043e
\u0442\u043e\u043b\u044c\u043a\u043e
\u0432\u0440\u0435\u043c\u0435\u043d\u043d\u043e.",
"ref": "either",
"task": "MT",
"model": "“
}
```

Either, src or tgt

Siempre ruso en MT!!!
En el csv se ve
“Не волнуйся. Это только
временно.”

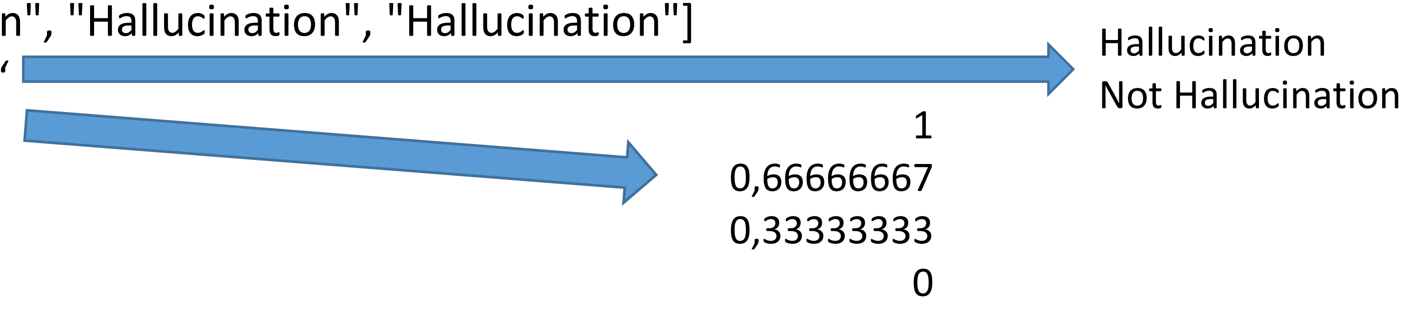
trial-v1

Trial dataset

80

Json con array con:

```
{
"hyp": "A district of Kowloon, China.“
"ref": "tgt“
"src": "The City <define> Chiuchow </define> is Kowloon 's other top restaurant and is famous for its goose dishes and other specialties from the Chiuchow region ( you may also wish to try the beef satay done in a creamy sauce ) .“
"tgt": "The Chaoshan region where the Teochew dialect is spoken.“
"model": "“
"task": "DM“
"labels": ["Hallucination", "Hallucination", "Hallucination"]
"label": "Hallucination“
"p(Hallucination)": 1.0
}
```



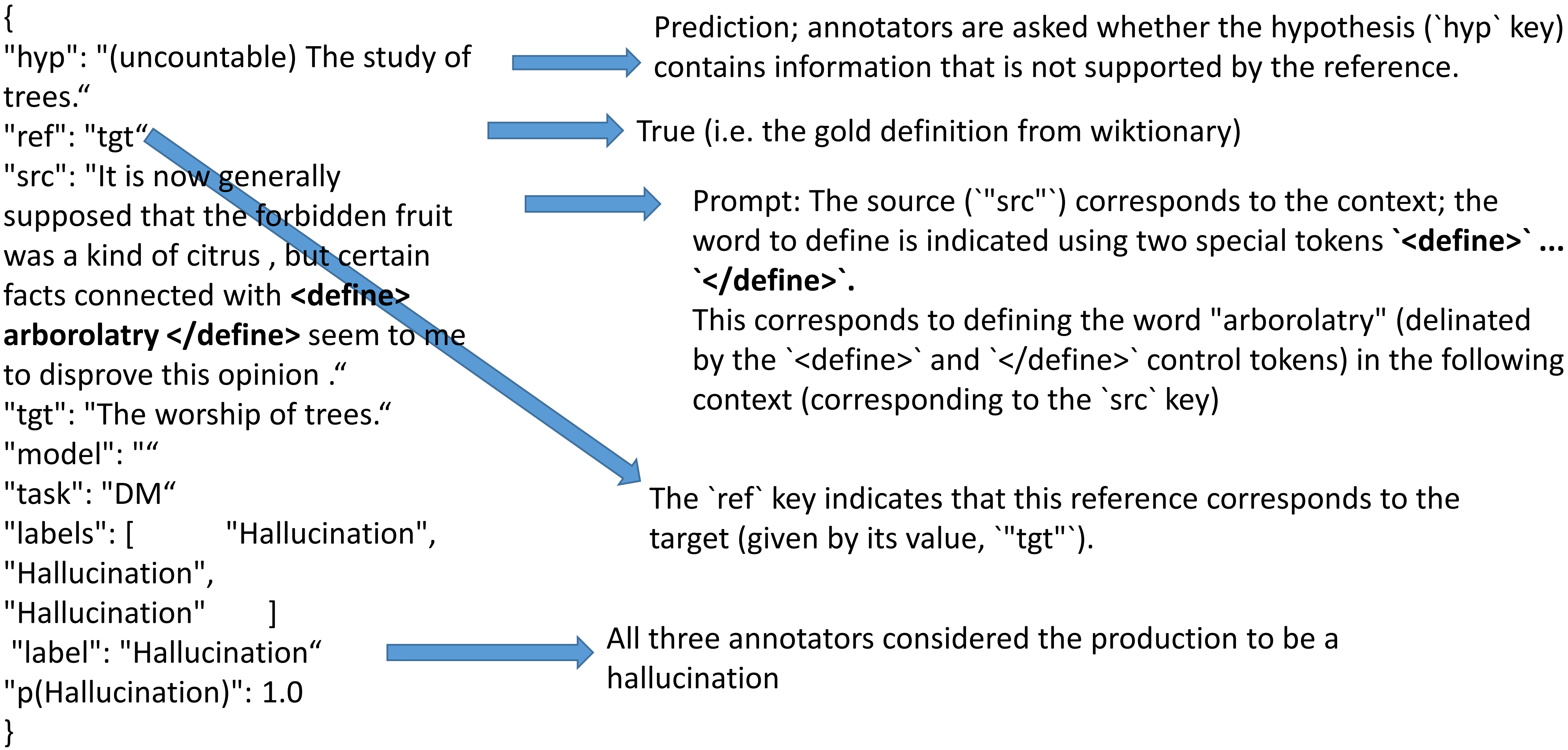
Hallucination
Not Hallucination

1
0,66666667
0,33333333
0

Trial readme example (DM)

Definition Modeling

README TRIAL



```
{
"hyp": "(uncountable) The study of
trees.“
"ref": "tgt“
"src": "It is now generally
supposed that the forbidden fruit
was a kind of citrus , but certain
facts connected with <define>
arborolatry </define> seem to me
to disprove this opinion .“
"tgt": "The worship of trees.“
"model": "“
"task": "DM“
"labels": [          "Hallucination",
"Hallucination",
"Hallucination"       ]
 "label": "Hallucination“
"p(Hallucination)": 1.0
}
```

Prediction; annotators are asked whether the hypothesis (`hyp` key) contains information that is not supported by the reference.

True (i.e. the gold definition from wiktionary)

Prompt: The source (`"src"`) corresponds to the context; the word to define is indicated using two special tokens **`<define>` ... `</define>`.**
This corresponds to defining the word "arborolatry" (delinated by the `<define>` and `</define>` control tokens) in the following context (corresponding to the `src` key)

The `ref` key indicates that this reference corresponds to the target (given by its value, `"tgt"`).

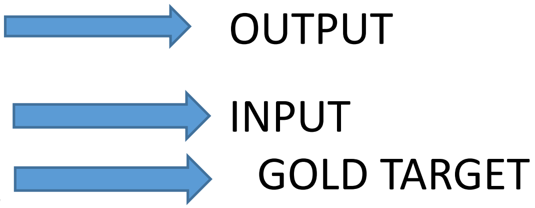
All three annotators considered the production to be a hallucination

README TRIAL

Trial readme example (PG)

Paraphrase Generation

```
{
"hyp": "When did you see him?",
"ref": "either",
"src": "When\u2019d you last see him?",
"tgt": "When was the last time you saw him?",
"model": "tuner007/pegasus_paraphrase",
"task": "PG",
"labels": [          "Not Hallucination",
"Not Hallucination",          "Not Hallucination"
],
"label": "Not Hallucination",
"p(Hallucination)": 0.0
}
```



OUTPUT

INPUT

GOLD TARGET

For PG datapoints, we also indicate the huggingface model that was used to generate the hypothesis

Trial readme example (MT)

Machine Translation

```
{
"hyp": "I have nothing to do with it.",
"ref": "either",
"src": "J'en ai rien \u00e0 secouer.",
"tgt": "I don't give a shit about it.",
"model": "",
"task": “MT",
"labels": ["Hallucination",           "Not
Hallucination",           "Hallucination"],
"label": "Hallucination",
"p(Hallucination)": 0.6666666666666666
}
```

A Traducir: SRC
Salida: HYP
TGT es lo correcto

README TRIAL

The trial set covers datapoints from definition modeling (DM), machine translation (MT) and paraphrase generation (PG). **All other sets should also include text simplification (TS) datapoints**.

Furthermore

- **The train set will not contain manual annotations.**
- The validation and evaluation sets will involve five annotators per datapoint.

Val folder

- Sólo los tres tipos que conocemos (MT, PG, DM)
- 5 anotadores
- Agnostic: 218 hallucination/281 no hallucination **499**
- Aware: 206 hallucination/295 no hallucination **501**
- Mismos modelos que antes

Enfoque tecnológico

¿Qué líneas sugerís?

- Distinto enfoque por modelo? (nada garantiza nuevos en test)
- Distinto enfoque por tarea?
- Bases de datos externas? KG, Wikidata, Wikipedia…

¿Organizarse en subgrupos independientes?

- Por tarea
- Por disponibilidad temporal
- Por modelo