From COMP0087 to the EMNLP Conference

A journey to EMNLP

...writing the paper

Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models

Kenza Benkirane[1*], Laura Gongas[1*],
Shahar Pelles[1], Naomi Fuchs[1], Joshua Darmon[1],
Pontus Stenetorp[1], David Ifeoluwa Adelani[2,3], Eduardo Sanchez[1,4]

[1]UCL, [2]Mila, Quebec AI institute, [3]McGill University, [4]Meta
*: Equal contributions

Empirical Methods for Natural Language Processing (EMNLP) 2024

Objectives

- Give all the resources, context and tips for your coursework submission to be at the level of a paper
- Introduction to the research framework of the publication of a research paper
- Give insights and learning material for embedding spaces and frameworks for LLM evaluation

Who are we?

Laura Gongas

- From Medellin, Colombia
- BSc Biomedical Engineering from Universidad de Las Andes
- AI Engineer and Product Manager for 4 years
- **MSc AI for Biomedicine and Healthcare**
- Now: AI Lead Engineer @MedTech Startup

Kenza Benkirane

- From Casablanca, Morocco
- MSc Biomedical Engineering from INSA Lyon (France)
- Digital Health Consultant for 2 years
- **MSc AI for Biomedicine and Healthcare**
- Now: AI Lead @MedTech Startup

Table of content

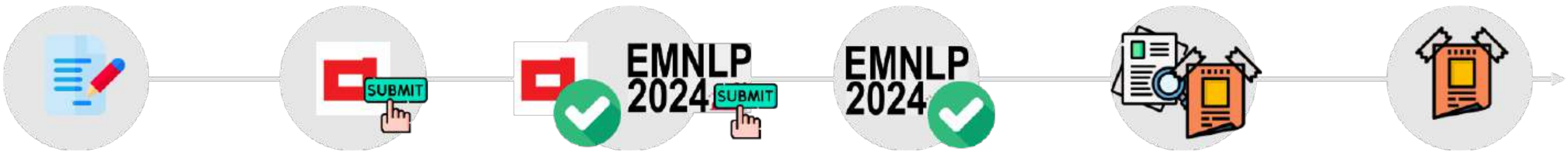
1. DEFINING THE PROJECT

Ideation phase

2. SOLVING THE RESEARCH QUESTION

Implementation phase

3. WRITING THE RESEARCH



4. PRACTICAL TIPS

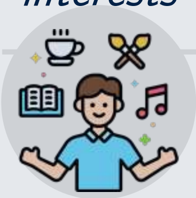
What we've learned on the way

Learning

1. DEFINING THE PROJECT

1. DEFINING THE PROJECT

Define group's interests

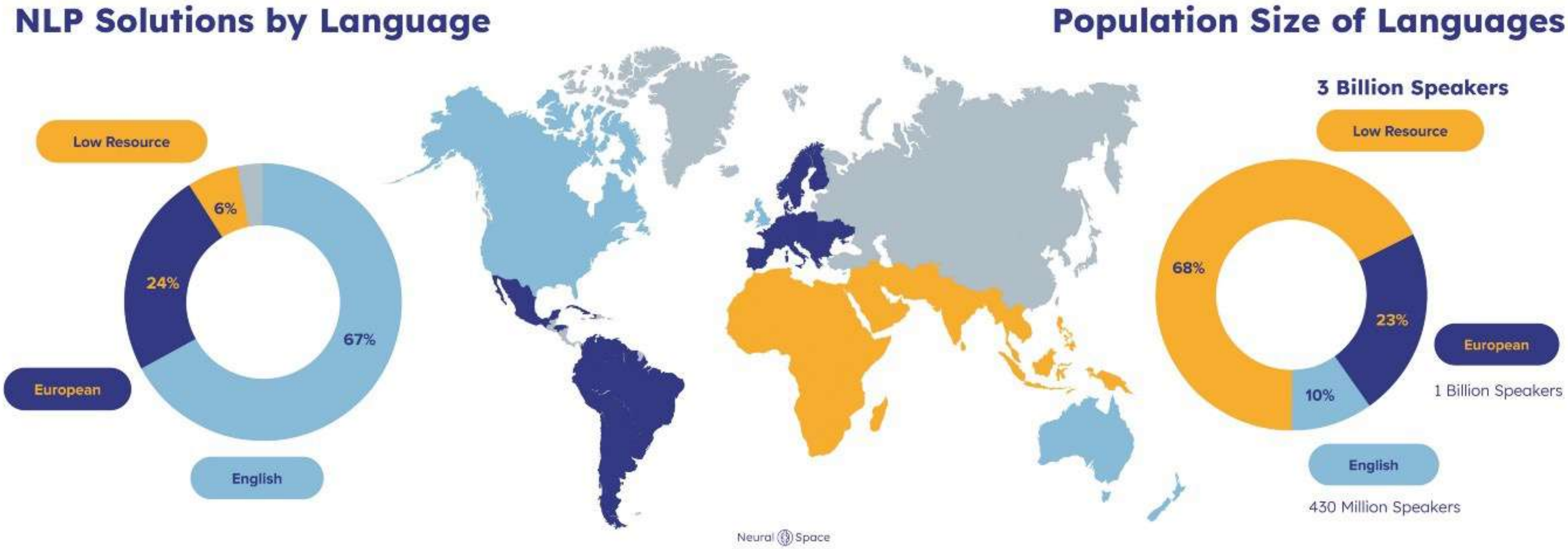


Identify limitations

Literature review

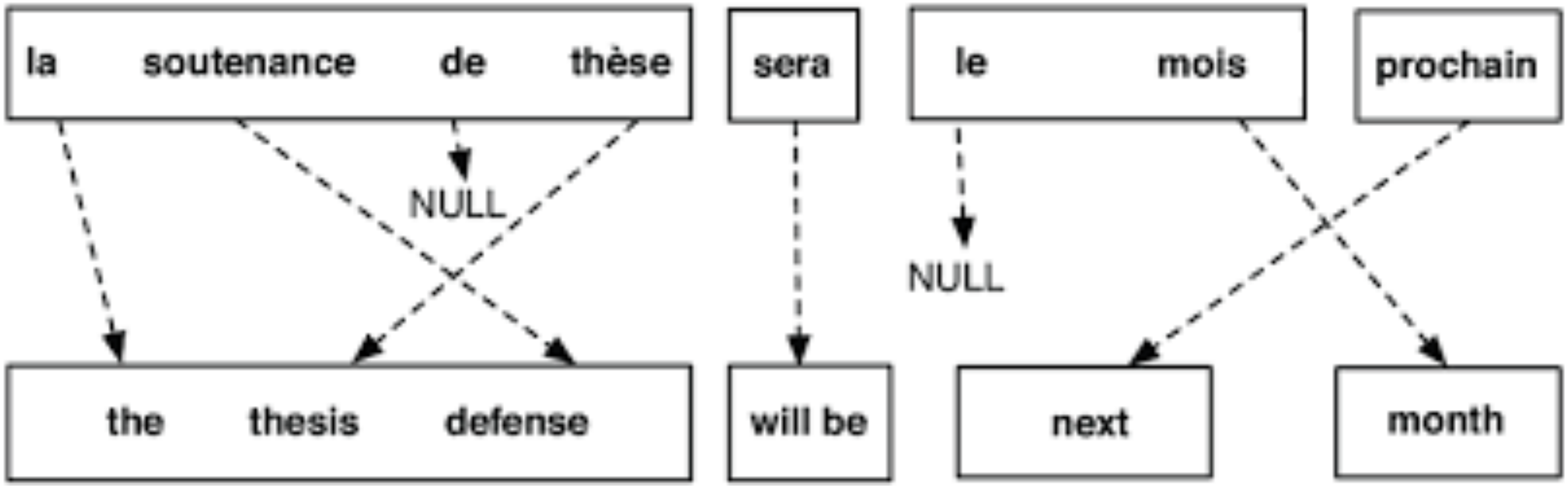
The Research Question

Key Term: Low Resource Languages



Medium | Low Resource Language - what does it mean?

Key Term: Machine Translation



Emu: Enhancing Multilingual Sentence Embeddings with Semantic Similarity

Key Term: Hallucination

A refund of 194.73EUR was done on January 12th.

Se realizó el reembolso por 19.73 en Enero 12.

Do not administer insulin if the patient’s blood glucose level is below 70 mg/dL

Administrer de l'insuline si la glycémie du patient est inférieure à 70 mg/dL

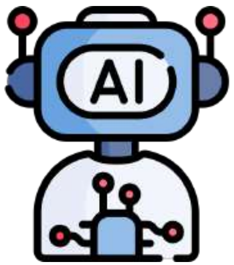
Hallucinations in machine translation are translations that contain information completely un- related to the input.

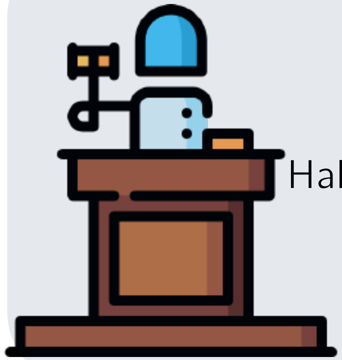
HalOmi: a manually annotated benchmark for multilingual hallucination and omission detection in Machine Translation

1. DEFINING THE PROJECT

HalOmi

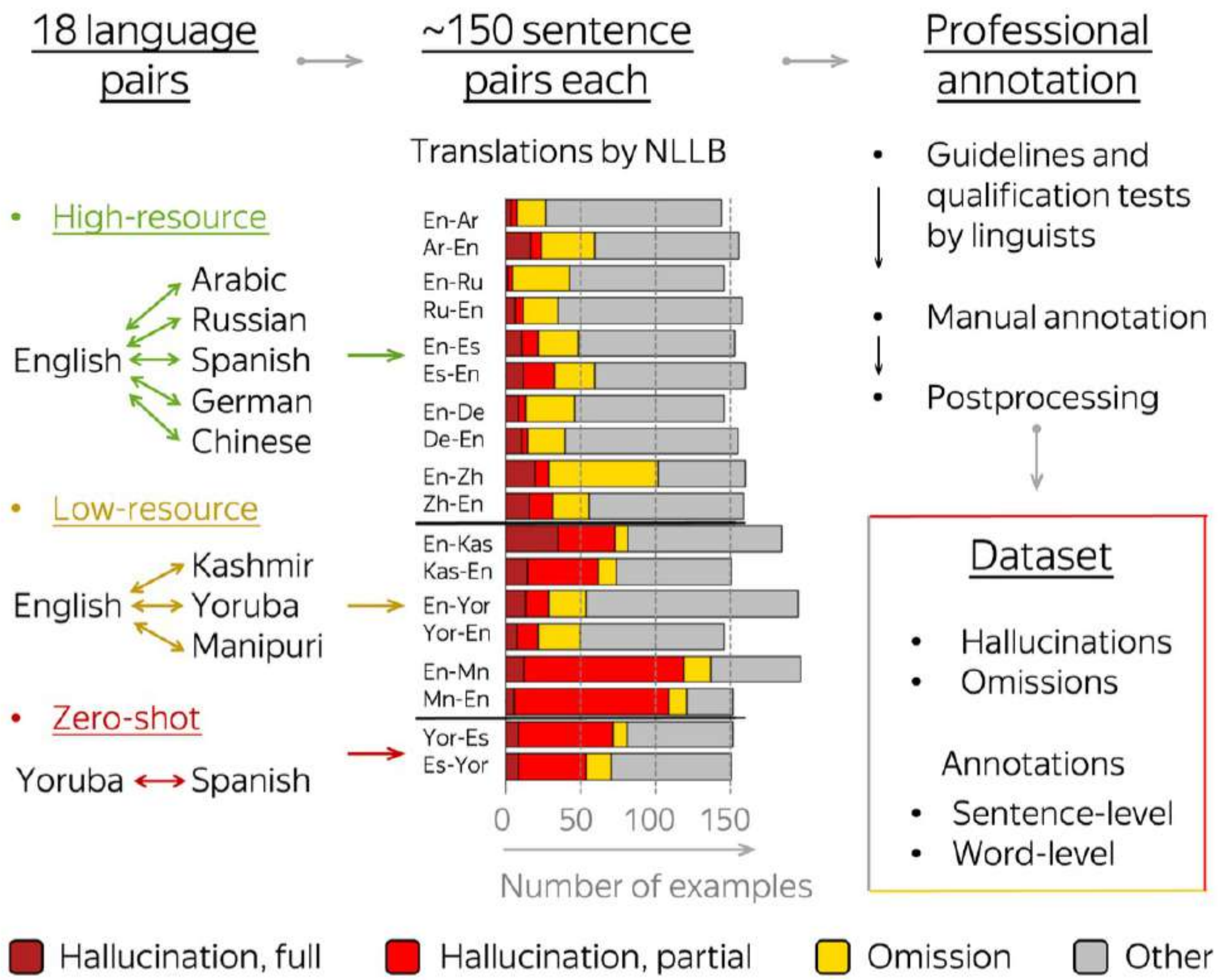
A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation

Source text → 

→ Translated text →  Hallucination in translated text?

Pro-tip: ensure you have the data or model access

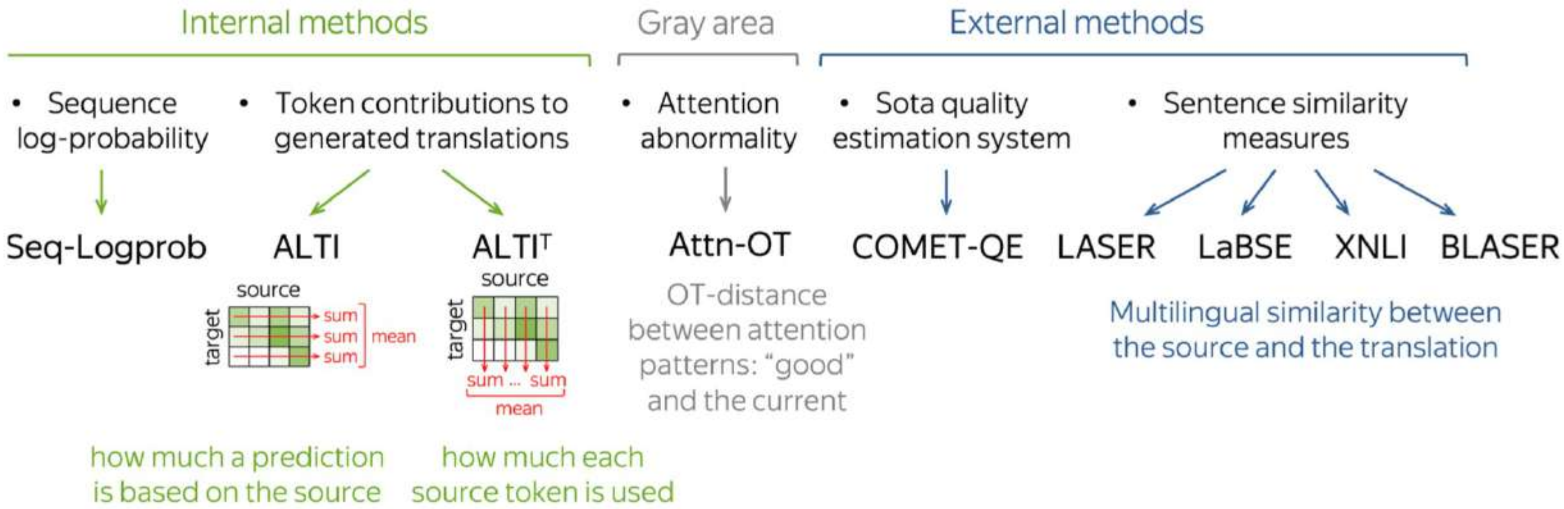
1. DEFINING THE PROJECT



Context and dataset:

- HalOmi is a hallucination and omission dataset for 18 sentence pairs, including three **low-resource languages.**
- The benchmark presents **BLASER** as the best method for hallucination ranking, for both high and low resource languages

1. DEFINING THE PROJECT



Internal

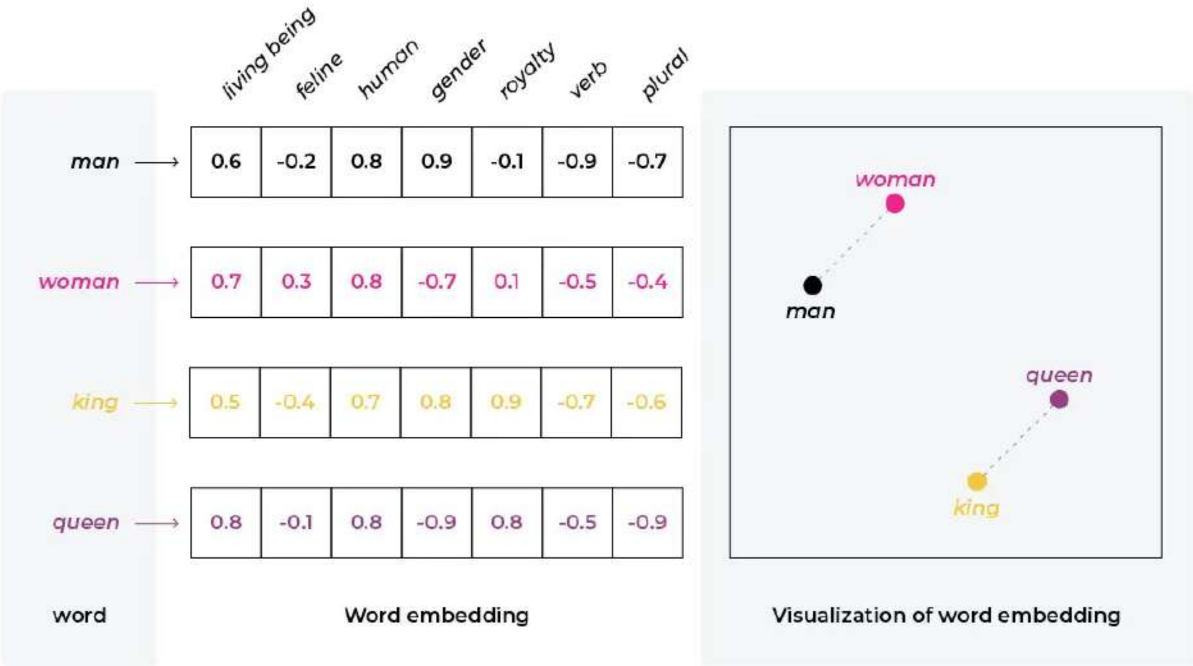
→ **Seq-logprob**: if a model is not confident in the translation then it might be a hallucination

External

→ **Sentence similarity measures**: how similar are the source text and machine translated text?

External

→ **Sentence similarity measures**: how similar are the source text and machine translated text?

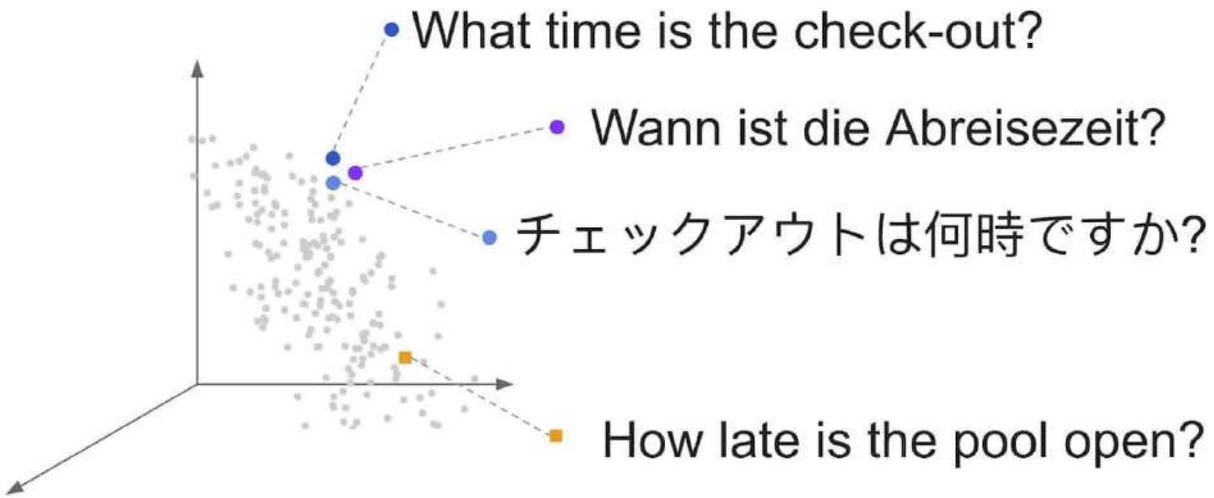


Arize AI | Embeddings, how to compute them

Step 1:
Map the source and translation sentences to an embedding space.

External

→ **Sentence similarity measures**: how similar are the source text and machine translated text?



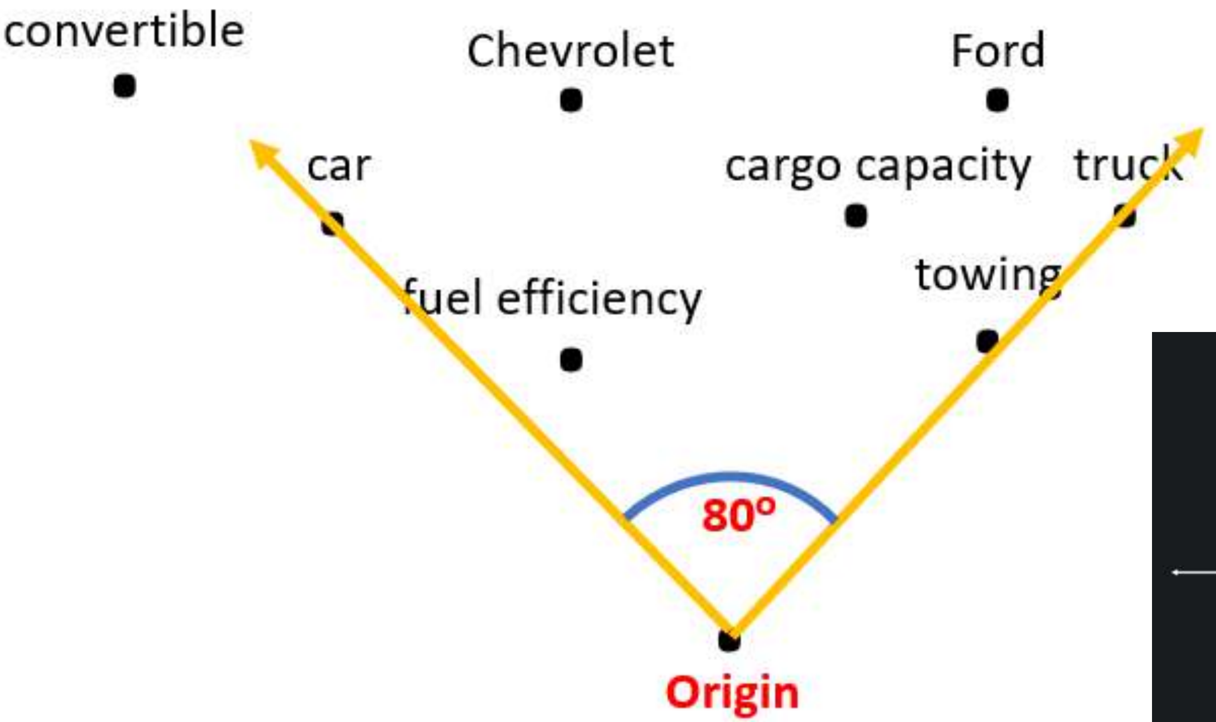
Shah, Kashif. (2012). Model adaptation techniques in machine translation.

Step 2:
Measure the distance between the embedded source and translated sentences.

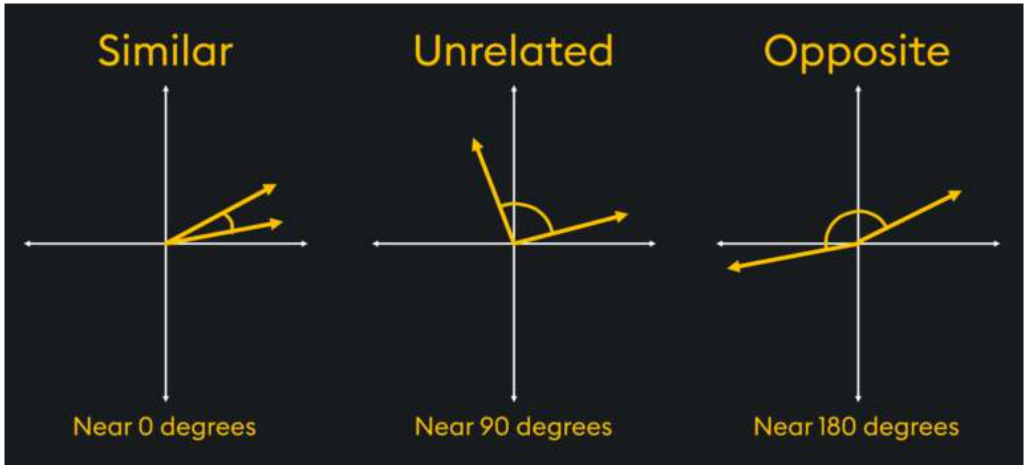
Distance measures:

- → Cosine similarity
- → **BLASER**: neural network trained to calculate the distance between embedding points

➔ **Cosine similarity**: mapping vectors within a graphical context and then calculating the angle between these vectors, ultimately delivering the cosine of that angle as a measure of their similarity.



Medium.| Distances in Machine Learning



Hallucinations (among all translations)

		Internal				External				
		Seq-Logprob	ALTI	ALTI^T	Attn-OT	COMET-QE	LaBSE	LASER	XNLI	BLASER-QE
High-resource	English-Arabic	0.89	0.78	0.54	0.36	0.84	0.84	0.89	0.83	0.90
	Arabic-English	0.83	0.75	0.72	0.51	0.80	0.88	0.83	0.87	0.94
	English-Russian	0.95	0.36	0.37	0.61	0.99	0.86	0.80	0.82	0.89
	Russian-English	0.86	0.76	0.56	0.55	0.78	0.83	0.85	0.77	0.92
	English-Spanish	0.87	0.85	0.69	0.59	0.85	0.88	0.84	0.62	0.85
	Spanish-English	0.92	0.89	0.67	0.37	0.86	0.94	0.88	0.89	0.87
	English-German	0.85	0.97	0.69	0.55	0.88	0.97	0.88	0.94	0.87
	German-English	0.90	0.80	0.59	0.65	0.92	0.95	0.86	0.88	0.97
	English-Chinese	0.88	0.82	0.47	0.60	0.84	0.86	0.79	0.84	0.78
	Chinese-English	0.89	0.88	0.65	0.46	0.89	0.88	0.82	0.78	0.87
	High-resource mean	0.88	0.79	0.59	0.52	0.86	0.89	0.84	0.81	0.89
Low-resource	English-Kashmir	0.68	0.71	0.74	0.54	0.64	0.76	0.70	0.68	0.81
	Kashmir-English	0.59	0.67	0.65	0.65	0.69	0.58	0.59	0.47	0.73
	English-Yoruba	0.68	0.81	0.49	0.54	0.57	0.80	0.76	0.35	0.83
	Yoruba-English	0.70	0.64	0.49	0.57	0.58	0.57	0.59	0.46	0.78
	English-Manipuri	0.77	0.74	0.59	0.65	0.68	0.56	0.57	0.44	0.79
	Manipuri-English	0.78	0.72	0.54	0.43	0.58	0.66	0.69	0.47	0.80
	Low-resource mean	0.70	0.72	0.58	0.56	0.62	0.66	0.65	0.48	0.79
Zero-shot	Yoruba-Spanish	0.60	0.65	0.47	0.44	0.54	0.56	0.60	0.51	0.58
	Spanish-Yoruba	0.61	0.66	0.52	0.55	0.55	0.66	0.64	0.52	0.68
	Overall mean	0.79	0.75	0.58	0.53	0.75	0.78	0.75	0.67	0.83

Omissions (among non-hallucinations)

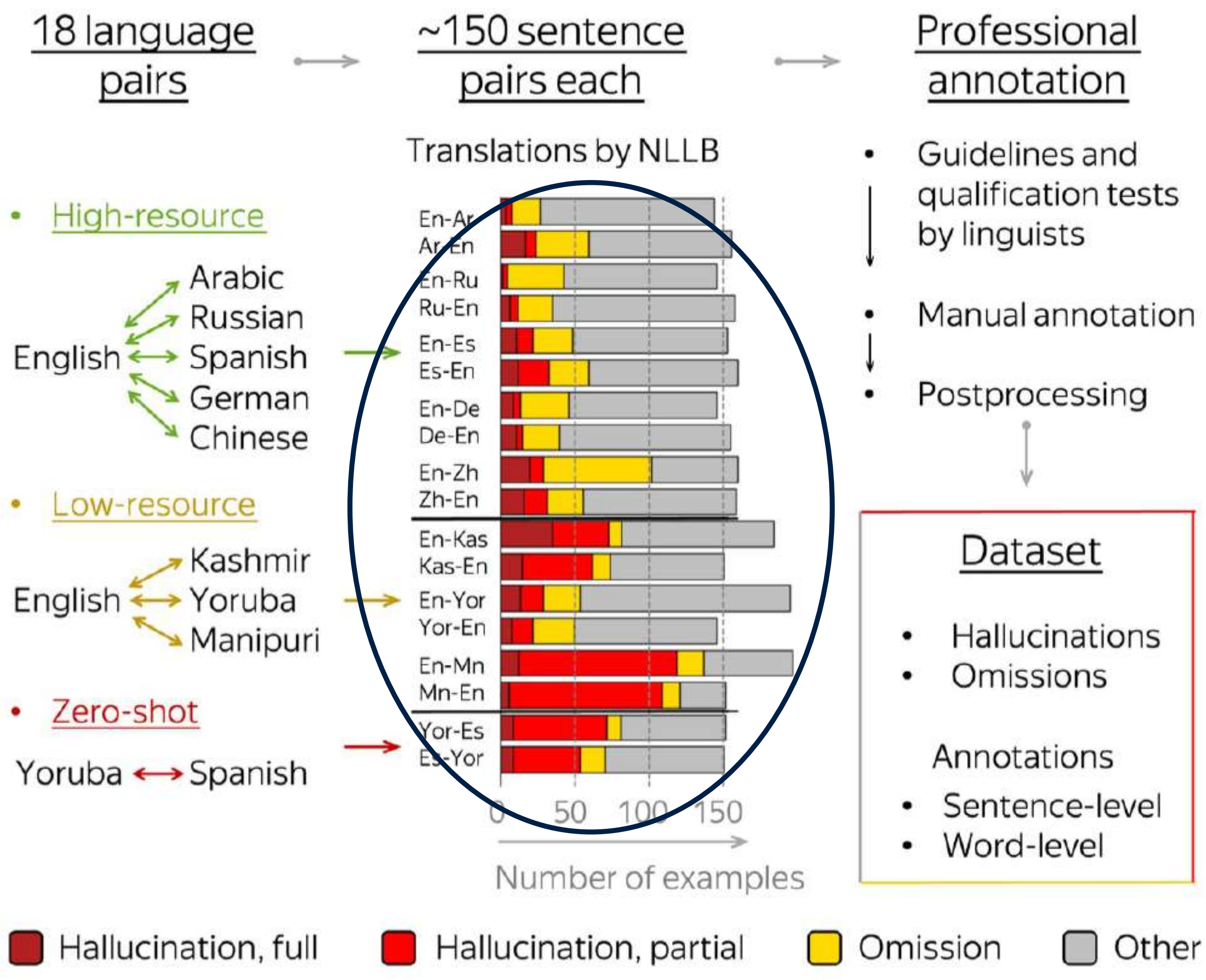
		Internal				External				
		Seq-Logprob	ALTI	ALTI^T	Attn-OT	COMET-QE	LaBSE	LASER	XNLI	BLASER-QE
High-resource	English-Arabic	0.63	0.44	0.76	0.64	0.61	0.79	0.77	0.76	0.85
	Arabic-English	0.57	0.49	0.82	0.72	0.63	0.83	0.62	0.86	0.77
	English-Russian	0.53	0.56	0.79	0.75	0.53	0.85	0.74	0.80	0.84
	Russian-English	0.59	0.52	0.86	0.76	0.57	0.76	0.75	0.80	0.80
	English-Spanish	0.39	0.31	0.89	0.89	0.65	0.86	0.64	0.85	0.80
	Spanish-English	0.61	0.50	0.59	0.66	0.69	0.73	0.62	0.70	0.71
	English-German	0.50	0.46	0.77	0.74	0.54	0.66	0.57	0.56	0.85
	German-English	0.48	0.38	0.82	0.83	0.62	0.70	0.64	0.77	0.70
	English-Chinese	0.60	0.51	0.73	0.67	0.57	0.69	0.73	0.63	0.88
	Chinese-English	0.73	0.61	0.77	0.64	0.66	0.75	0.74	0.79	0.73
	High-resource mean	0.56	0.48	0.78	0.73	0.61	0.76	0.68	0.75	0.79
Low-resource	English-Kashmir	0.50	0.52	0.90	0.81	0.80	0.76	0.55	0.60	0.77
	Kashmir-English	0.36	0.50	0.64	0.63	0.61	0.59	0.44	0.53	0.45
	English-Yoruba	0.45	0.52	0.80	0.73	0.51	0.82	0.68	0.64	0.80
	Yoruba-English	0.42	0.33	0.68	0.63	0.51	0.52	0.45	0.52	0.74
	English-Manipuri	0.77	0.50	0.70	0.72	0.59	0.65	0.41	0.54	0.67
	Manipuri-English	0.68	0.60	0.65	0.55	0.57	0.66	0.76	0.56	0.56
	Low-resource mean	0.53	0.50	0.73	0.68	0.60	0.67	0.55	0.57	0.66
Zero-shot	Yoruba-Spanish	0.62	0.47	0.85	0.69	0.40	0.62	0.63	0.31	0.66
	Spanish-Yoruba	0.68	0.50	0.83	0.60	0.51	0.69	0.77	0.42	0.67
	Overall mean	0.56	0.48	0.77	0.70	0.59	0.72	0.64	0.65	0.74

Figure 5: Results for sentence-level detection of hallucinations (left) and omissions (right).

⚠️

Limitation Identification: Suboptimal performance for low resource languages with SOTA.

1. DEFINING THE PROJECT



⚠

Limitation Identification:
Highly imbalanced dataset

1. DEFINING THE PROJECT

2. SOLVING THE RESEARCH QUESTION

How can we improve the hallucination detection for machine translation in low-resource languages?

In other words:
Beat BLASER (SOTA)

Major challenges

- ✔ Constraints on computational resources and access
- ✔ No access to training data

We can’t train a neural network stronger than BLASER.

Define group's interests

Identify limitations

Literature review

Set an objective

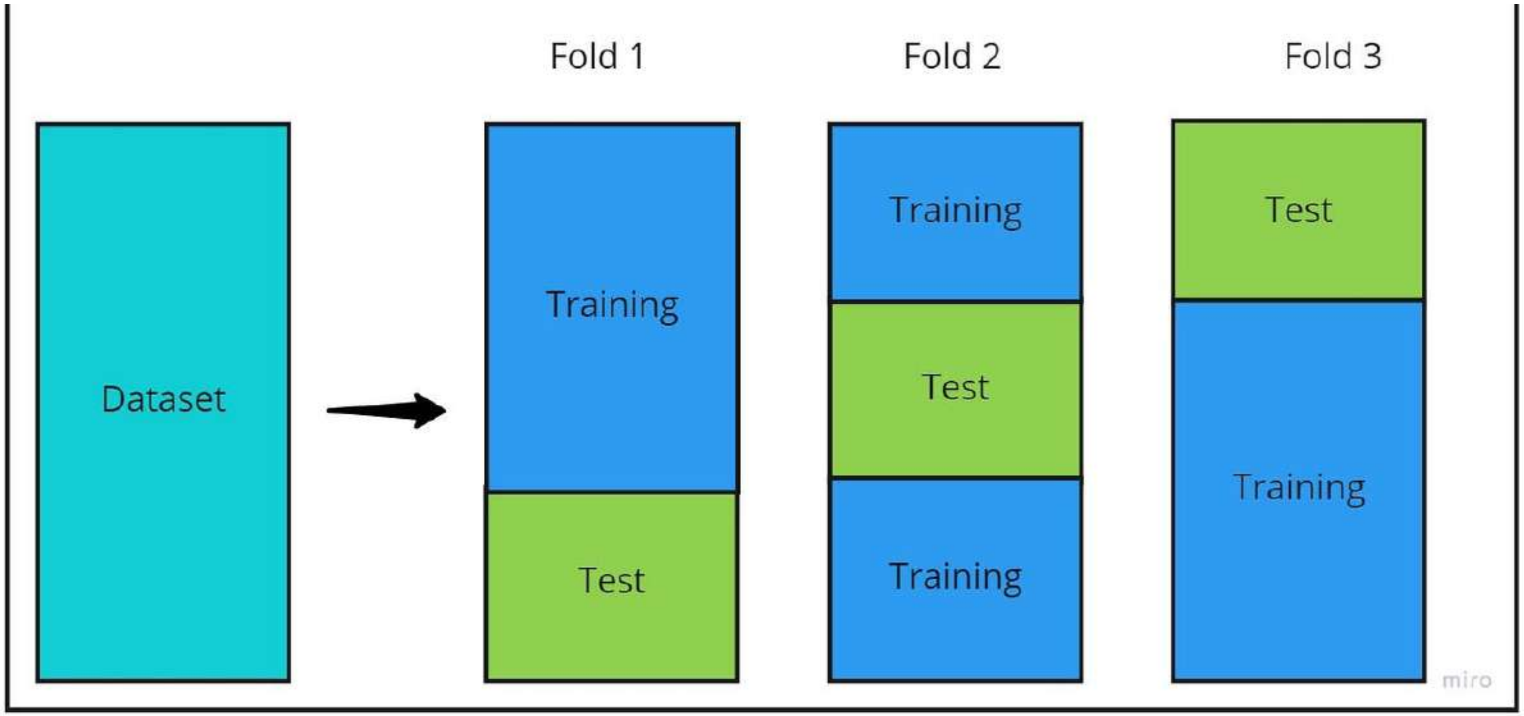
START SMALL

2. SOLVING THE RESEARCH QUESTION

Alternative Approaches

- BLASER proven to outperform simpler models BUT with highly imbalanced data -> analyse other metrics that capture performance under this dataset characteristic
- If BLASER works well in a robust embedding space (SONAR), then a simpler distance measure (cosine similarity) might be enough in SONAR -> reproducibility check
- Let's start small scale:
  - Binary hallucination detection
  - Only 1 language pair: English-Yoruba
  - Cross validation to compensate for highly imbalanced limited dataset

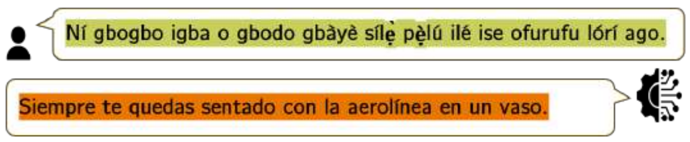
Key Term: Cross Validation



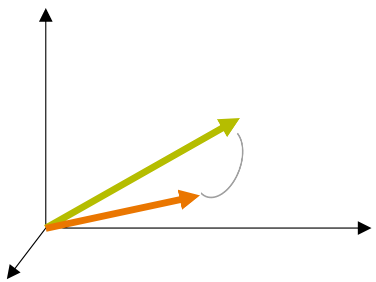
Medium | Cross-Validation

Methodology – Embeddings

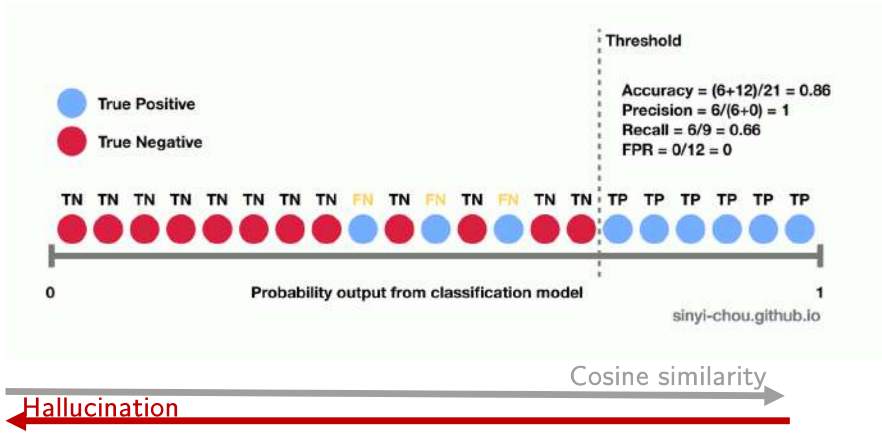
Step 1:
Encoding the source and translation sentences



Step 2:
Calculate cosine similarity for each sentence pair



Step 3: Validation
Calculate the optimal threshold to maximise the F1-Score



Step 4: Test
Use this threshold for binary classification using the test set

CoT: Chain of Thoughts
MCC: Matthew's Correlation Coefficient

Results - Embeddings

Cosine similarity for:

- SONAR
- Cohere 3 Multilingual Embedding from Embed 3 *cohere-embed multilingual-v3_0*
- OpenAI latest embeddings *text-embedding-3-large*

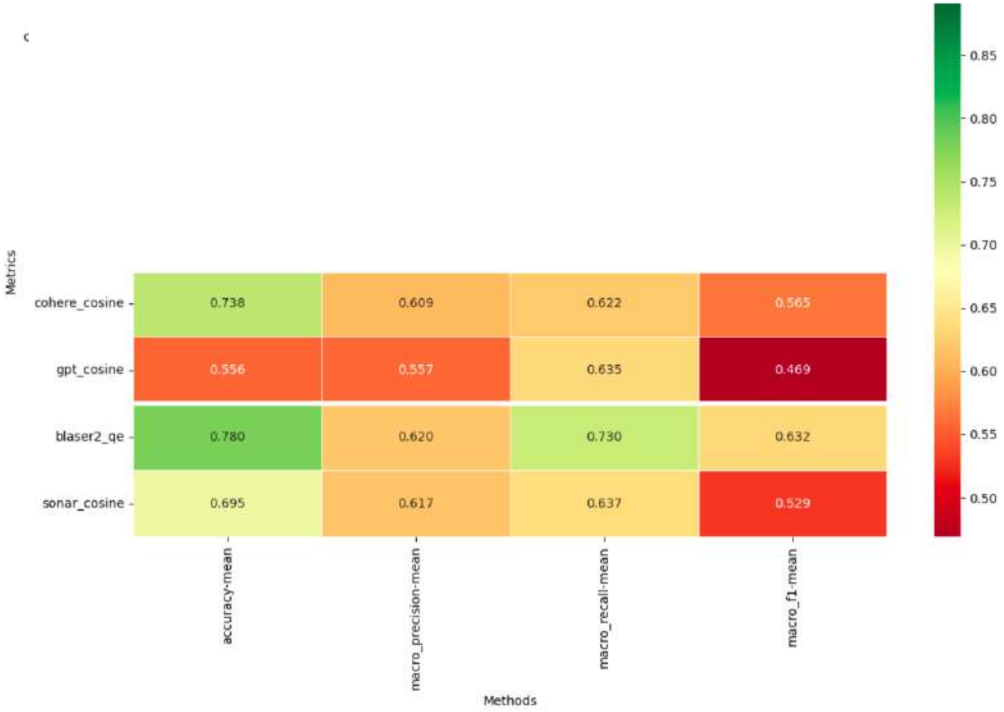


Figure 5: Mean Test scores for LLMs, Embeddings, and *HalOmi* best performing methods.

Conclusion: Embeddings are not enough in order to beat BLASER

What next?

Methodology – LLMs

Building prompting experiments with four Large Language Models

- Cohere Command-R Plus
- Cohere Command-R
- Aya
- OpenAI GPT3.5

Input	**Task introduction**	G-Eval1; G-Eval2
Input	**Chain of Thought (CoT)**	No CoT; CoT
Examples	**Number of examples**	0 examples (zero-shot); 10 examples; 18 examples
Examples	**Examples order: alternative**	Hallucination – No Hallucination; No Hallucination – Hallucination
Examples	**Label format**	1/0; Hallucination / No Hallucination

Prompting hyperparameters

We compared the best performing experiment of each LLM

- 3-fold cross-validation
- 174 validation samples/fold,
- 120 experiments per sample
  ➔ **62 640 operations per LLM**

Pro-tip:
- Find experts advice
- Calculate costs before running

Methodology – LLMs

Building prompting experiments with four Large Language Models

Input	**Task introduction** G-Eval1 G-Eval2
	Chain of Thought (CoT) No CoT CoT
Examples	**Number of examples** 0 examples (zero-shot) 10 examples 18 examples
	Examples order: alternative Hallucination – No Hallucination No Hallucination – Hallucination
	Label format 1/0 Hallucination / No Hallucination

Prompting hyperparameters

Chain of Thought (CoT) Zero/Few-shot

Can we automate reasoning text writing?

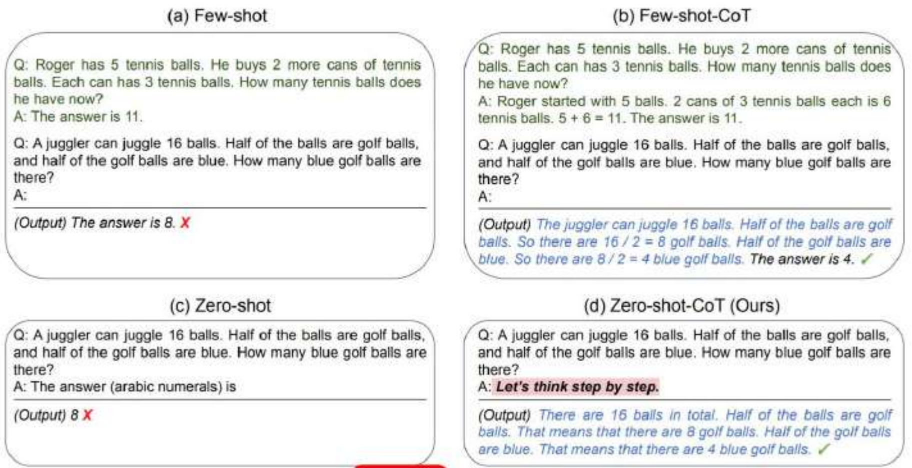


Figure 1: Example inputs and outputs of GPT-3 with (a) standard Few-shot ([Brown et al., 2020]), (b)

Example order for few shot

Order Sensitivity (positional bias?) of Prompts

Review: I like this movie.
Sentiment: Good

Review: I hate this movie.
Sentiment: Bad

Review: Excellent!.
Sentiment: ???

Accuracy: 50 %

Review: I hate this movie.
Sentiment: Bad

Review: I like this movie.
Sentiment: Good

Review: Excellent!.
Sentiment: ???

Accuracy: 85 %

Labels

Let's discuss labels

Label words	Accuracy
	mean (std)
great/terrible	**92.7 (0.9)**
good/bad	92.5 (1.0)
cat/dog	91.5 (1.4)
dog/cat	86.2 (5.4)
terrible/great	83.2 (6.9)

Yao, L. (2024, March 4). Prompting. Lecture presented at UCL.

Methodology - LLMs

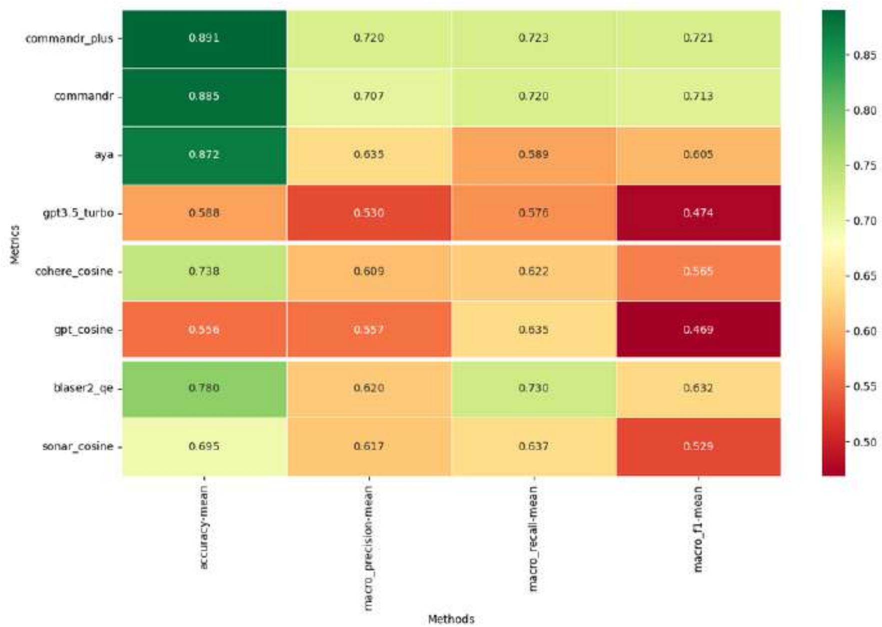
Prompt example

Input	**Task Introduction**	G-Eval1 / G-Eval2
	Chain of Thought (CoT)	No CoT / CoT
Examples	**Number of examples**	0 examples (zero-shot) / 10 examples / 18 examples
	Examples order: alternative	Hallucination – No Hallucination / No Hallucination – Hallucination
	Label format	1/0 / Hallucination / No Hallucination

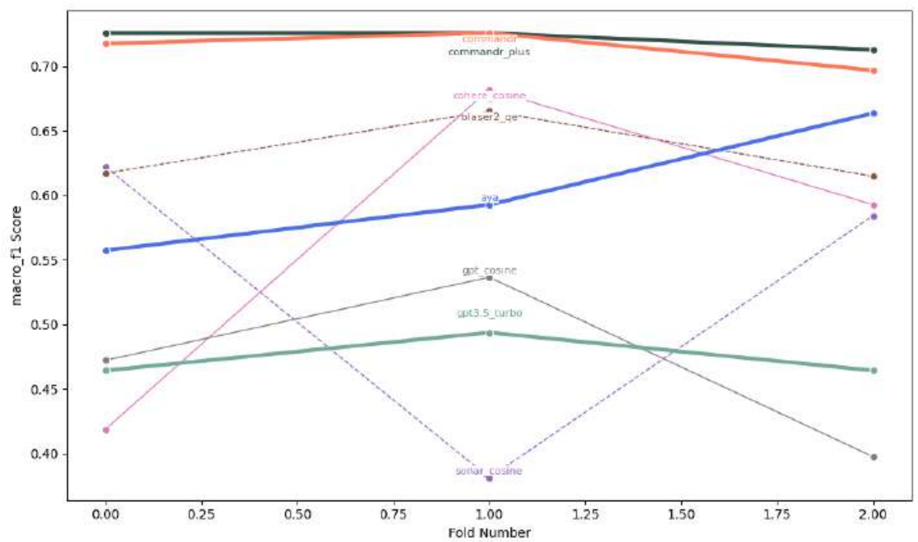
Results

- ✓ LLMs are the best-performing models, beating BLASER
- ✓ Command-R Plus and Command-R outperform all models by 10 points
- ✓ Cohere's embeddings are comparable to BLASER but don't outperform it
- ✓ GPT lags behind, both the LLM and the embeddings

<u>Command-R models outperforming from far</u>



<u>Best performance is also the most robust across folds</u>



We beat BLASER*

**binary, English to Yoruba*

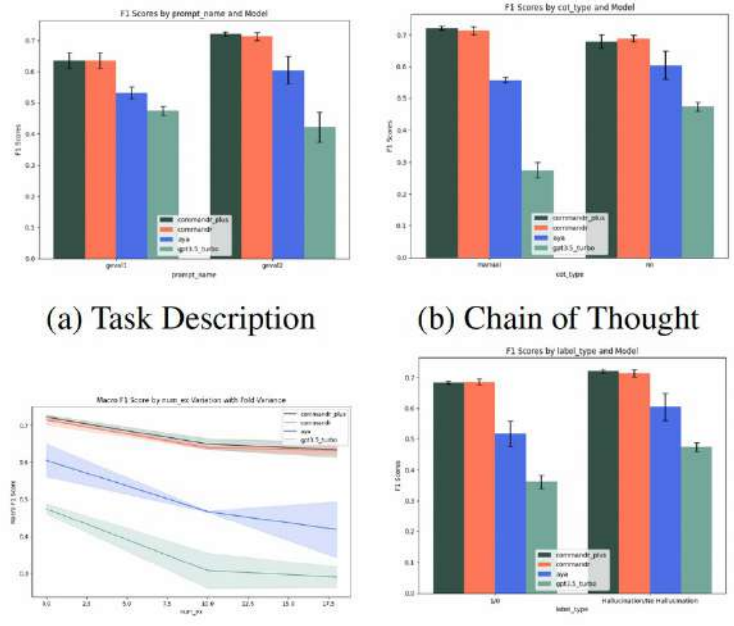
Results

Great difference between stable (Command-R models) and unstable LLMs (Aya and GPT)
G-Eval2 with CoT is preferred for stable models
All models performed better in zero-shot learning

Best performing model for each LLM - accuracy

	CommandR Plus	CommandR	Aya	GPT3.5
1st performing experiment	G-Eval2 CoT 0 H/NH	G-Eval2 CoT 0 H/NH	G-Eval2 *no CoT* 0 H/NH	G-Eval2 *no CoT* 0 H/NH
	72%	71%	60%	47%
2nd perfoming experiment TIP:	G-Eval2 CoT 0 1/0	G-Eval2 CoT 0 1/0	G-Eval2 CoT 0 H/NH	*G-Eval1* CoT *18* *starts 0* H/NH
	68%	69%	56%	46%

LLMs performances depending on hyperparameters



(a) Task Description
(b) Chain of Thought
(c) Number of examples
(d) Format of the label

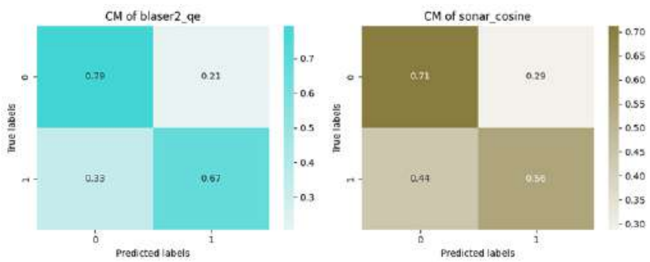
Pro-tip: Find which cases your model performs well and which ones wrong to try to find patterns. It's normally not enough to say something works better, but why?

Results

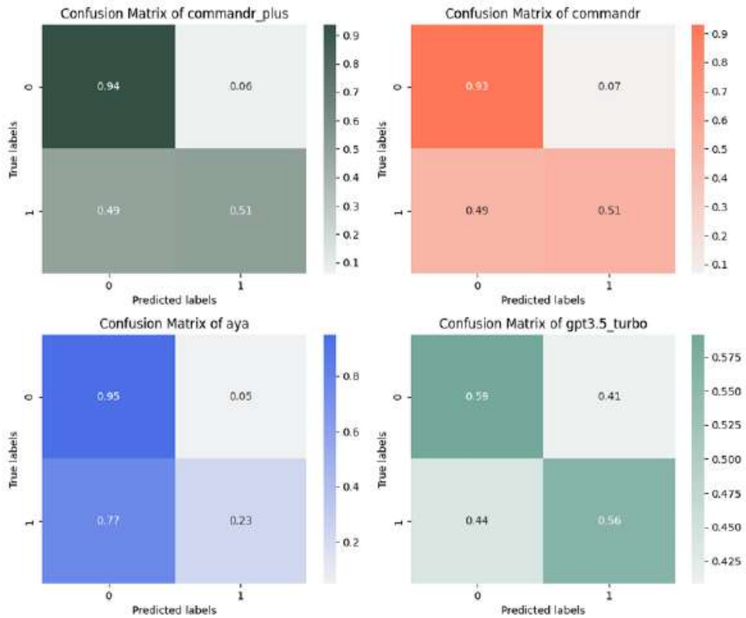
LLMs are the best performing models, beating BLASER

Command-R Plus and Command-R are outperforming all models by 10 points

Baseline



Confusion Matrices



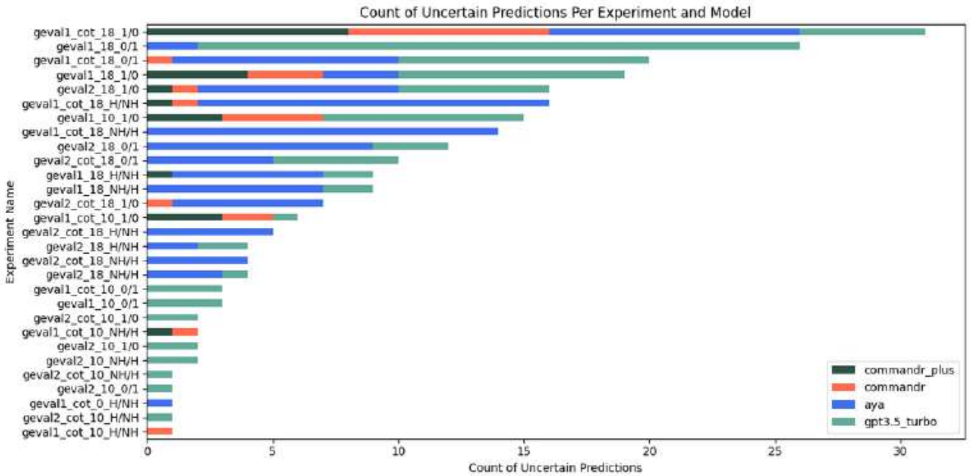
Limits

Why was GPT performing so low? Or explaining why robustness is a condition for accuracy

From embedding space to binary classification

	Number of answers changing prediction across folds
CommandR Plus	3
CommandR	3
Aya	14
GPT3.5	20

Evaluating unclassification within experiments and LLMs



Research extension proposed:

Assessing consistency and improving robustness

- ✔ Scale up to all translation directions
- ✔ Include additional LLMs like Llama
- ✔ Include comparability with previous works
- ✘ Leveraging RAG methods with accurate translations
- ✘ Automatic Prompt Engineering with Resampling

Pro-tip: Speak up to academic team

Pro-tip: Prioritise future work to decide how to scale towards a publishable paper

3. WRITING RESEARCH PAPER

3. WRITING THE RESEARCH PAPER

Writing the Coursework

Mid-April
Write the paper

15th June
ACL Submission

15-20th August
ACL Reviews Results

EMNLP Submission

20th September
EMNLP Results

3rd October
Camera-ready paper
and
Poster

Submissions

Mid-November
Conference

AI Centre,
Dept. of Computer Science,
University College London (UCL), UK

Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models

Kenza Benkirane[1*], Laura Gongas[1*],
Shahar Pelles[1], Naomi Fuchs[1], Joshua Darmon[1],
Pontus Stenetorp[1], David Ifeoluwa Adelani[2,3], Eduardo Sanchez[1,4]

[1]UCL, [2]Mila, Quebec AI institute, [3]McGill University, [4]Meta
*: Equal contributions

First, let's dive into a use case

Ní gbogbo igba o gbodo gbàyè sílẹ̀ pẹ̀lú ilé ise ofurufu lórí ago.
English Human translation: *Always you should reserve a sit with an airline company via a telephone*

Siempre te quedas sentado con la aerolínea en un vaso.
English Human translation: *You always remain seated with the airline in a glass*

Ground Truth: Hallucination

	Embeddings		LLMs			
	SONAR	Cohere Embed v3	GPT4o	Claude Sonnet	Claude Opus	Llama3-70B
Correct hallucination classification:	✗	✗	✓	✓	✓	✓

Research objective:

Evaluate hallucination detection in machine translation (MT)

- across diverse languages, including LRL
- using LLMs and embeddings
- as a binary classification task (hallucination vs. no hallucination)

Dataset and Baseline

Scope:

18 language directions, including high-resource (HRL) and low-resource (LRL) pairs

- High-Resource Languages (HRL): EN ↔ (AR, *DE*, RU, ES, ZH)
- Low-Resource Languages (LRL): EN ↔ (KA, YO, MN)
- Non-English centric: ES ↔ YO

AR: Arabic, DE: German, RU: Russian, ES: Spanish, ZH: Chinese
KA: Kashmiri, YO: Yoruba, MN: Manipuri

Dataset:

HalOmi [1] benchmark for Machine Translation (MT) hallucination detection

Validation set (DE ↔ EN) :	**Test set:**
301 sentence pairs	2,865 sentence pairs

Baseline:

BLASER-QE (previous state-of-the-art)

[1] HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation

Why binary classification?

Direction	Total	1 No	2 Small	3 Partial	4 Full
EN→AR	**144**	136	2	2	4
		94.44%	*1.39%*	*1.39%*	*2.78%*
AR→EN	**156**	132	5	2	17
		84.62%	*3.21%*	*1.28%*	*10.90%*
EN→RU	**146**	141	1	2	2
		96.58%	*0.68%*	*1.37%*	*1.37%*
RU→EN	**158**	146	3	2	7
		92.41%	*1.90%*	*1.27%*	*4.43%*
EN→ES	**153**	131	8	3	11
		85.62%	*5.23%*	*1.96%*	*7.19%*
ES→EN	**160**	127	17	4	12
		79.38%	*10.63%*	*2.50%*	*7.50%*
EN→ZH	**160**	131	5	4	20
		81.88%	*3.13%*	*2.50%*	*12.50%*
ZH→EN	**159**	127	9	7	16
		79.87%	*5.66%*	*4.40%*	*10.06%*
EN→KA	**184**	111	8	30	35
		60.33%	*4.35%*	*16.30%*	*19.02%*
KA→EN	**151**	89	15	32	15
		58.94%	*9.93%*	*21.19%*	*9.93%*
EN→YO	**195**	166	4	11	14
		85.13%	*2.05%*	*5.64%*	*7.18%*
YO→EN	**146**	124	4	10	8
		84.93%	*2.74%*	*6.85%*	*5.48%*
EN→MN	**197**	78	52	54	13
		39.59%	*26.40%*	*27.41%*	*6.60%*
MN→EN	**152**	43	45	58	6
		28.29%	*29.61%*	*38.16%*	*3.95%*
ES→YO	**151**	97	16	29	9
		64.24%	*10.60%*	*19.21%*	*5.96%*
YO→ES	**152**	80	26	37	9
		52.63%	*17.11%*	*24.34%*	*5.92%*
Total	**2564**	1859	220	287	198
		72.47%	*8.58%*	*11.19%*	*7.72%*

- Original dataset was based on **severity ranking** (No, Small, Partial, Full Hallucination)
- Further data analysis showed a **high dataset imbalance**, prompting us to switch to a binary classification paradigm
- We still evaluated the severity ranking analogy with our methods as an **ablation study**

Methodology

We used the Matthew Correlation Coefficient (MCC) *providing a single, easily interpretable value between -1 and +1. This value encapsulates the model's performance for the confusion matrix scores, making it more robust to class imbalance.*

$$MCC = \frac{TN \times TP - FP \times FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}}$$

CoT: Chain of Thoughts

MCC: Matthew's Correlation Coefficient

Ablation study results

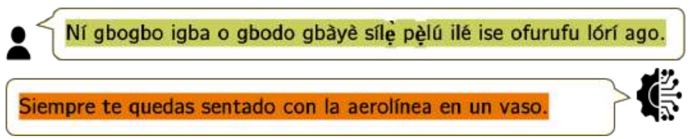
	EN→HRL				HRL→EN				EN→LRL			LRL→EN			ES→YO	YO→ES	*AVG*		
Model	AR	RU	ES	ZH	AR	RU	ES	ZH	KA	YO	MN	KA	YO	MN			**HRL**	**LRL**	**Overall**
GPT text-embedding-3-large	0.89	0.82	0.84	**0.92**	0.91	0.94	0.87	**0.87**	0.71	0.7	0.54	0.56	0.68	0.6	0.62	0.51	**0.88**	0.62	0.75
Cohere Embed v3	0.84	0.87	0.83	0.88	0.9	**0.96**	**0.89**	0.83	0.75	0.73	0.54	0.58	0.74	0.64	0.65	0.59	0.88	**0.65**	0.76
Mistral-embed	**0.92**	0.88	0.82	0.85	0.92	0.86	0.86	0.83	0.72	0.7	0.56	0.53	0.68	0.61	0.63	0.53	0.87	0.62	0.74
SONAR	0.89	0.79	**0.85**	0.77	0.93	0.93	0.85	0.87	0.81	0.8	0.69	0.73	**0.79**	0.73	**0.69**	**0.62**	0.86	0.73	0.8
GPT4-Turbo	0.8	0.72	0.65	0.8	0.86	0.91	0.86	0.79	0.61	0.57	0.26	0.47	0.43	0.31	0.38	0.4	0.8	0.43	0.61
GPT4o	0.71	0.74	0.65	0.8	0.86	0.86	0.74	0.8	0.64	0.58	0.3	0.47	0.59	0.4	0.45	0.41	0.77	0.48	0.63
Command R	0.56	0.88	0.61	0.83	0.86	0.84	0.77	0.68	0.47	0.51	0.19	0.16	0.19	0.33	0.37	0.3	0.75	0.32	0.53
Command R+	0.59	0.56	0.65	0.7	0.91	0.91	0.76	0.74	0.34	0.39	0.04	0.41	0.43	0.26	0.15	0.4	0.73	0.3	0.51
Mistral 8x22b	0.25	0.59	0.53	0.67	0.84	0.94	0.74	0.77	0.51	0.4	0.08	0.46	0.52	0.5	0.33	0.46	0.67	0.41	0.54
Sonnet	0.7	0.75	0.61	0.8	0.84	0.89	0.7	0.69	0.64	0.62	0.41	0.56	0.58	0.55	0.53	0.47	0.75	0.55	0.65
Opus	0.6	**0.91**	0.69	0.83	0.88	0.9	0.83	0.76	0.66	0.54	0.2	0.49	0.7	0.53	0.33	0.49	0.8	0.49	0.65
Llama3-70B	0.6	**0.91**	0.69	0.83	0.88	0.9	0.83	0.76	0.66	0.54	0.2	0.49	0.7	0.53	0.33	0.49	0.8	0.49	0.65
BLASER 2.0-QE	0.9	0.89	0.85	0.78	**0.94**	0.92	0.87	0.87	**0.81**	**0.83**	**0.79**	**0.73**	0.78	**0.8**	0.68	0.58	0.88	**0.75**	**0.81**

Table 5: ROC-AUC results for severity hallucination ranking across HRL and LRL directions. *Bold values indicate the best performing prompt per model.*

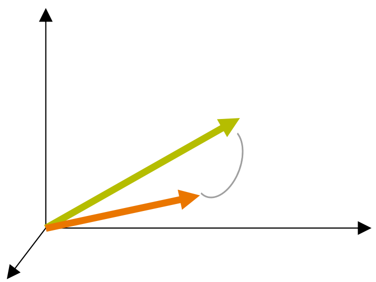
Original dataset was based on **severity ranking** (No, Small, Partial, Full Hallucination)

Methodology – Embedding spaces

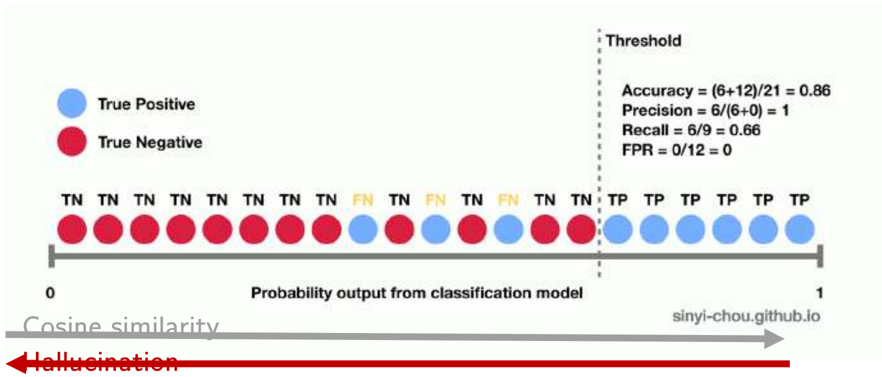
Step 1:
Encoding the source and translation sentences



Step 2:
Calculate cosine similarity for each sentence pair



Step 3: Validation
Calculate the optimal threshold to maximise the F1-Score



Step 4: Test
Use this threshold for binary classification using the test set

CoT: Chain of Thoughts
MCC: Matthew's Correlation Coefficient

Methodology – Large Language Models

System prompts variations:

Three prompts | Two CoT variations | Zero shot learning

```
Human Evaluation of Machine Translation
Systems:
Hallucination Evaluation Criteria: Does the
translated text contain information completely
unrelated to the source text?
- Hallucination: there is hallucination.
-  No Hallucination: there is no hallucination
```

```
Instructions for Evaluating Machine Translation:
You will be given a source text in src_lang and a
machine translated text in tgt_lang. Your task is
to identify if the machine translated text has
hallucination or not. Please make sure you read
and understand these instructions carefully.
Please keep this document open while reviewing,
and refer to it as needed.
Evaluation Criteria:
Hallucination: Does the translated text contain
information completely unrelated to the source
text?
- Hallucination: there is hallucination.
- No Hallucination: there is no hallucination.
```

```
Instructions for Evaluating Machine
Translation:
You will be given a source text in src_lang
and a machine translated text in mt_lang.
Your task is to identify if the machine
translated
text has hallucination or not. Please
make sure you read and understand these
instructions
carefully. Please keep this document
open while reviewing, and refer to it as
needed.
Definition of Hallucination: The translated
text is considered a hallucination if it
introduces
information that is completely unrelated
to the source text.
Hallucination labels:
• Hallucination: there is hallucination.
• No hallucination: there is no hallucination.
```

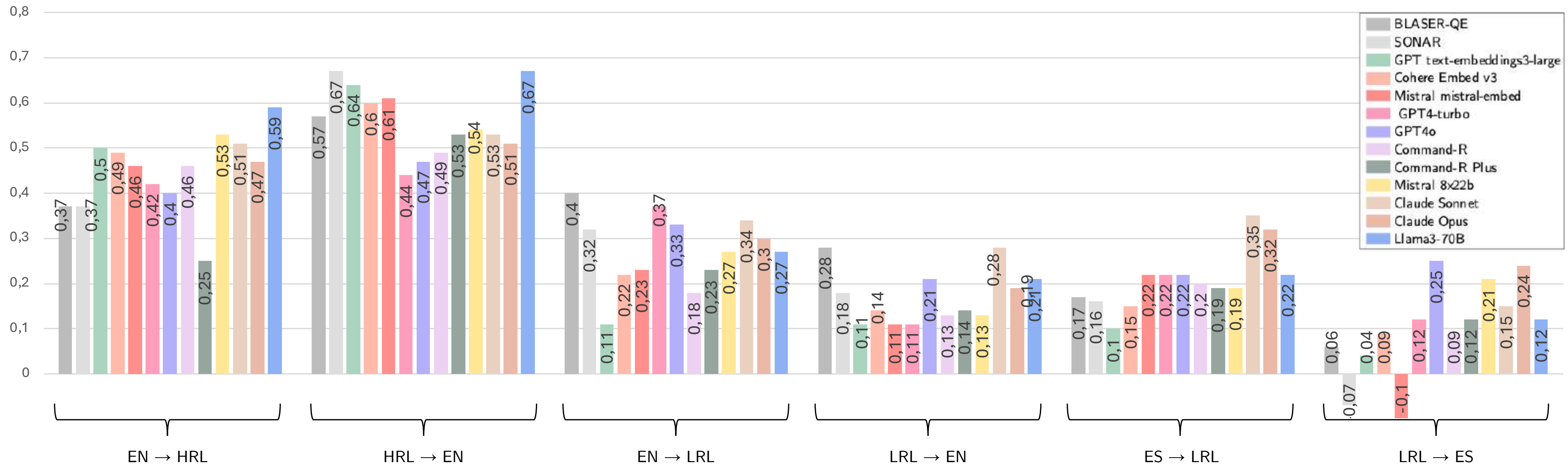
User

```
Source Text: src_text
Translated Text: mt_text
Does the translation contain hallucination?
Answer (label ONLY: ’Hallucination’ OR ’No
Hallucination’):
```

Prompt structure:
Task descriptions, hallucination definitions, and language-specific instructions
Validation criteria: Best prompt per model selected on highest MCC score

Overall results

- → LLMs outperformed BLASER-QE in 13/16 language directions
- → New state-of-the-art achieved for most language pairs



Findings

Capabilities	Resources	Performance across languages
LLMs can outperform previous models without being explicitely trained on	Embedding spaces are competitive for	High Resource Low Resource Non-Latin Scripts

Results summarised

High-resource languages (HRLs):

- **Best overall:** Llama3-70B (0.63 MCC, +16 points over BLASER-QE)
- 10/12 evaluated methods surpassed BLASER-QE
- Embedding methods competitive, especially for HRL→EN

Low-resource languages (LRLs):

- **Best overall:** Claude Sonnet
- GPT4o most consistent across all languages (lowest standard deviation)

Non- Latin scripts:

Embeddings performed well for HRL→EN

Non-English translations:

Opus showed promise, outperforming embeddings

Conclusion and future work

Performance factors

→ Resource level (HRL vs. LRL)

→ Translation direction (to/from English, non-English centric)

→ Script type (Latin vs. non-Latin)

Challenges

→ LRL performance still lags behind HRL

→ Dataset imbalances affect evaluation

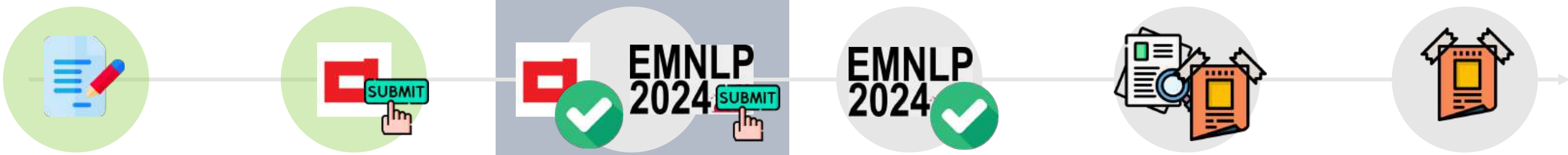
→ Non-English centric translations remain difficult

→ Fine-grained hallucination span detection to improve interpretability

3. WRITING THE RESEARCH PAPER

Writing the Coursework

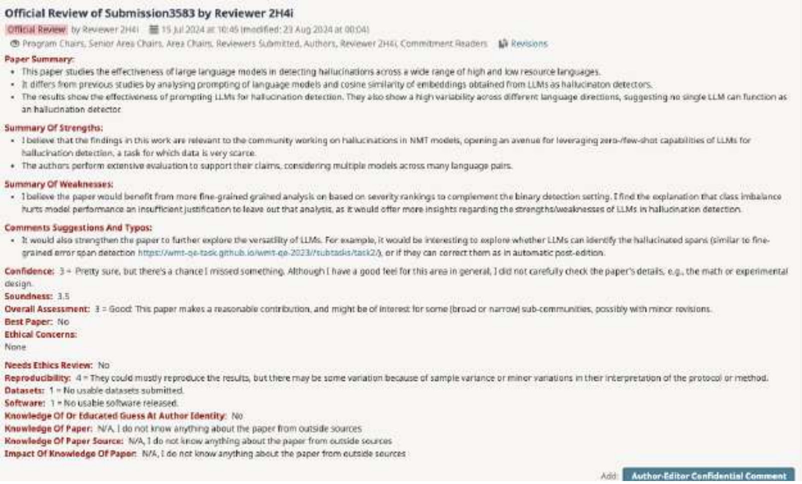
Mid-April | *15th June* | *15-20th August* | *20th September* | *3rd October* | *Mid-November*



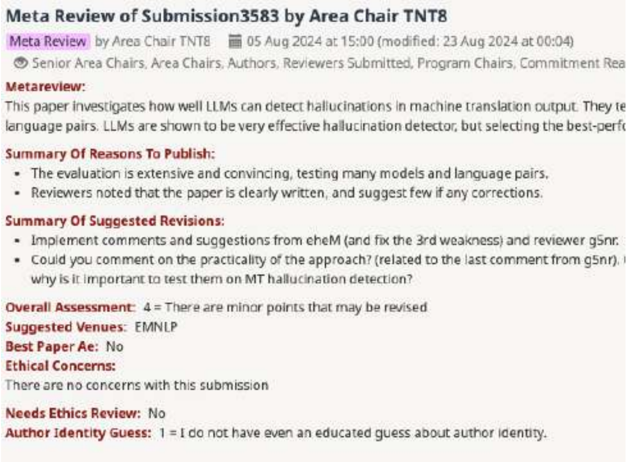
- *Mid-April*: Write the paper
- *15th June*: **ACL Submission**
- *15-20th August*: ACL Reviews Results — **EMNLP Submission**
- *20th September*: EMNLP Results
- *3rd October*: **Camera-ready paper** and **Poster** **Submissions**
- *Mid-November*: Conference

ACL Rolling Review (ARR)

Step 1: Receive and answer reviews



Step 2: Receive the meta-review



✓ Recieve reviews and have 5 days to answer, most of the time during the weekend
✓ A few weeks later, receive the meta review with an overall assesment

Paper accepted!

Workshops

Workshop: interactive session focused on hands-on activities, discussions, or in-depth learning about a specific topic, often designed to engage participants actively.

Pro-tip:
- Great way to submit a paper and be exposed to your specific field
- Funding opportunities

Workshops

W1. BlackboxNLP 2024: Analysing and interpreting neural networks for NLP

- Organizers: Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Hanjie Chen and Yonatan Belinkov
- Date: Fri, Nov 15

W2. Seventh Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2024)

- Organizers: Maciej Ogrodniczuk, Sameer Pradhan, Anna Nedoluzhko, Massimo Poesio and Vincent Ng
- Date: Fri, Nov 15

W3. Seventh Workshop on Fact Extraction and VERification (FEVER)

- Organizers: Michael Sejr Schlichtkrull, Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Zhenyun Deng, Arpit Mittal, James Thorne and Andreas Vlachos
- Date: Fri, Nov 15

W4. Workshop on the Future of Event Detection

EMNLP 2024

NINTH CONFERENCE ON MACHINE TRANSLATION (WMT24)

Widening NLP
Natural Language Processing

EMNLP 2024 am Registration Participants Info Calls Venue Blog Committees Spons

CALL FOR
Diversity & Inclusion Subsidies
Main Conference Papers
System Demonstrations
Industry Track

Call for EMNLP 2024 Diversity and Inclusion Subsidies

3. WRITING THE RESEARCH PAPER

Writing the Coursework

Mid-April
Write the paper

15th June
ACL Submission

15-20th August
ACL Reviews Results

EMNLP Submission

20th September
EMNLP Results

3rd October
Camera-ready paper and **Poster**

Submissions

Mid-November
Conference

The conference

In total, there are 1271 papers accepted to the Main Conference and 1029 papers accepted to Findings. The acceptance rate for Main Conference papers is 20.8%

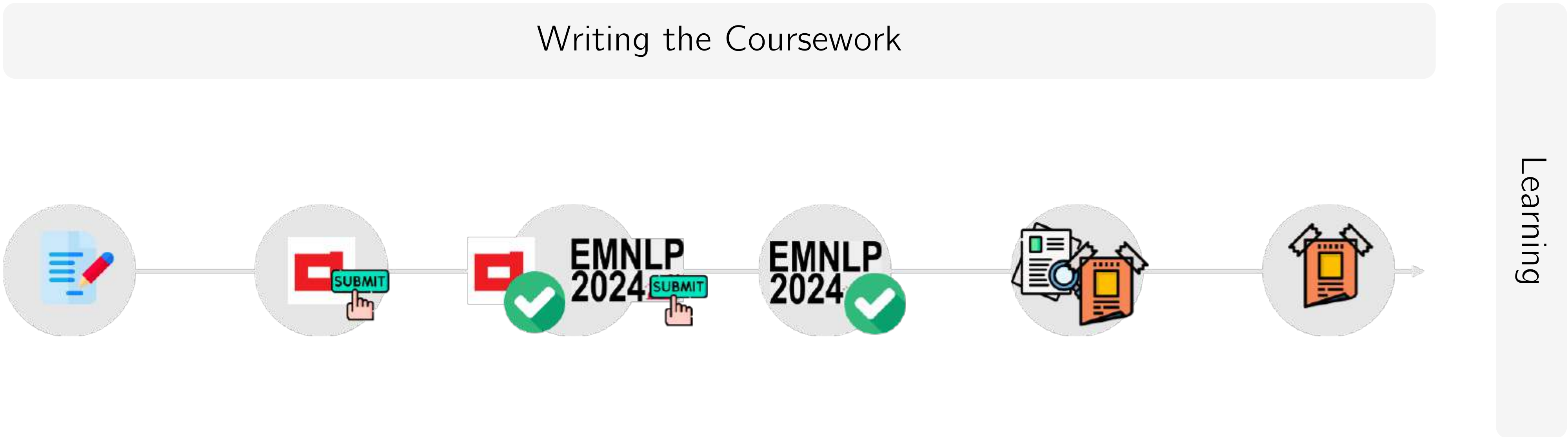
Pro-tip:

- Try to give yourself as much exposure as possible
- List before the conference the people you'd like to meet and ideally contact them for coffee chat.
- Be proactive

3. PRACTICAL TIPS

What have we learned and how you can use it

For your coursework, for your thesis, for your next job, etc...



Be aware of timing

- Start working early
- **Keep notes** of everything
- Everything takes **1.25 more time**
- Less is more

Go deep on the problem, not the solution

- **Read papers**, make sure you understand them profoundly, and reproduce them if necessary
- The best way to learn about a subject is to read the paper.
- Mentors give great feedback, make sure to come prepare to make the most out of your time

Be curious

- Each lecture will give you a new lens from which to see the problem: be proactive
- There's lot of news in the field, try to inspect the companies working on your problem, the techniques they use, the papers they publish, etc.

What helped us during our NLP module

- ✓ *Speech and Language Processing* – 2025 new pdf shared
- ✓ **Read paper:** Go straight to the source
- ✓ **GitHub**: read code, reproduce code, etc.
- ✓ Read news, but the right ones, and the right amount

How to stay up to date

For Business: Newsletters

About AI

TLDR Tech, TLDR AI, TLDR DevOps

TheScroll *DataScienceWeekly*

AlphaSignal *The Tech Buzz* *NoCode.AI*

About Healthcare AI

What The Health? *DoctorPenguin*

HealthTech Pigeon

Machine Learning for MDs

Apps

Meco

substack

For Developers: Social Platforms

@GoogleAI, @GoogleHealth, @GoogleDeepMind
@OpenAI **@AnthropicAI**
@MistralAILabs **@Cohere**

X / Twitter

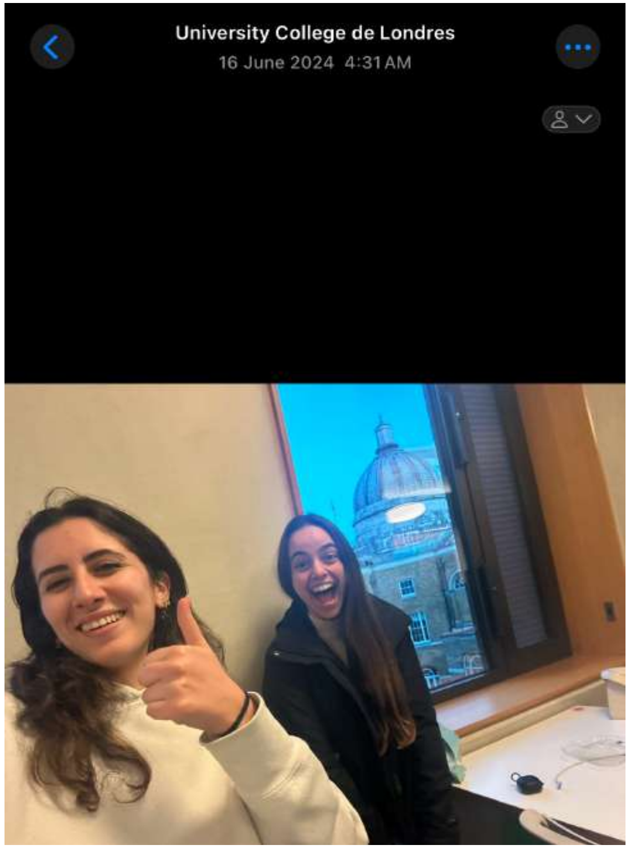
@dwarkesh_sp
@AIBreakfast
@trending_repos

@DAIR_AI
@DeepLearningAI
@karpathy

Mistral, Perplexity, Google Developers, Cohere, OpenAI, HuggingFace,
All have public discord workspaces with links available on their website.

Discord

It's worth the work



Thank you for your attention!

Check out our paper
for more details!

Laura Gongas

laura.gkogka.23@ucl.ac.uk

Kenza Benkirane

kenza.benkirane.23@ucl.ac.uk

Kenza Benkirane[1], Laura Gongas[1*],*
Shahar Pelles[1], Naomi Fuchs[1], Joshua Darmon[1],
Pontus Stenetorp[1], David Ifeoluwa Adelani[2,3], Eduardo Sanchez[1,4]
**: Equal contributions*