



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 Pelles¹, Naomi Fuchs¹, Joshua Darmon¹,
Pontus Stenetorp¹, David Ifeoluwa Adelani^{2,3}, Eduardo Sanchez^{1,4}

¹UCL, ²Mila, Quebec AI institute, ³McGill University, ⁴Meta

*: 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

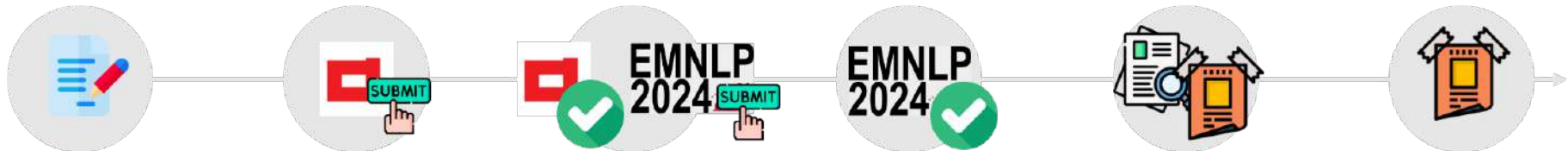
1. DEFINING THE PROJECT

Ideation phase

2. SOLVING THE RESEARCH QUESTION

Implementation phase

3. WRITING THE RESEARCH



4. PRACTICAL TIPS

What we've learned on the way



Learning



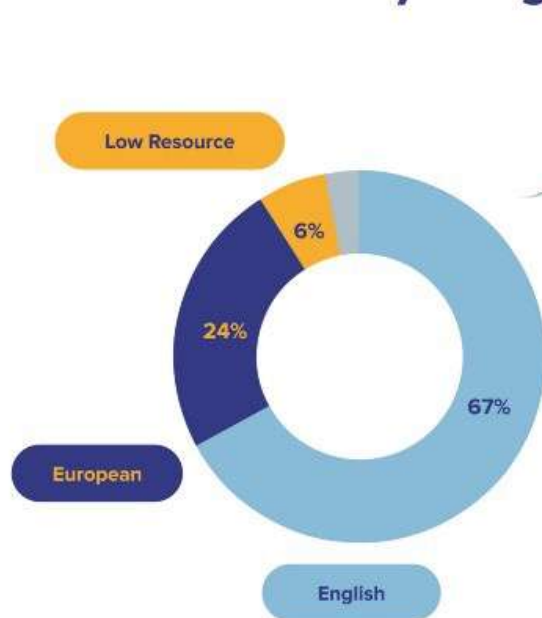
1. DEFINING THE PROJECT



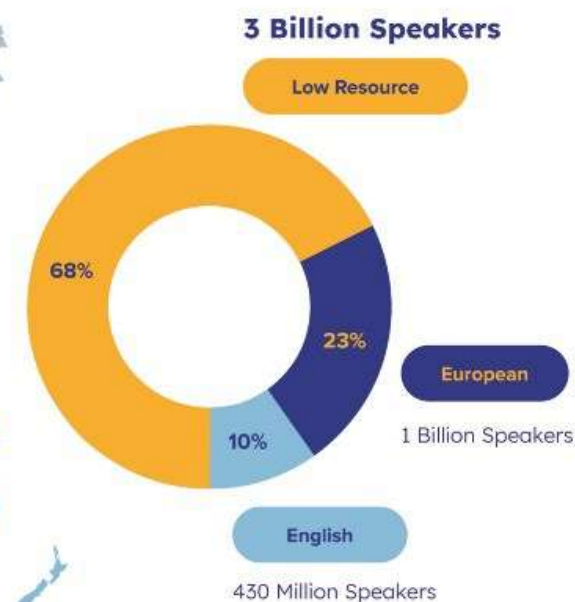
1. DEFINING THE PROJECT

Key Term: Low Resource Languages

NLP Solutions by Language



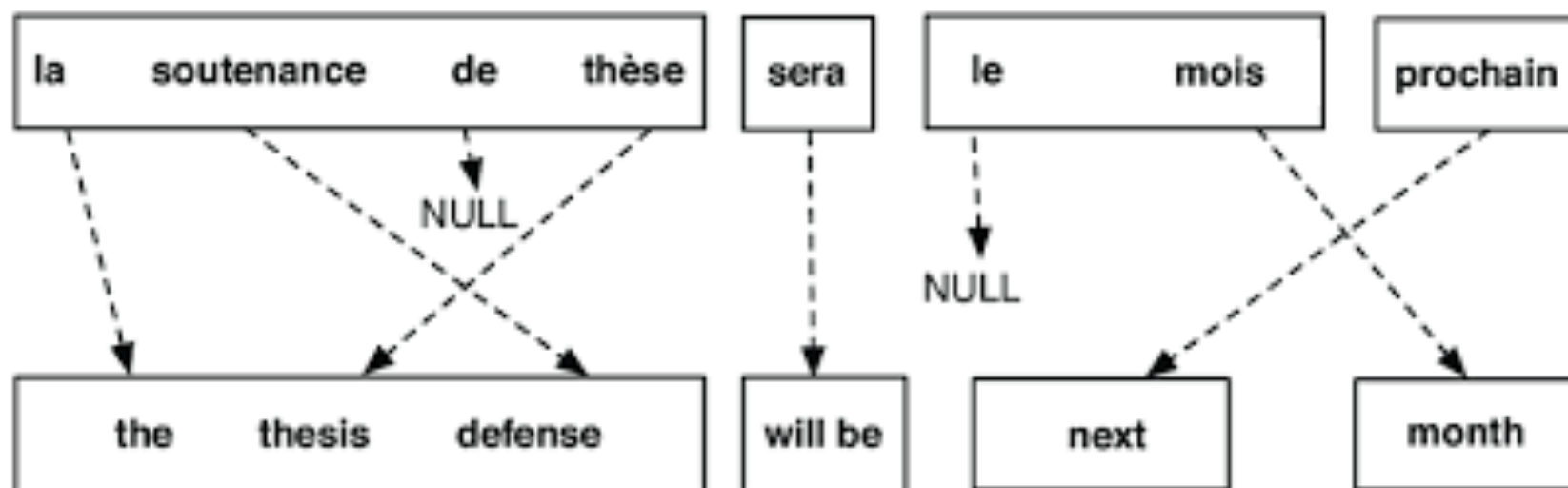
Population Size of Languages



Medium / Low Resource Language - what does it mean?

1. DEFINING THE PROJECT

Key Term: Machine Translation



Emu: Enhancing Multilingual Sentence Embeddings with Semantic Similarity

1. DEFINING THE PROJECT

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

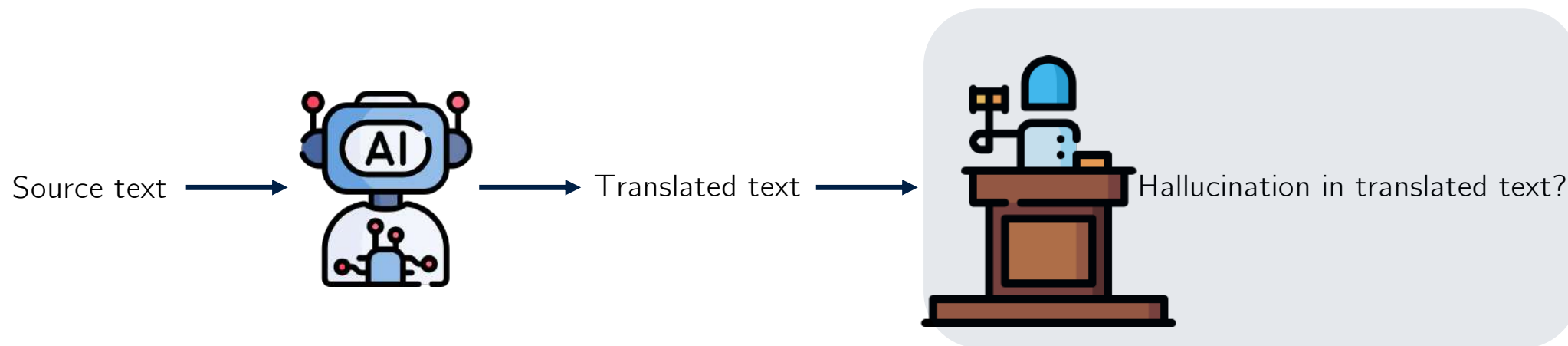
1. DEFINING THE PROJECT



1. DEFINING THE PROJECT

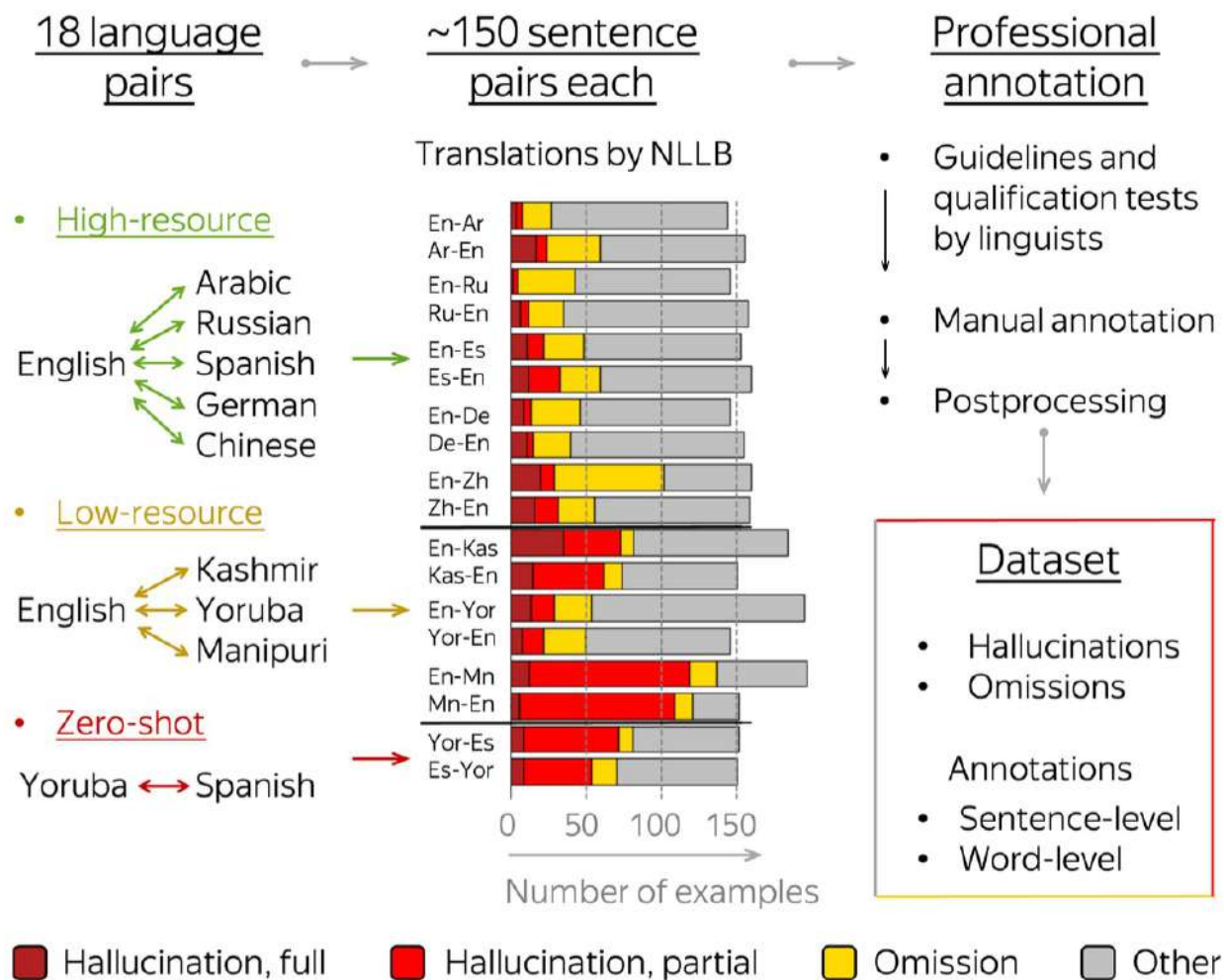
HalOmi

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



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

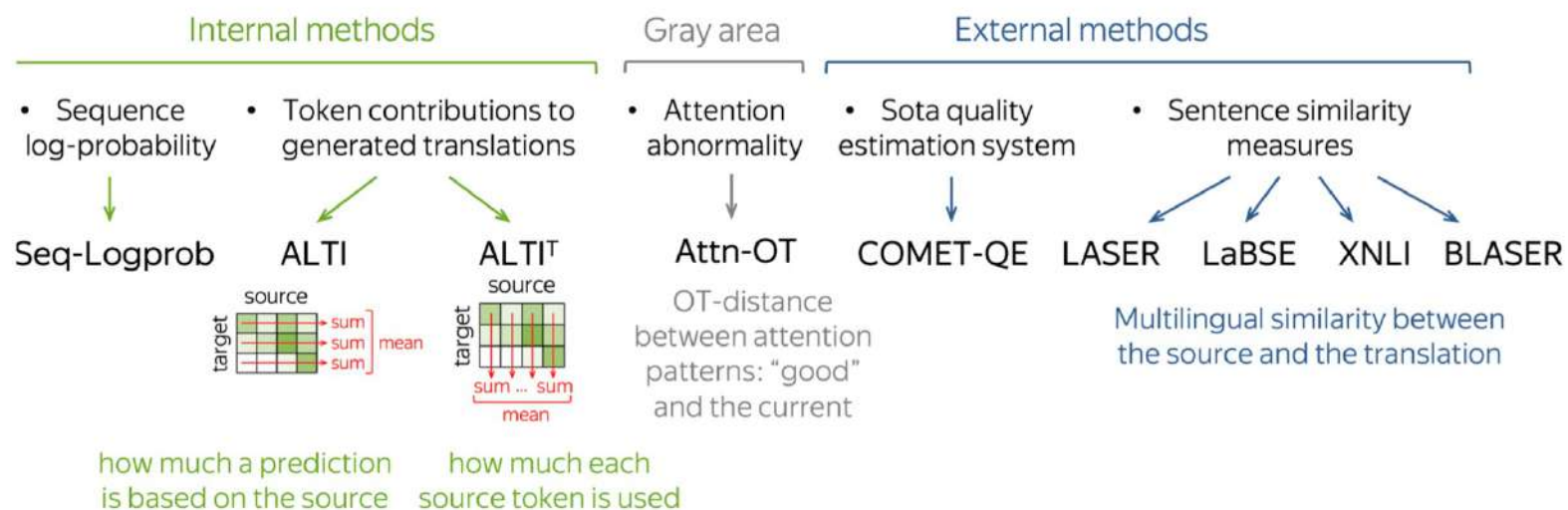
1. DEFINING THE PROJECT



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

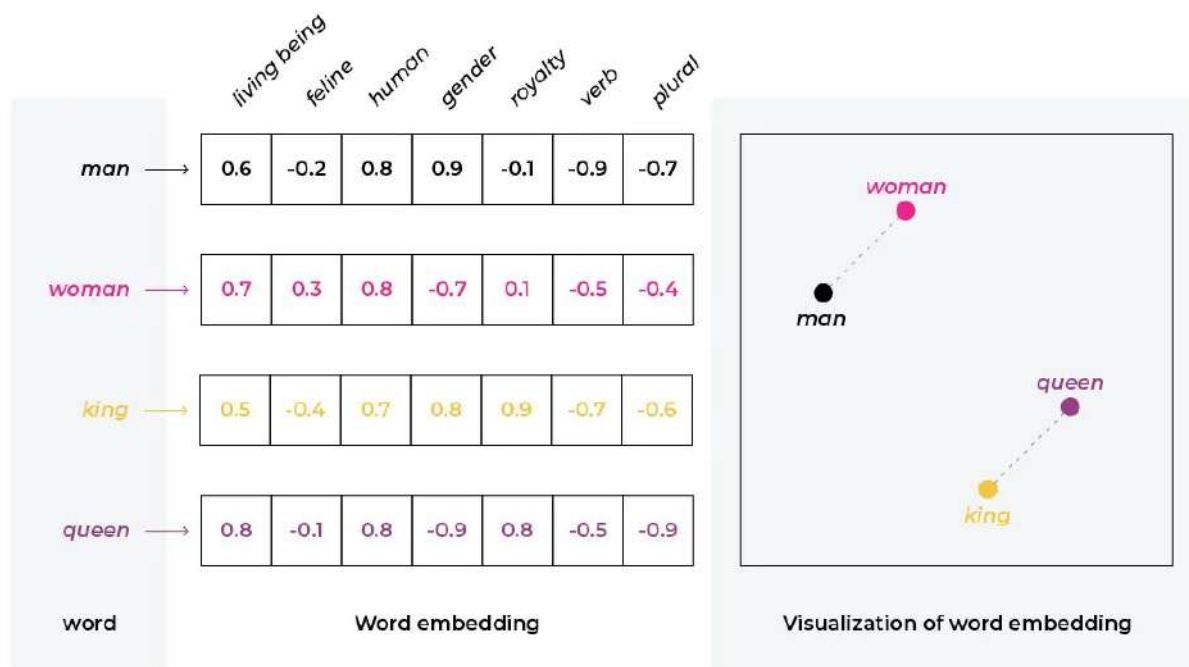
External

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

1. DEFINING THE PROJECT

External

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



Step 1:

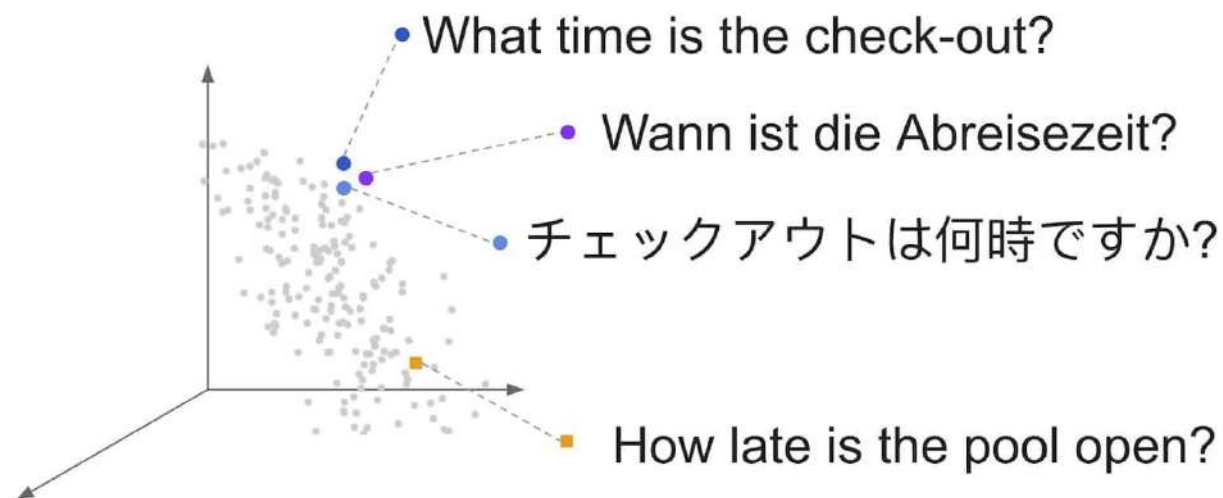
Map the source and translation sentences to an embedding space.

Arize AI / Embeddings, how to compute them

1. DEFINING THE PROJECT

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:

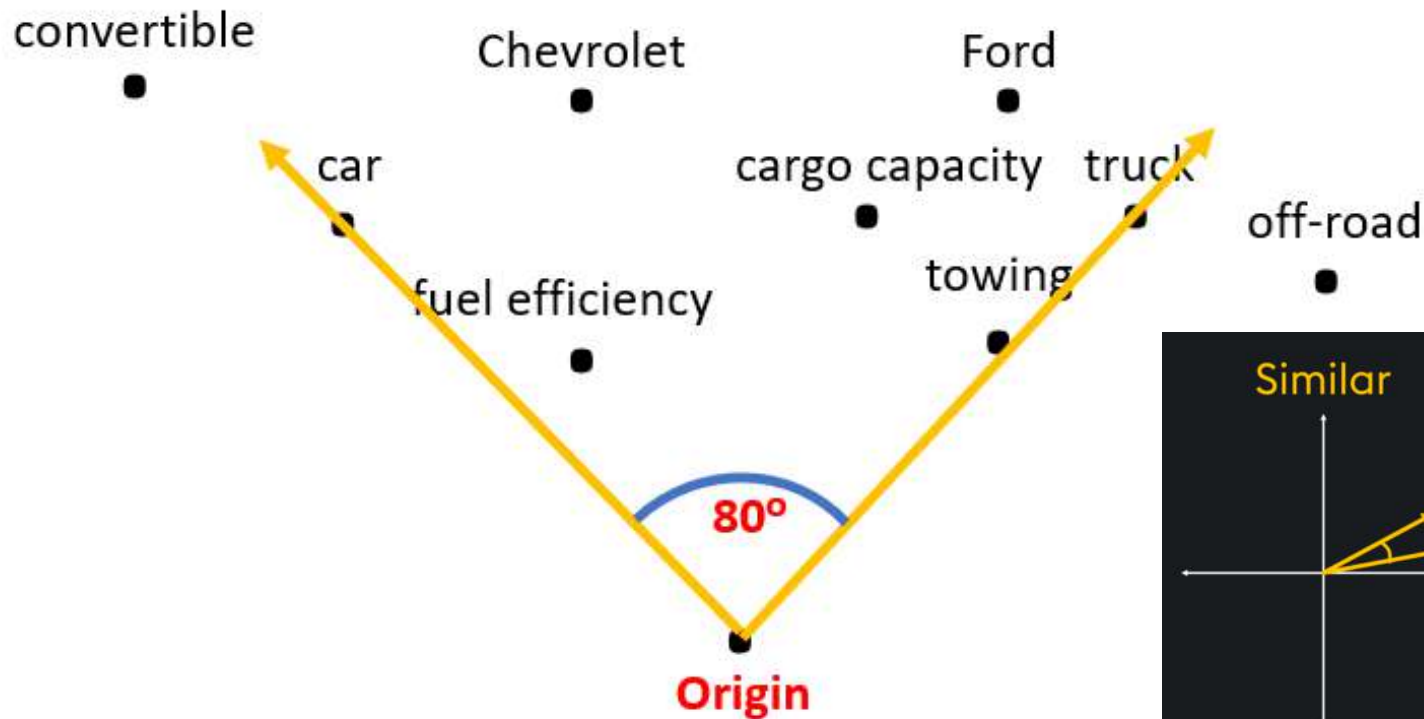
Measure the distance between the embedded source and translated sentences.

Distance measures:

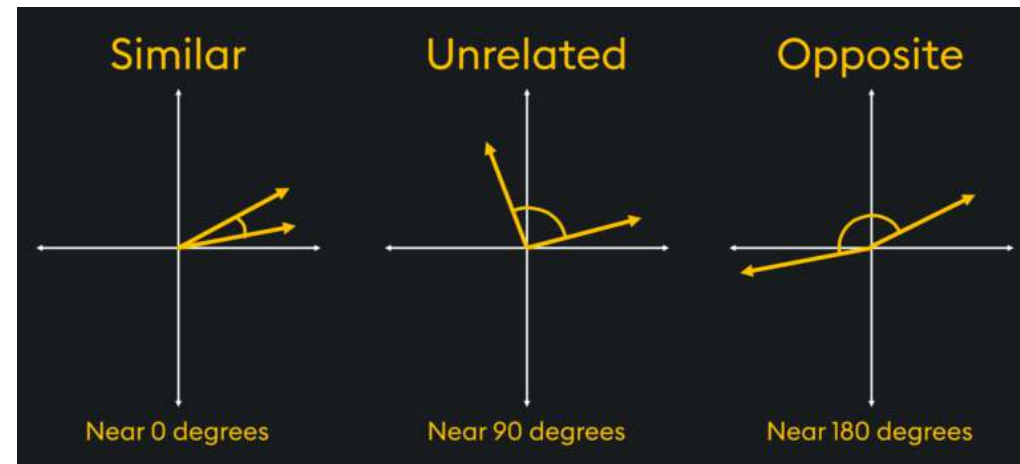
- Cosine similarity
- **BLASER**: neural network trained to calculate the distance between embedding points

1. DEFINING THE PROJECT

→ **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



1. DEFINING THE PROJECT

		Internal				External							Internal				External					
		Seq-Logprob	ALTI	ALTI [†]	Attn-OT	COMET-QE	LaBSE	LASER	XNLI	BLASER-QE			Seq-Logprob	ALTI	ALTI [†]	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			0.63	0.44	0.76	0.64	0.61	0.79	0.77	0.76	0.85	
	Arabic-English	0.83	0.75	0.72	0.51	0.80	0.88	0.83	0.87	0.94			0.57	0.49	0.82	0.72	0.63	0.83	0.62	0.86	0.77	
	English-Russian	0.95	0.36	0.37	0.61	0.99	0.86	0.80	0.82	0.89			0.53	0.56	0.79	0.75	0.53	0.85	0.74	0.80	0.84	
	Russian-English	0.86	0.76	0.56	0.55	0.78	0.83	0.85	0.77	0.92			0.59	0.52	0.86	0.76	0.57	0.76	0.75	0.80	0.80	
	English-Spanish	0.87	0.85	0.69	0.59	0.85	0.88	0.84	0.62	0.85			0.39	0.31	0.89	0.89	0.65	0.86	0.64	0.85	0.80	
	Spanish-English	0.92	0.89	0.67	0.37	0.86	0.94	0.88	0.89	0.87			0.61	0.50	0.59	0.66	0.69	0.73	0.62	0.70	0.71	
	English-German	0.85	0.97	0.69	0.55	0.88	0.97	0.88	0.94	0.87			0.50	0.46	0.77	0.74	0.54	0.66	0.57	0.56	0.85	
	German-English	0.90	0.80	0.59	0.65	0.92	0.95	0.86	0.88	0.97			0.48	0.38	0.82	0.83	0.62	0.70	0.64	0.77	0.70	
	English-Chinese	0.88	0.82	0.47	0.60	0.84	0.86	0.79	0.84	0.78			0.60	0.51	0.73	0.67	0.57	0.69	0.73	0.63	0.88	
Chinese-English	0.89	0.88	0.65	0.46	0.89	0.88	0.82	0.78	0.87	0.73	0.61		0.77	0.64	0.66	0.75	0.74	0.79	0.73			
High-resource mean		0.88	0.79	0.59	0.52	0.86	0.89	0.84	0.81	0.89			0.56	0.48	0.78	0.73	0.61	0.76	0.68	0.75	0.79	
Low-resource	English-Kashmir	0.68	0.71	0.74	0.54	0.64	0.76	0.70	0.68	0.81			0.50	0.52	0.90	0.81	0.80	0.76	0.55	0.60	0.77	
	Kashmir-English	0.59	0.67	0.65	0.65	0.69	0.58	0.59	0.47	0.73			0.36	0.50	0.64	0.63	0.61	0.59	0.44	0.53	0.45	
	English-Yoruba	0.68	0.81	0.49	0.54	0.57	0.80	0.76	0.35	0.83			0.45	0.52	0.80	0.73	0.51	0.82	0.68	0.64	0.80	
	Yoruba-English	0.70	0.64	0.49	0.57	0.58	0.57	0.59	0.46	0.78			0.42	0.33	0.68	0.63	0.51	0.52	0.45	0.52	0.74	
	English-Manipuri	0.77	0.74	0.59	0.65	0.68	0.56	0.57	0.44	0.79			0.77	0.50	0.70	0.72	0.59	0.65	0.41	0.54	0.67	
Manipuri-English	0.78	0.72	0.54	0.43	0.58	0.66	0.69	0.47	0.80	0.68	0.60		0.65	0.55	0.57	0.66	0.76	0.56	0.56			
Low-resource mean		0.70	0.72	0.58	0.56	0.62	0.66	0.65	0.48	0.79			0.53	0.50	0.73	0.68	0.60	0.67	0.55	0.57	0.66	
Zero-shot	Yoruba-Spanish	0.60	0.65	0.47	0.44	0.54	0.56	0.60	0.51	0.58			0.62	0.47	0.85	0.69	0.40	0.62	0.63	0.31	0.66	
	Spanish-Yoruba	0.61	0.66	0.52	0.55	0.55	0.66	0.64	0.52	0.68			0.68	0.50	0.83	0.60	0.51	0.69	0.77	0.42	0.67	
Overall mean		0.79	0.75	0.58	0.53	0.75	0.78	0.75	0.67	0.83			0.56	0.48	0.77	0.70	0.59	0.72	0.64	0.65	0.74	

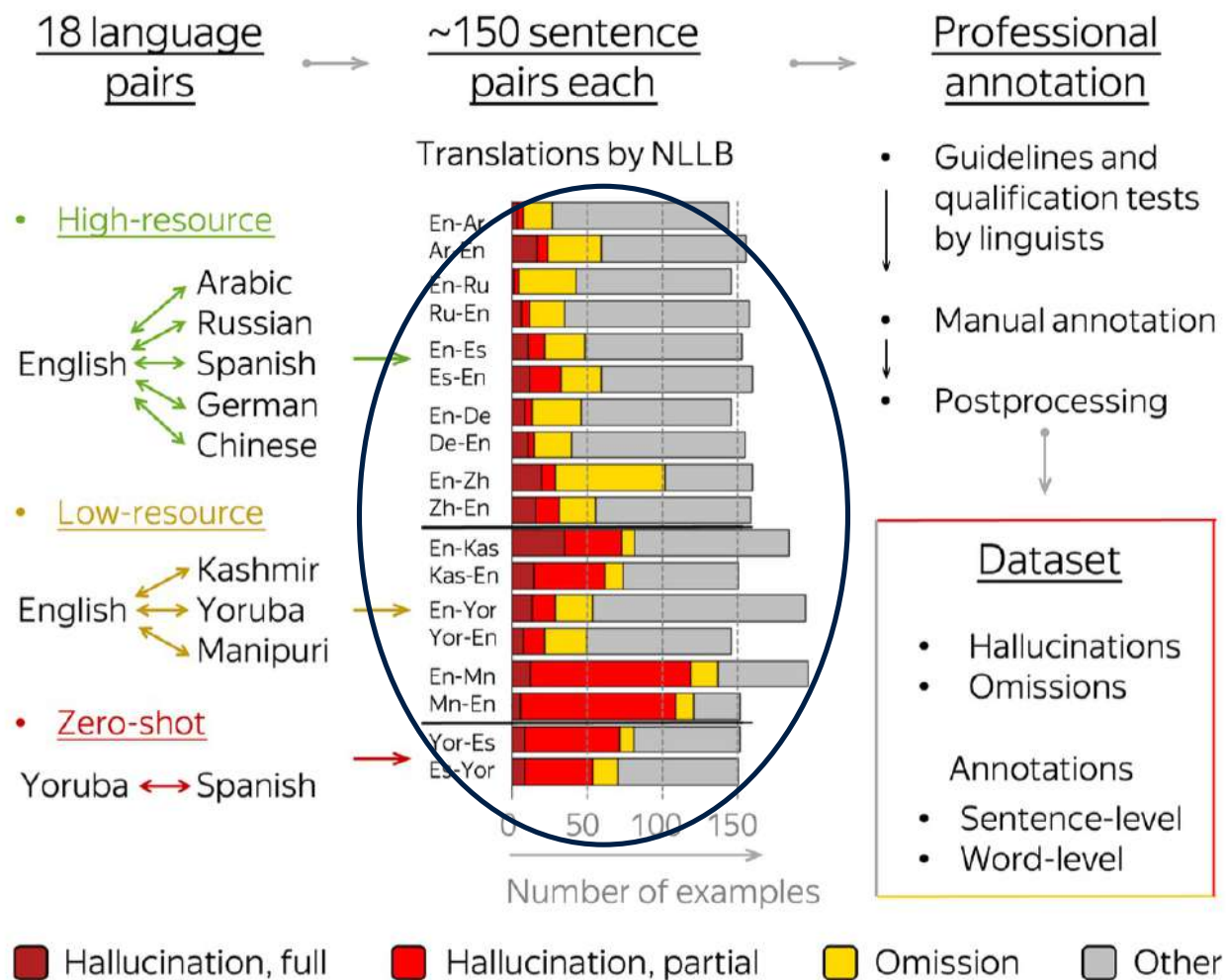
Hallucinations
(among all translations)

Omissions
(among non-hallucinations)

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

*Define group's
interests*



Identify limitations



Literature review

*The Research
Question*





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.

2. SOLVING THE RESEARCH QUESTION

*Define group's
interests*



Identify limitations



Literature review

Set an objective



START SMALL



2. SOLVING THE RESEARCH QUESTION

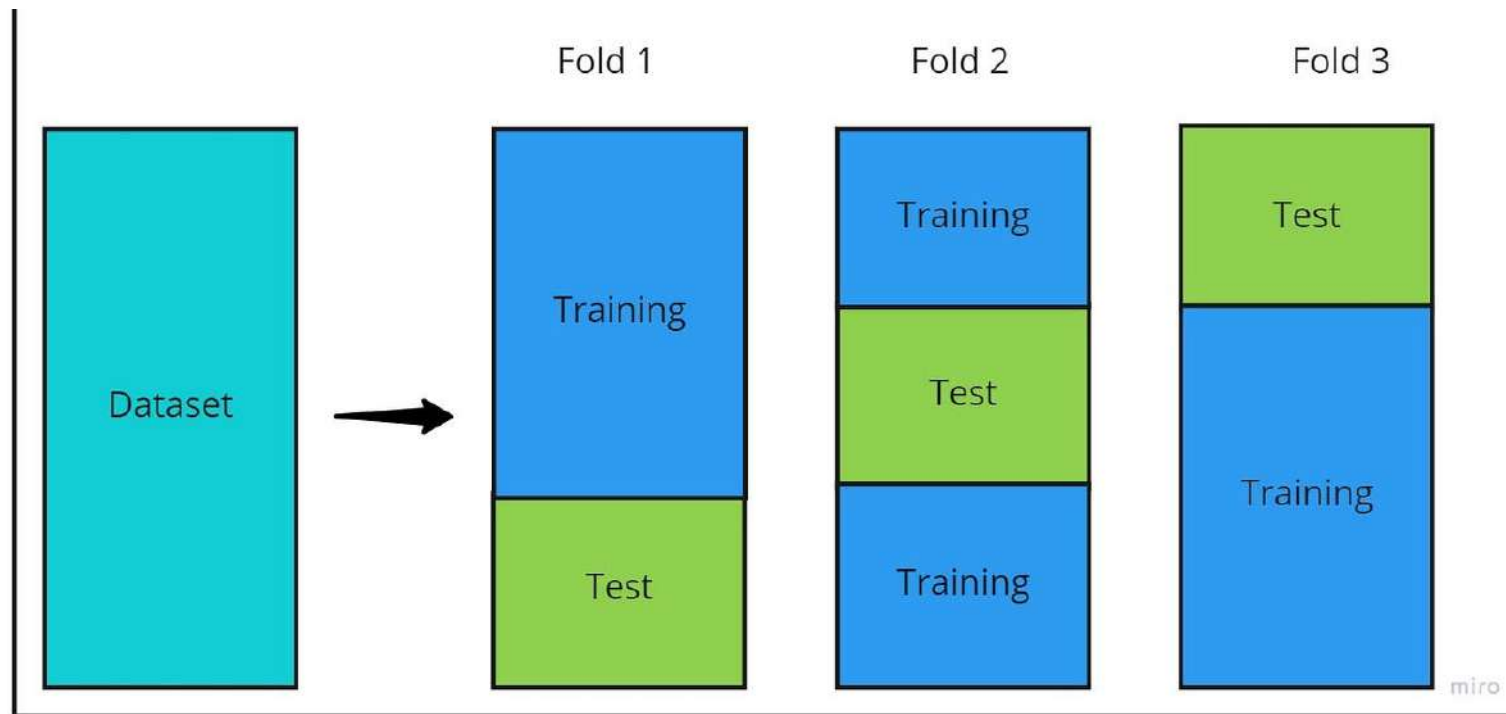
“IF THE ONLY TOOL YOU HAVE IS A HAMMER,
IT IS TEMPTING TO TREAT EVERYTHING
AS IF IT WERE A NAIL”



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



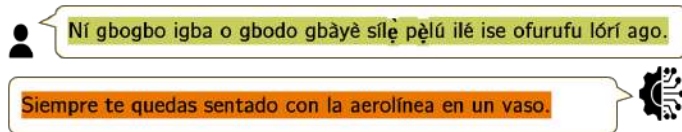
Medium / Cross-Validation

2. SOLVING THE RESEARCH QUESTION

Methodology – Embeddings

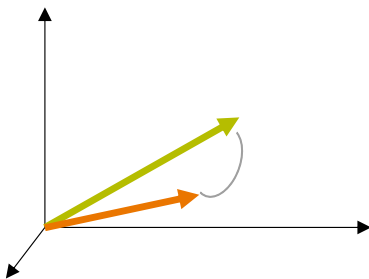
Step 1:

Encoding the source and translation sentences



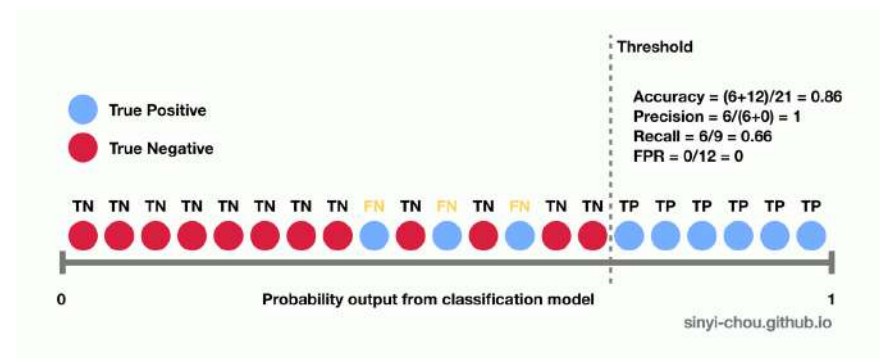
Step 2:

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

2. SOLVING THE RESEARCH QUESTION

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*

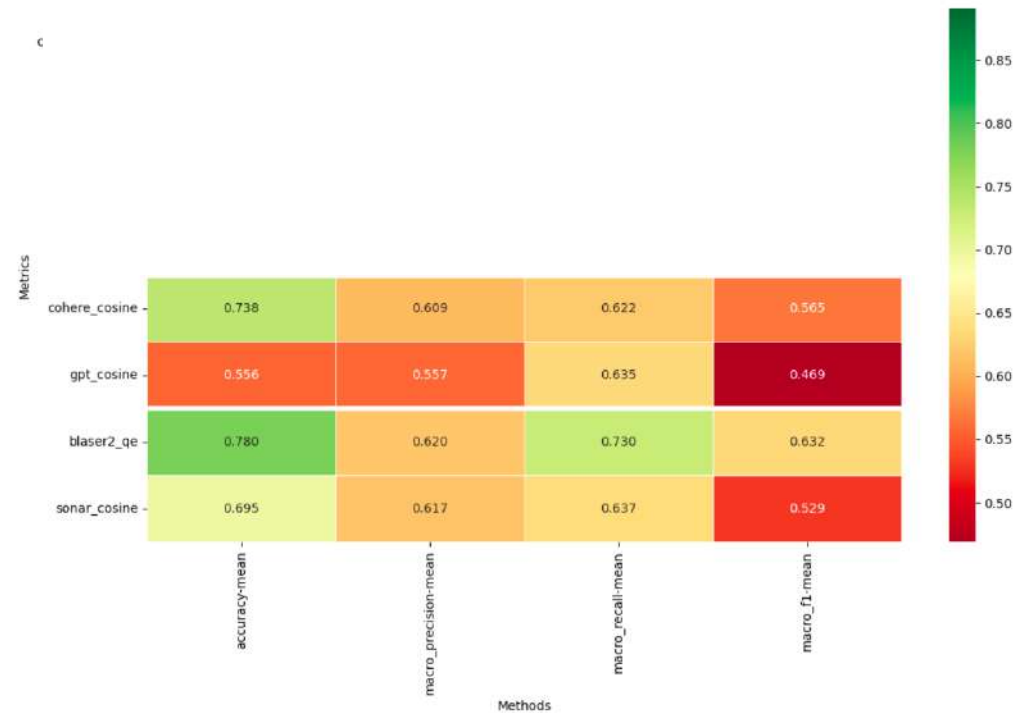


Figure 5: Mean Test scores for LLMs, Embeddings, and *HalOmi* best performing methods.

Conclusion: Embeddings are not enough
in order to beat BLASER

What next?

2. SOLVING THE RESEARCH QUESTION

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
Examples	Chain of Thought (CoT)
	No CoT CoT
	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

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

2. SOLVING THE RESEARCH QUESTION

Methodology – LLMs

Building prompting experiments with four Large Language Models

Input	Task introduction	
	G-Eval1	G-Eval2
Examples	Chain of Thought (CoT)	
	No CoT	CoT
	Number of examples	
	0 examples (zero-shot) 10 examples 18 examples	
Label format	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

Labels

Can we automate reasoning text writing?

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

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: I like 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 %

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)

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

2. SOLVING THE RESEARCH QUESTION

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

Task introduction	G-Eval1	Human Evaluation of Machine Translation Systems: Hallucination Evaluation Criteria: Does the translated text contain information completely unrelated to the source text? - {{pos_label}}: there is hallucination. - {{neg_label}}: there is no hallucination. {{eval_steps}}
Chain of Thought (CoT)	Yes	Evaluation Steps: 1. Read the source text and the translated text carefully. 2. To decide whether the translated text contains hallucinations check if the source tokens "correspond" to erroneous target tokens. For each token answer: - Does this source word fall into the common meaning category as this target word? - Does this source word have a semantic connection with this target word? - Can you try to come up with a reasonable theory on how this source word is associated with this target word? 3. If "no" to all the questions above, then hallucination ({{pos_label}}).
Number of examples		Source Text: 'You got some mail too' Translated Text: 'Ivo nã ti gba létã.' Answer: No Hallucination Source Text: 'Once again, sorry' Translated Text: 'Léèkan sí í, e jòó, o toro áforíjì.' Answer: Hallucination ... Source Text: ':::Yes, they must be just the same person!' Translated Text: 'Ó dájú pé enì kan nã ní wón!' Answer: No Hallucination

Order of examples
Starts with a negative value

Label format
String instead of digit

2. SOLVING THE RESEARCH QUESTION

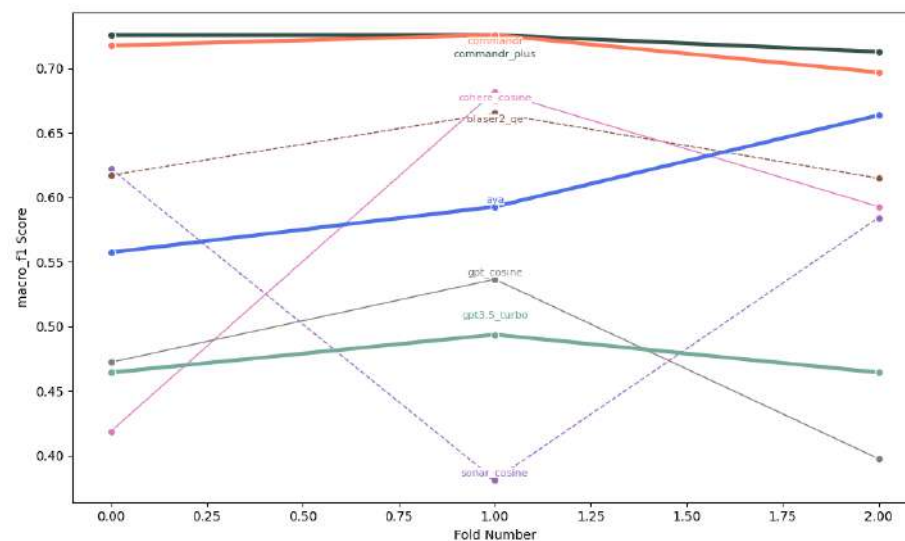
Results


- ✓ 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

Command-R models outperforming from far



Best performance is also the most robust across folds





1. THE COURSEWORK

We beat BLASER*

**binary, English to Yoruba*

2. SOLVING THE RESEARCH QUESTION

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

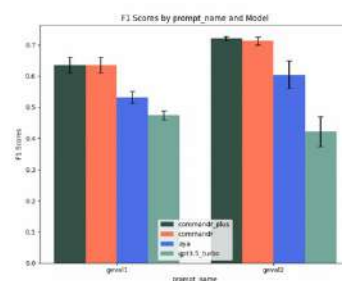
Best performing model for each LLM - accuracy

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

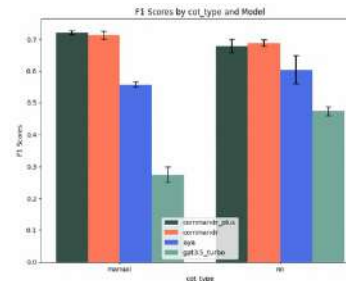


Pro-tip: Find which cases your model performs well and which ones wrong to try to find patterns. It's normally not enough to say something works better, but why?

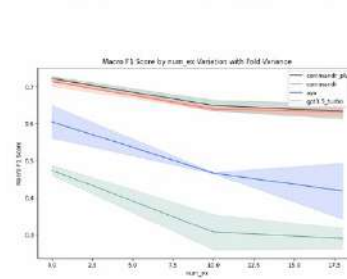
LLMs performances depending on hyperparameters



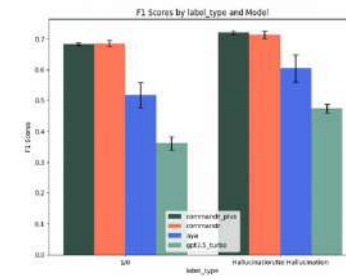
(a) Task Description



(b) Chain of Thought



(c) Number of examples



(d) Format of the label

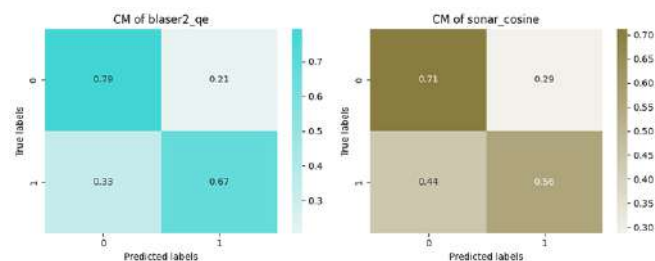
2. SOLVING THE RESEARCH QUESTION

Results

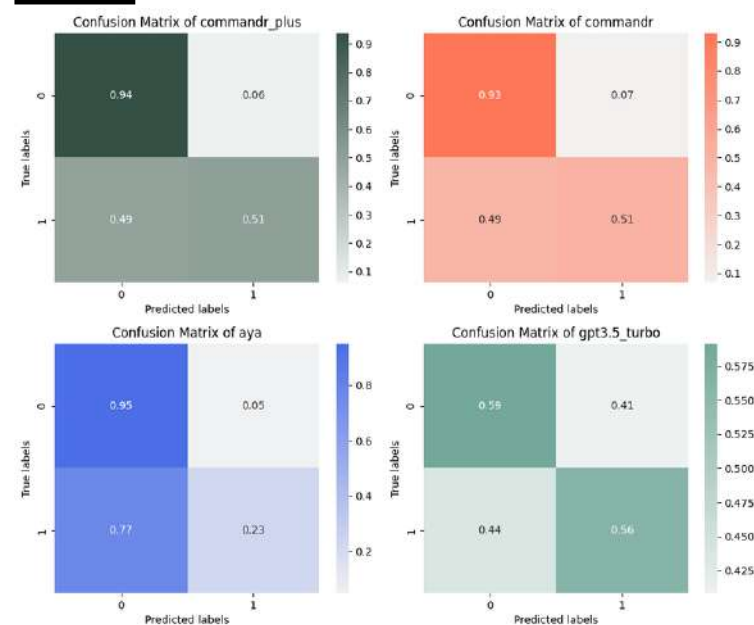
LLMs are the best performing models, beating BLASER

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

Baseline



Confusion Matrices



2. SOLVING THE RESEARCH QUESTION

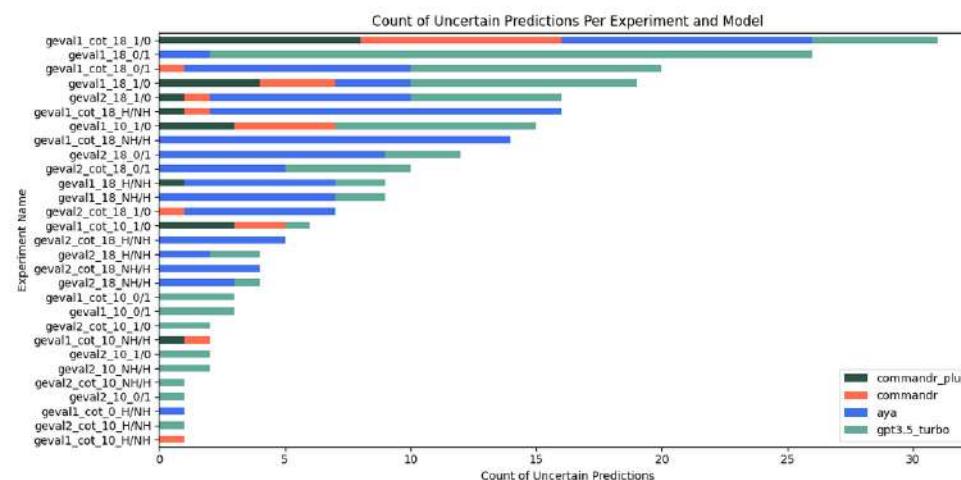
Limits

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



2. SOLVING THE RESEARCH QUESTION

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

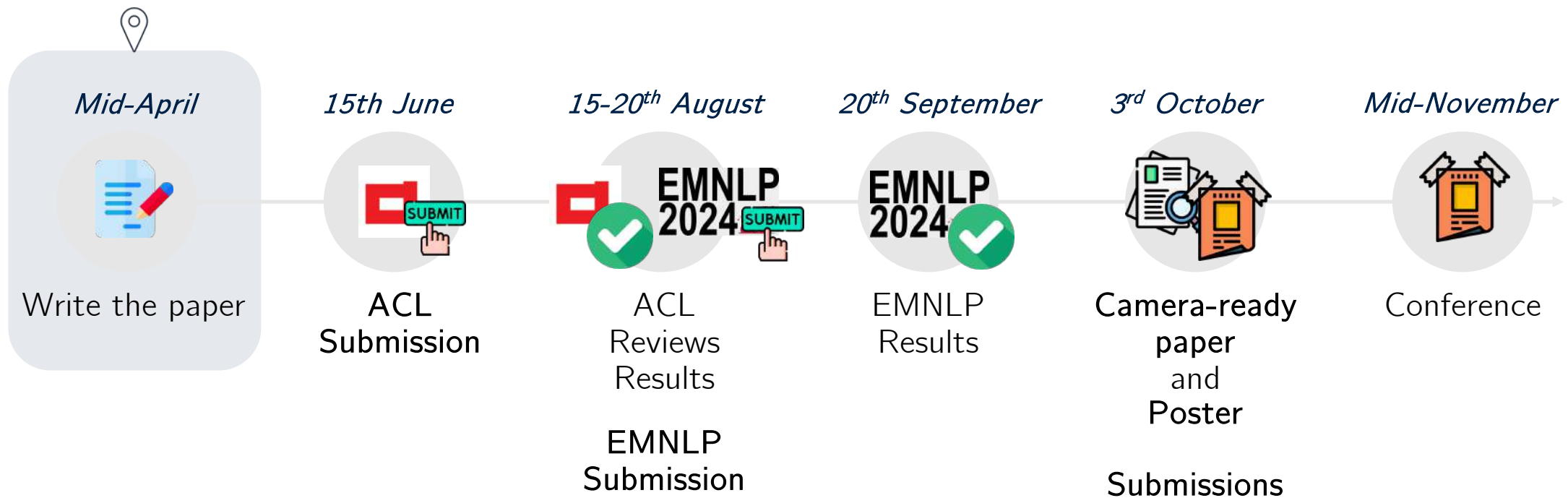


Pro-tip: Prioritise future work to decide how to scale towards a publishable paper

3. WRITING RESEARCH PAPER

3. WRITING THE RESEARCH PAPER

Writing the Coursework



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

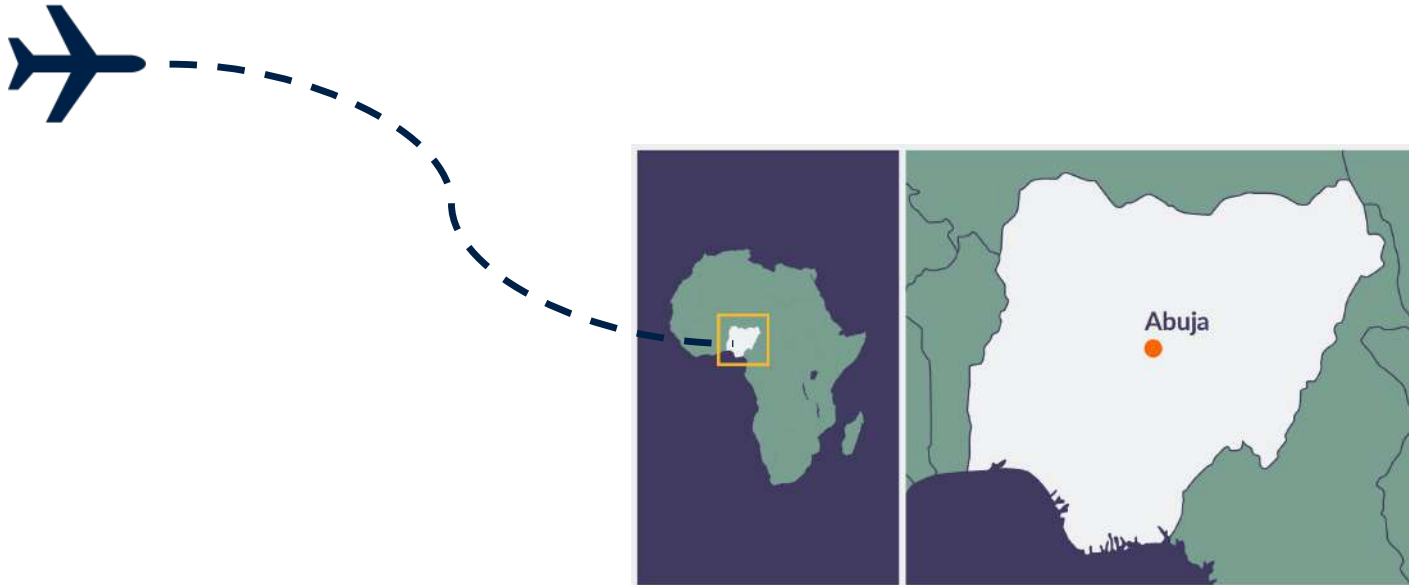
Kenza Benkirane^{1*}, Laura Gongas^{1*},

Shahar Pelles¹, Naomi Fuchs¹, Joshua Darmon¹,
Pontus Stenetorp¹, David Ifeoluwa Adelani^{2,3}, Eduardo Sanchez^{1,4}

¹UCL, ²Mila, Quebec AI institute, ³McGill University, ⁴Meta

*: Equal contributions

First, let's dive into a use case



3. WRITING THE RESEARCH PAPER



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

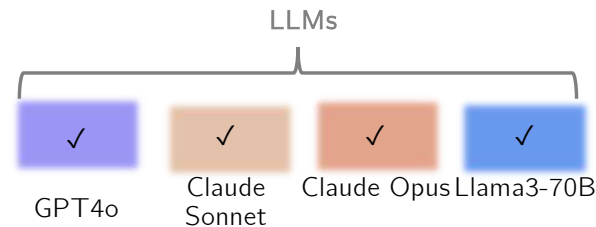
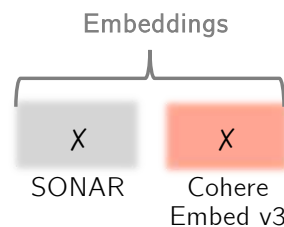
Siempre te quedas sentado con la aerolínea en un vaso.

English Human translation: You always remain seated with the airline in a glass



Ground Truth: **Hallucination**

Correct hallucination classification:



3. WRITING THE RESEARCH PAPER

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 \leftrightarrow (AR, *DE*, RU, ES, ZH)
- Low-Resource Languages (LRL): EN \leftrightarrow (KA, YO, MN)
- Non-English centric: ES \leftrightarrow 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 \leftrightarrow EN) :
301 sentence pairs

Test set:
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 94.44%	2 1.39%	2 1.39%	4 2.78%
AR→EN	156	132 84.62%	5 3.21%	2 1.28%	17 10.90%
EN→RU	146	141 96.58%	1 0.68%	2 1.37%	2 1.37%
RU→EN	158	146 92.41%	3 1.90%	2 1.27%	7 4.43%
EN→ES	153	131 85.62%	8 5.23%	3 1.96%	11 7.19%
ES→EN	160	127 79.38%	17 10.63%	4 2.50%	12 7.50%
EN→ZH	160	131 81.88%	5 3.13%	4 2.50%	20 12.50%
ZH→EN	159	127 79.87%	9 5.66%	7 4.40%	16 10.06%
EN→KA	184	111 60.33%	8 4.35%	30 16.30%	35 19.02%
KA→EN	151	89 58.94%	15 9.93%	32 21.19%	15 9.93%
EN→YO	195	166 85.13%	4 2.05%	11 5.64%	14 7.18%
YO→EN	146	124 84.93%	4 2.74%	10 6.85%	8 5.48%
EN→MN	197	78 39.59%	52 26.40%	54 27.41%	13 6.60%
MN→EN	152	43 28.29%	45 29.61%	58 38.16%	6 3.95%
ES→YO	151	97 64.24%	16 10.60%	29 19.21%	9 5.96%
YO→ES	152	80 52.63%	26 17.11%	37 24.34%	9 5.92%
Total	2564	1859 72.47%	220 8.58%	287 11.19%	198 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

Model	EN→HRL				HRL→EN				EN→LRL			LRL→EN			ES→YO	YO→ES	AVG		
	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.

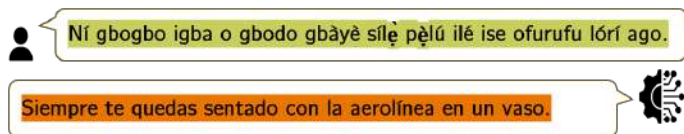
Bold values indicate the best performing prompt per model.

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

Methodology – Embedding spaces

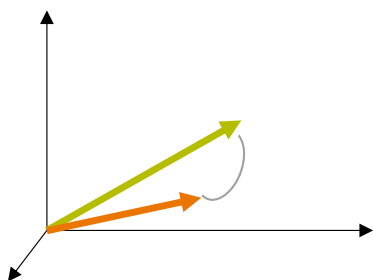
Step 1:

Encoding the source and translation sentences



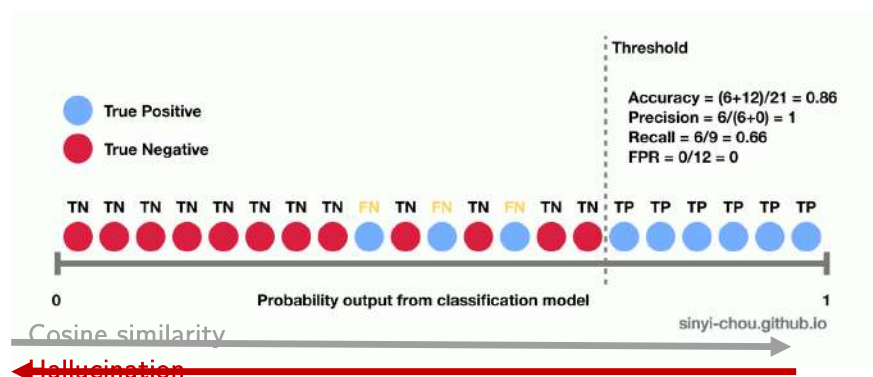
Step 2:

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

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:
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:

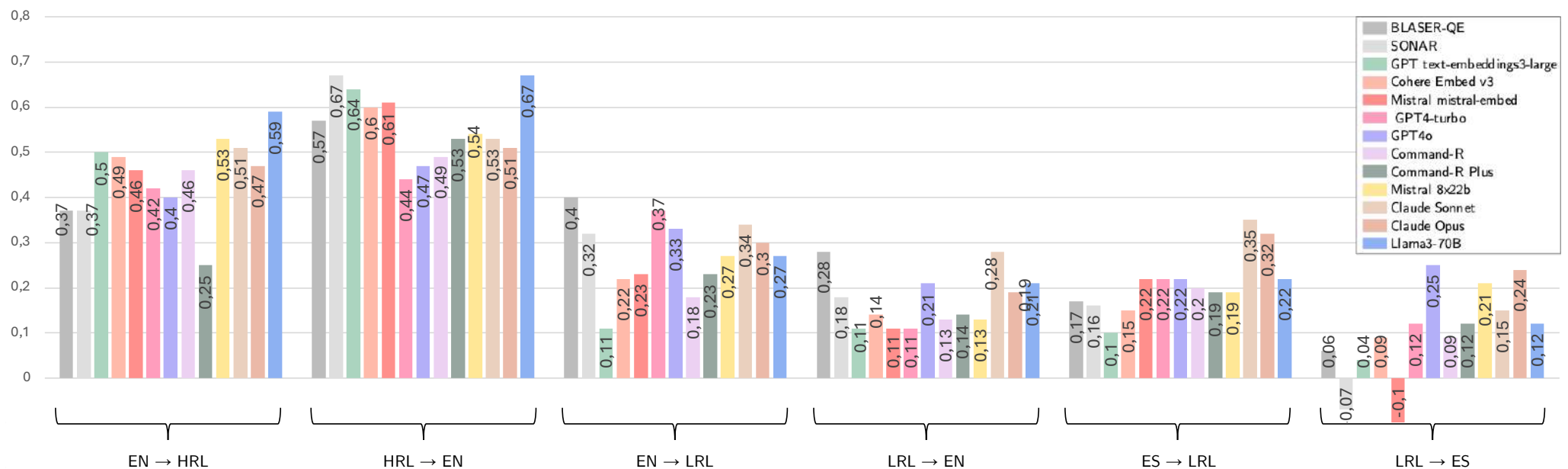
Task descriptions, hallucination definitions, and language-specific instructions

Validation criteria: Best prompt per model selected on highest MCC score

3. WRITING THE RESEARCH PAPER

Overall results

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



Findings

Capabilities

LLMs can outperform
previous models
without being
explicitly trained on

Resources

Embedding spaces
are competitive for

Performance across languages

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

Non- Latin scripts:

Embeddings performed well for HRL→EN

Low-resource languages (LRLs):

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

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)

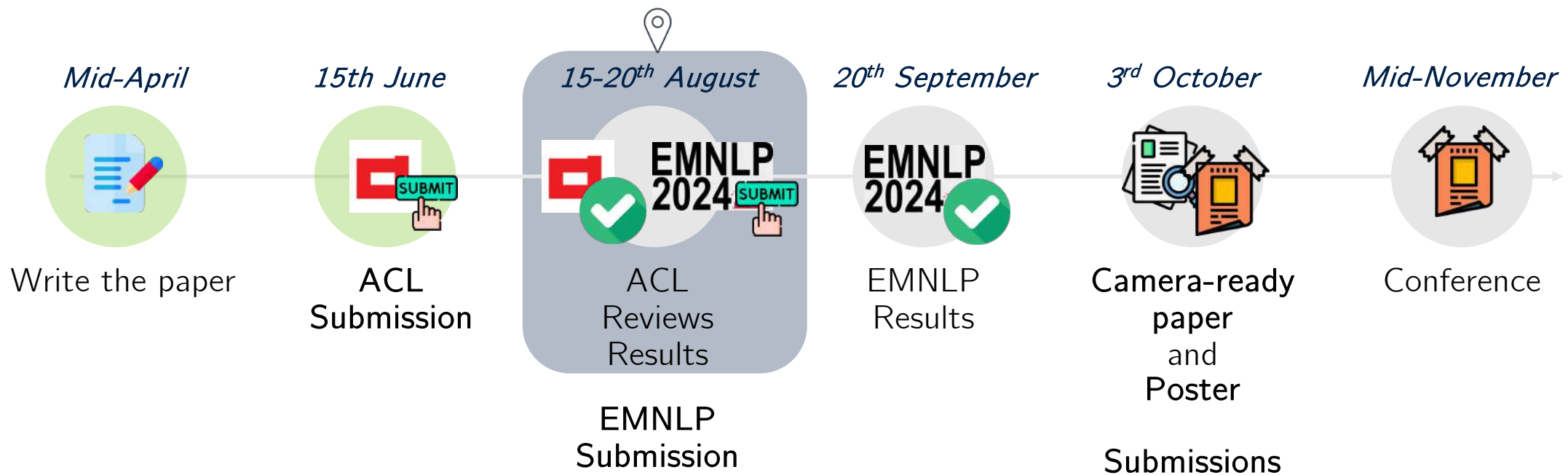


Challenges

- 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



3. WRITING THE RESEARCH PAPER

ACL Rolling Review (ARR)

 **Wikipedia**
https://en.wikipedia.org/wiki/Association_for_Computational_Linguistics

Association for Computational Linguistics

Activities. The ACL organizes several of the top conferences and workshops in the field of computational linguistics and natural language processing. These ...

History · Annual Meeting of the ACL · Activities · Special Interest Groups

#	Submission Summary	Official Review	Decision
3583	Machine Translation Hallucination Detection for Low and High Resource Languages using Large Language Models Download PDF Kenza Benkirane  , Laura Gongas  , Shahar Pelles  Naomi Fuchs  , Joshua Darnon  , Pontus Stenetorp  David Ifeoluwa Adelani  , Eduardo Sánchez  ACL ARR 2024 June Submission Show details	3 Official Reviews Submitted Reviewer g5nr: Overall assessment: 3 / Confidence: 4 Read Official Review Reviewer uheM: Overall assessment: 3.5 / Confidence: 4 Read Official Review Reviewer 2H4i: Overall assessment: 3 / Confidence: 3 Read Official Review Average Overall Assessment: 3.17 (Min: 3, Max: 3.5) Average Confidence: 3.67 (Min: 3, Max: 4)	ACL ARR 2024 June Submission No Recommendation

Step 1: Receive and answer reviews

Official Review of Submission3583 by Reviewer 2H4i
Official Review by Reviewer 2H4i 15 Jul 2024 at 16:45 (modified: 23 Aug 2024 at 00:04)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 2H4i, Commitment Readers, Revisions

Paper Summary:

- This paper studies the effectiveness of large language models in detecting hallucinations across a wide range of high and low resource languages.
- It differs from previous studies by analysing prompting of language models and cosine similarity of embeddings obtained from LLMs as hallucination detectors.
- The results show the effectiveness of prompting LLMs for hallucination detection. They also show a high variability across different language directions, suggesting no single LLM can function as an hallucination detector.

Summary Of Strengths:

- I believe that the findings in this work are relevant to the community working on hallucinations in NMT models, opening an avenue for leveraging zero-shot capabilities of LLMs for hallucination detection, a task for which data is very scarce.
- The authors perform extensive evaluation to support their claims, considering multiple models across many language pairs.

Summary Of Weaknesses:

- I believe the paper would benefit from more fine-grained analysis on based on severity rankings to complement the binary detection setting. I find the explanation that class imbalance hurts model performance an insufficient justification to leave out that analysis, as it would offer more insights regarding the strengths/weaknesses of LLMs in hallucination detection.

Comments Suggestions And Typoes:

- It would also strengthen the paper to further explore the versatility of LLMs. For example, it would be interesting to explore whether LLMs can identify the hallucinated spans (similar to fine-grained error span detection <https://arxiv.org/abs/2023.06.01>), or if they can correct them as in automatic post-edition.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 3.5

Overall Assessment: 3 = Good. This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Best Paper: No

Ethical Concerns: None

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources.

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources.

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources.

Add: [Author-Editor Confidential Comment](#)

Step 2: Receive the meta-review

Meta Review of Submission3583 by Area Chair TNT8
Meta Review by Area Chair TNT8 05 Aug 2024 at 15:00 (modified: 23 Aug 2024 at 00:04)
Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Metareview:

This paper investigates how well LLMs can detect hallucinations in machine translation output. They use language pairs. LLMs are shown to be very effective hallucination detector, but selecting the best-performing model is challenging.

Summary Of Reasons To Publish:

- The evaluation is extensive and convincing, testing many models and language pairs.
- Reviewers noted that the paper is clearly written, and suggest few if any corrections.

Summary Of Suggested Revisions:

- Implement comments and suggestions from eheM (and fix the 3rd weakness) and reviewer g5nr.
- Could you comment on the practicality of the approach? (related to the last comment from g5nr, why is it important to test them on MT hallucination detection?)

Overall Assessment: 4 = There are minor points that may be revised

Suggested Venues: EMNLP

Best Paper Ae: No

Ethical Concerns: There are no concerns with this submission

Needs Ethics Review: No

Author Identity Guess: 1 = I do not have even an educated guess about author identity.

- ✓ Receive 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 assessment

3. WRITING THE RESEARCH PAPER

Paper accepted!

3. WRITING THE RESEARCH PAPER

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
WiNLP
Natural Language Processing

EMNLP
2024

am Registration Participants Info Calls Venue Blog Committees Spons

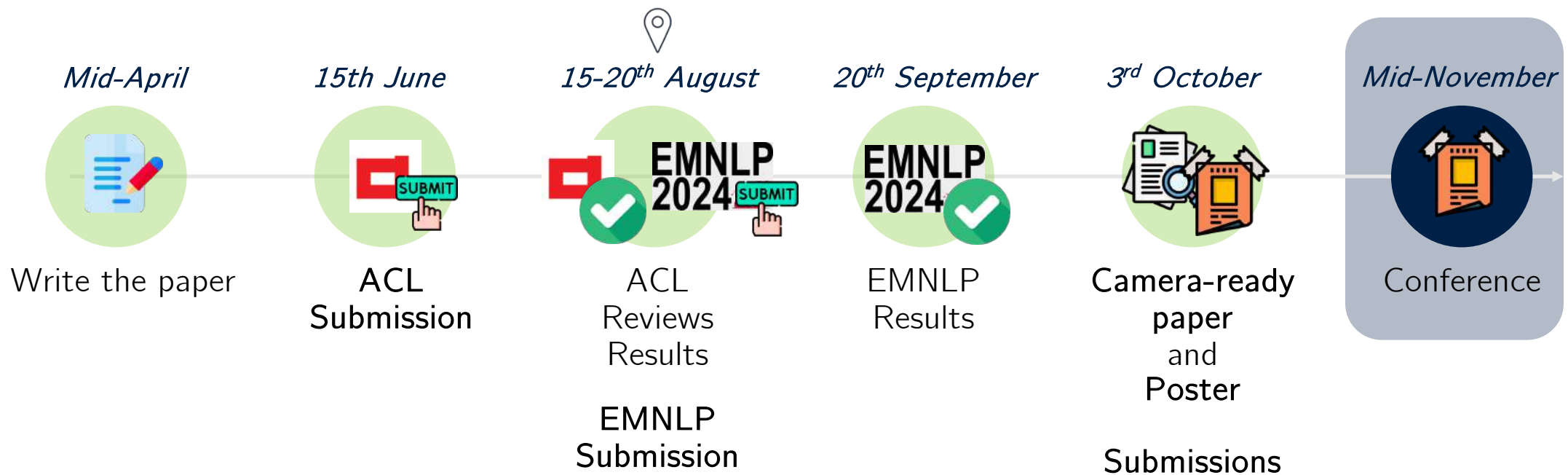
CALL FOR

Diversity & Inclusion Subsidies
Main Conference Papers
System Demonstrations
Industry Track

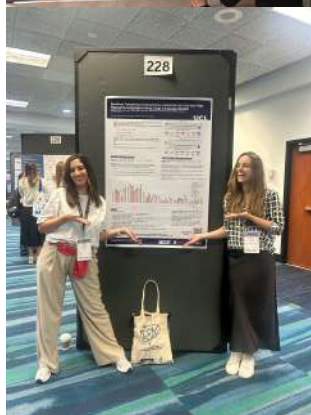
Call for EMNLP 2024 Diversity and Inclusion Subsidies

3. WRITING THE RESEARCH PAPER

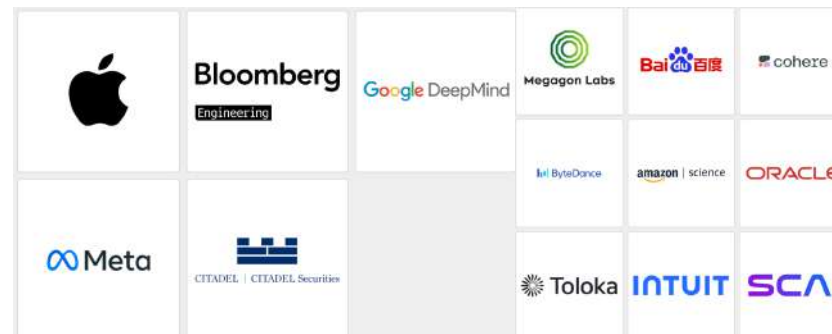
Writing the Coursework



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

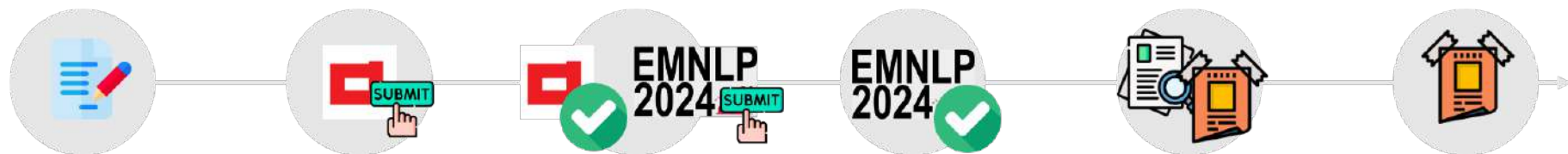
3. PRACTICAL TIPS



What have we learned and how you can use it

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

Writing the Coursework



Learning

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 prepared 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

3. PRACTICAL TIPS

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



 substack

For Developers: Social Platforms



X / Twitter

@GoogleAI, @GoogleHealth, @GoogleDeepMind

@OpenAI @AnthropicAI

@MistralAILabs @Cohere

@dwarkesh_sp

@AIBreakfast

@trending_repos

@DAIR_AI

@DeepLearningAI

@karpathy



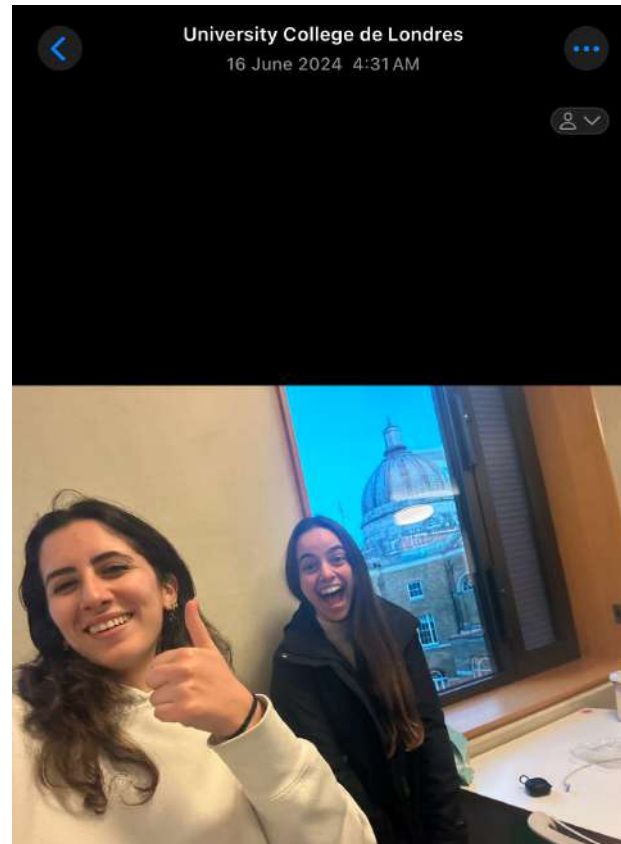
Discord

Mistral, Perplexity, Google Developers, Cohere, OpenAI, HuggingFace,

All have public discord workspaces with links available on their website.

3. PRACTICAL TIPS

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 Pelles¹, Naomi Fuchs¹, Joshua Darmon¹,
Pontus Stenetorp¹, David Ifeoluwa Adelan^{2,3}, Eduardo Sanchez^{1,4}*

**: Equal contributions*