

# **GPTs Are Multilingual Annotators for Sequence Generation Tasks**

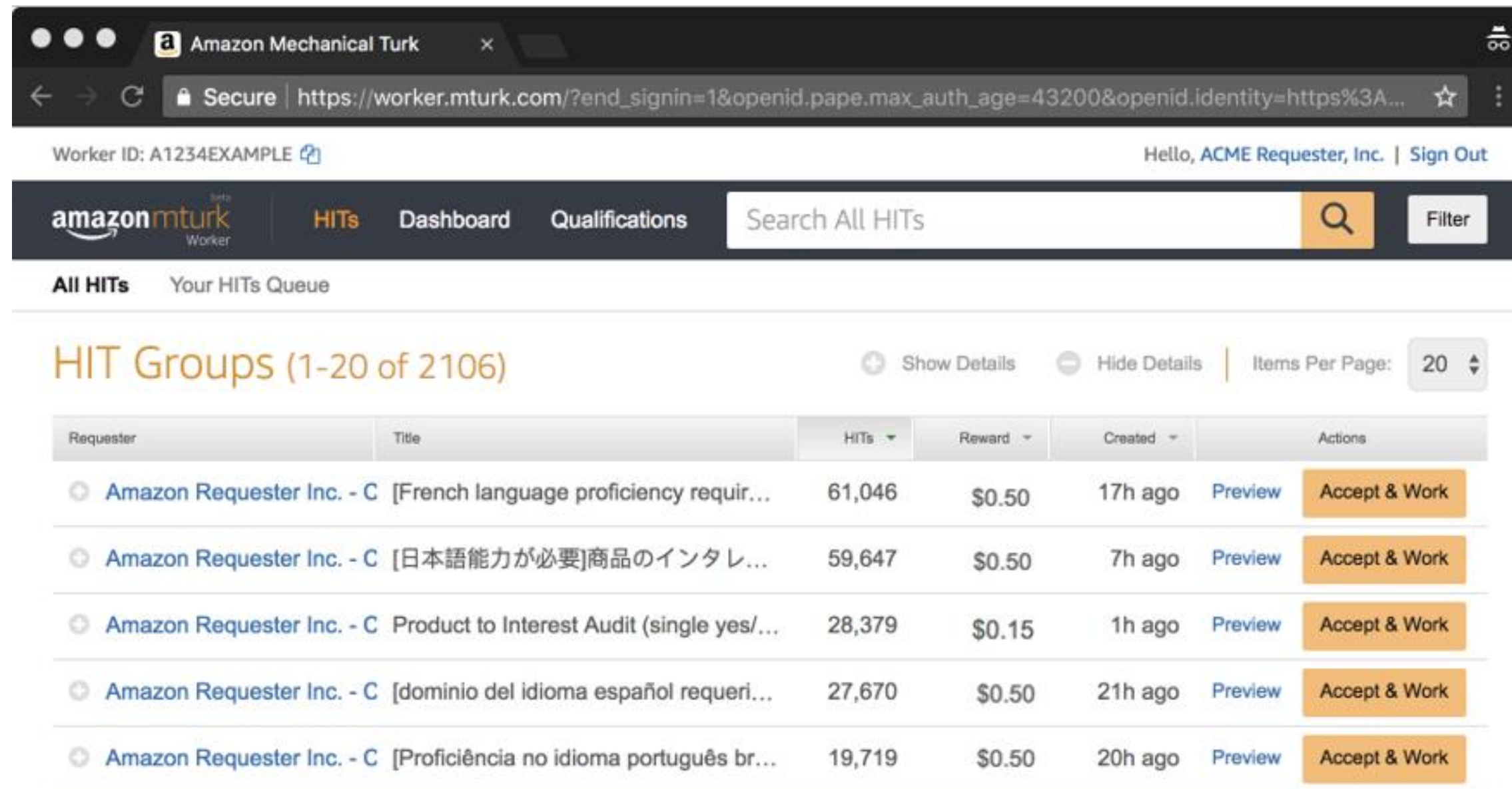
Juhwan Choi, Eunju Lee, Kyohoon Jin, and Youngbin Kim

Chung-Ang University

# Importance of Data Annotation

**Data annotation** is crucial for construction of new dataset

- In general, this process is performed by **human annotators** (i.e., crowdworkers)
- Platforms such as Amazon Mechanical Turk (MTurk) is mostly used for this purpose



The screenshot shows the Amazon Mechanical Turk worker interface. At the top, the browser address bar displays the URL: [https://worker.mturk.com/?end\\_signin=1&openid.pape.max\\_auth\\_age=43200&openid.identity=https%3A...](https://worker.mturk.com/?end_signin=1&openid.pape.max_auth_age=43200&openid.identity=https%3A...). Below the browser bar, the page header includes the Worker ID: A1234EXAMPLE, the greeting "Hello, ACME Requester, Inc. | Sign Out", and navigation links for "HITs", "Dashboard", and "Qualifications". A search bar labeled "Search All HITs" and a "Filter" button are also present.

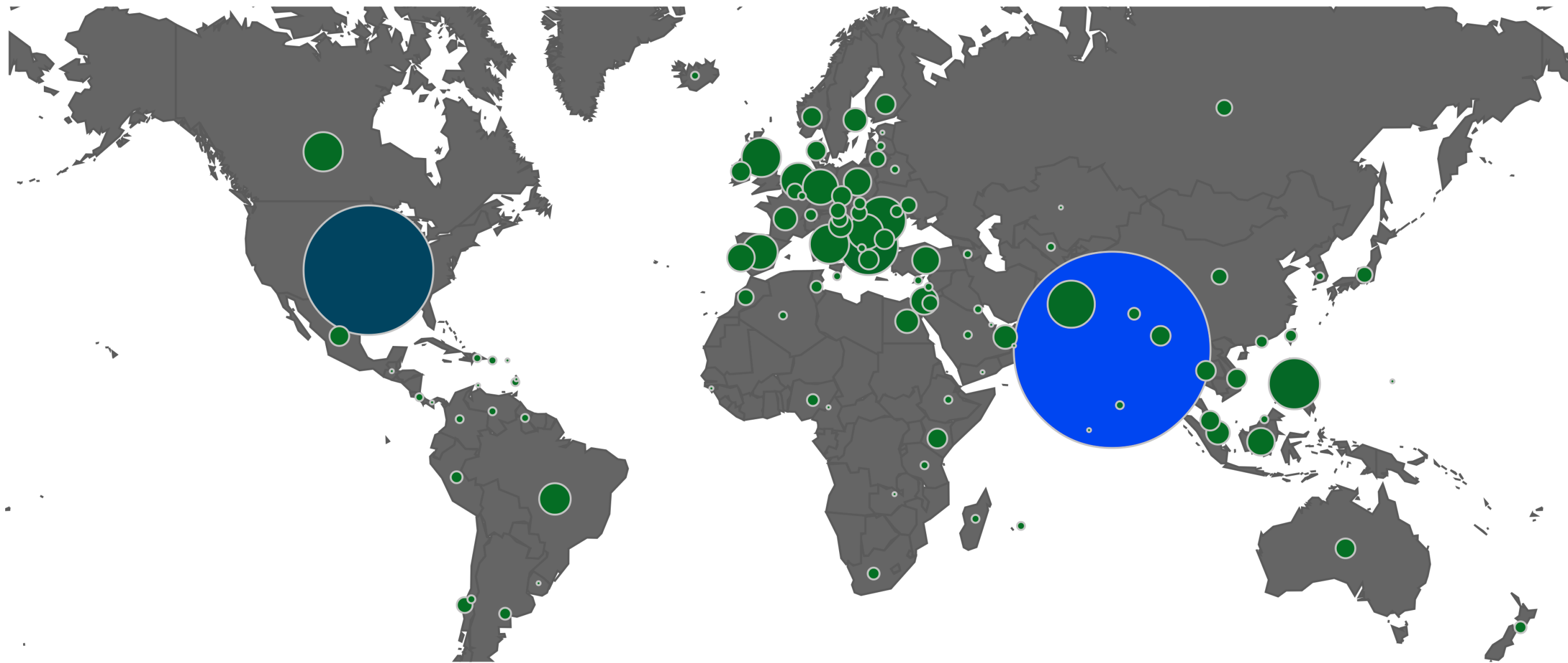
The main content area displays "HIT Groups (1-20 of 2106)". Above the table, there are controls for "Show Details" (plus icon), "Hide Details" (minus icon), and "Items Per Page: 20".

Requester	Title	HITs	Reward	Created	Actions	
Amazon Requester Inc. - C	[French language proficiency requir...	61,046	\$0.50	17h ago	<a href="#">Preview</a>	<a href="#">Accept &amp; Work</a>
Amazon Requester Inc. - C	[日本語能力が必要]商品のインタレ...	59,647	\$0.50	7h ago	<a href="#">Preview</a>	<a href="#">Accept &amp; Work</a>
Amazon Requester Inc. - C	Product to Interest Audit (single yes/...	28,379	\$0.15	1h ago	<a href="#">Preview</a>	<a href="#">Accept &amp; Work</a>
Amazon Requester Inc. - C	[dominio del idioma español requeri...	27,670	\$0.50	21h ago	<a href="#">Preview</a>	<a href="#">Accept &amp; Work</a>
Amazon Requester Inc. - C	[Proficiência no idioma português br...	19,719	\$0.50	20h ago	<a href="#">Preview</a>	<a href="#">Accept &amp; Work</a>

# Demographics of MTurk

However, there is a difference in the number of the speaker of each language<sup>1</sup>

- In other words, it could be **difficult to hire non-English annotators**



1. Pavlick et al., [The Language Demographics of Amazon Mechanical Turk](#), TACL 2014.

# What about Low-resource Languages?

Due to the limited language pool, it is especially challenging to hire annotators for low-resource languages

- This makes **dataset construction in the low-resource language expensive**
- As a result, it may discourage the construction of new dataset in low-resource languages

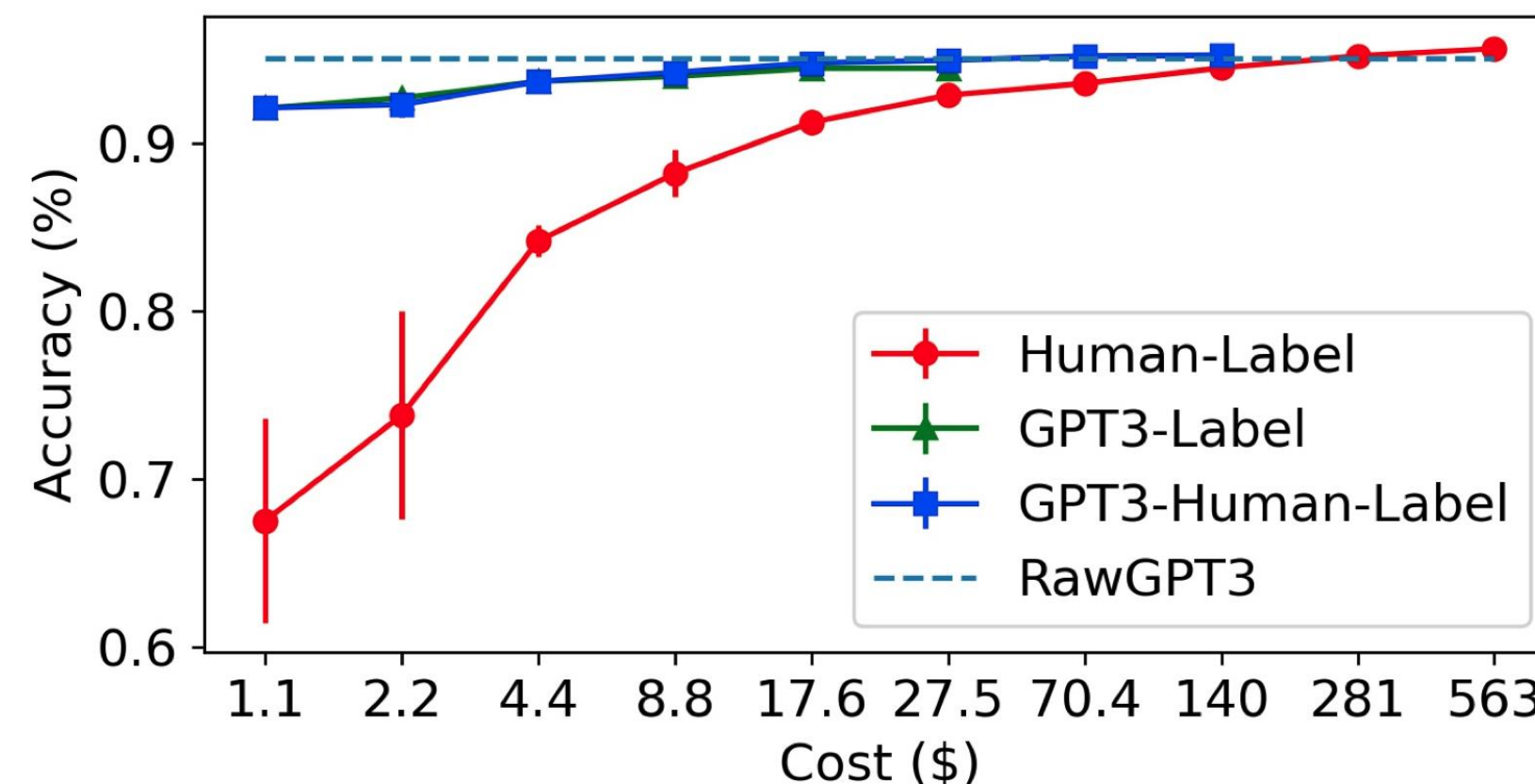
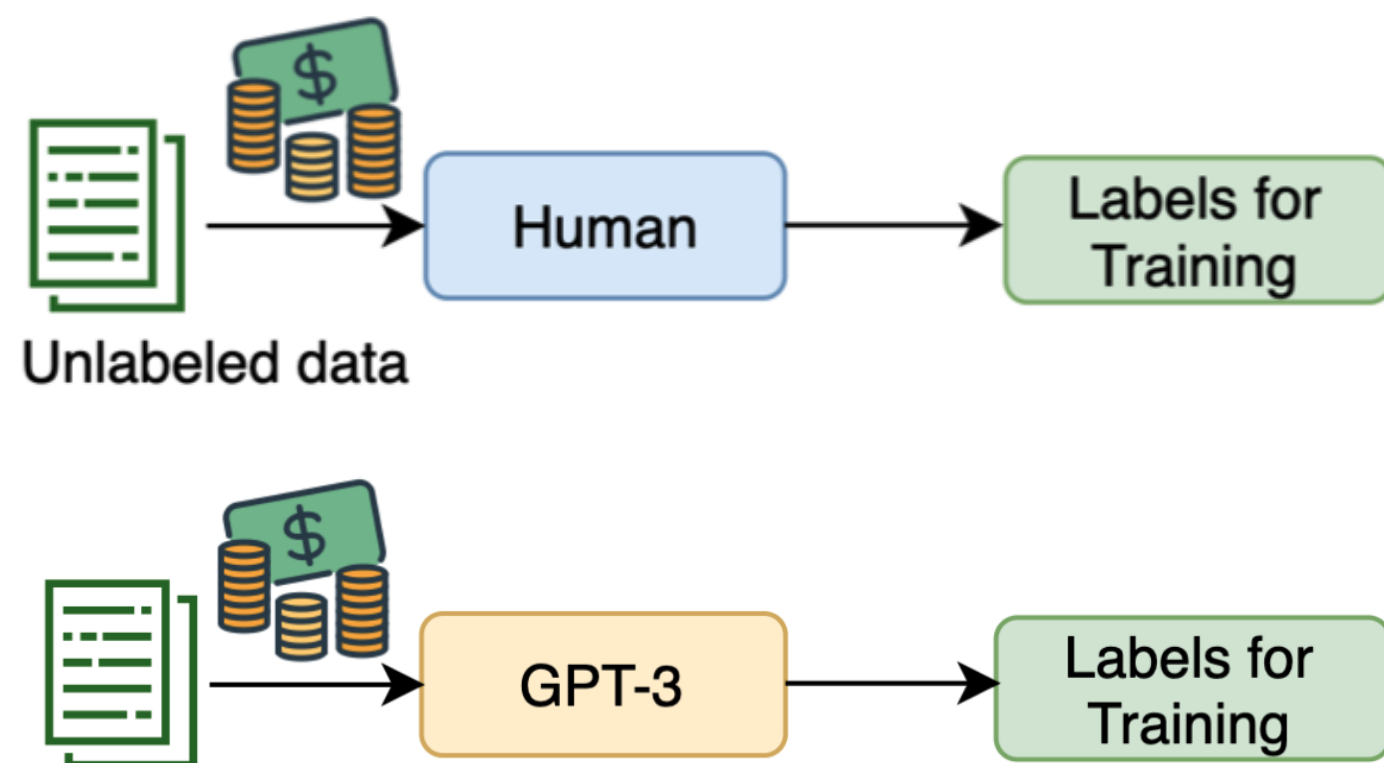




# Data Annotation with Large Language Model

Recently, researchers are exploring methods to apply LLMs for data annotation

- In the perspective of the cost, previous research has found that **LLM-based annotation is more cost-efficient**<sup>1</sup>
- This makes LLM-based data annotation more **effective where the budget is limited**



# Data Annotation with Large Language Model

LLM-based data annotation is achieved by few-shot prompting

- Well-designed prompt could assist to annotate unlabeled data<sup>1</sup>
- Previous studies primarily **focused on simple tasks such as text classification**
- Additionally, LLM-based data annotation is **less explored in languages other than English**

**Choose the sentiment of the given text from Positive and Negative.**

**Text:** a feast for the eyes

**Sentiment:** Positive

...

**Text:** boring and obvious

**Sentiment:** Negative

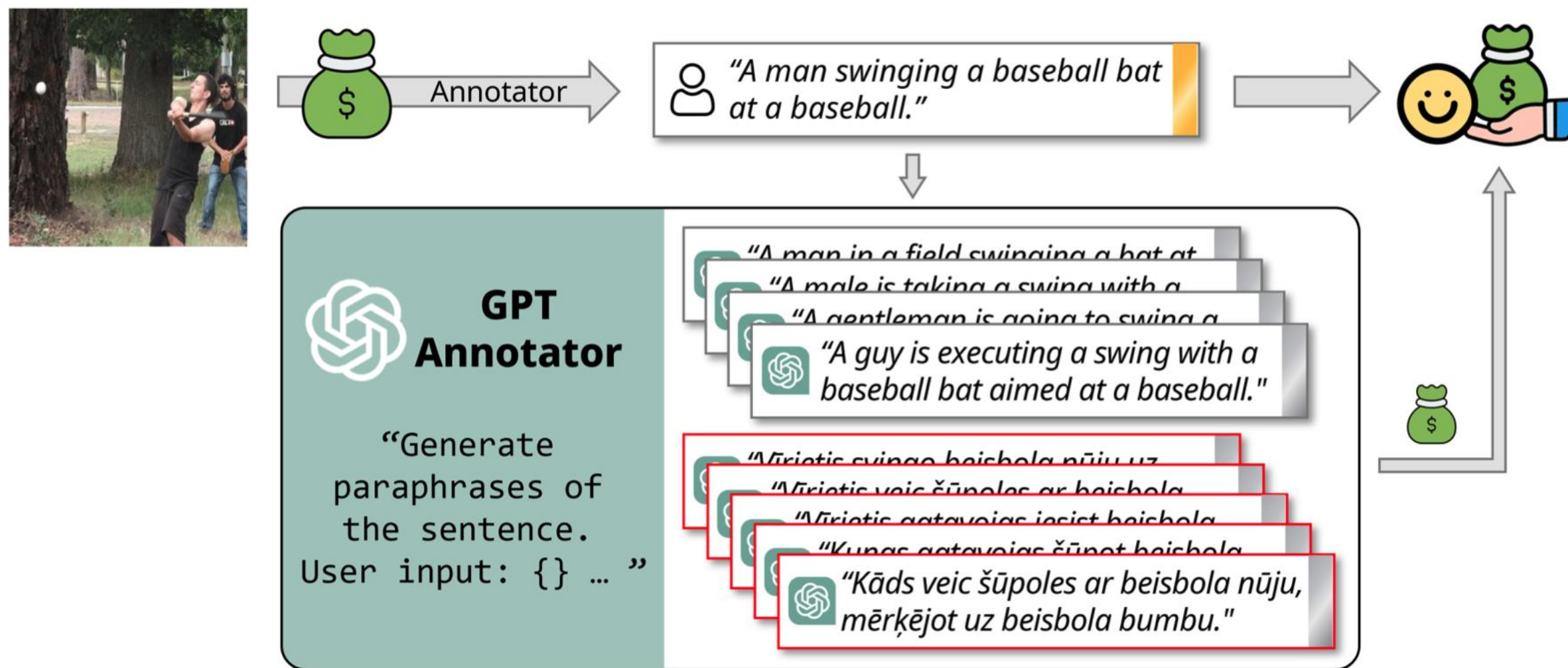
**Text:** [Unlabeled Data]

**Sentiment:** [Label]

# GPTs as Multilingual Data Annotators

In this study, we explored LLM-based data annotation in:

- **Generation tasks** — Image captioning and Text style transfer
- **Multilingual setting** — Various languages including low-resource languages
- We aim to establish LLM-based annotation as an **assistant** annotator to human annotator in this setup



# Two Sequence Generation Tasks

We established two different text generation tasks for our method

- **Image captioning** — Generating a description of given image
- **Text style transfer** — Altering the writing style of given text
- These tasks are more complex and costly than classification task

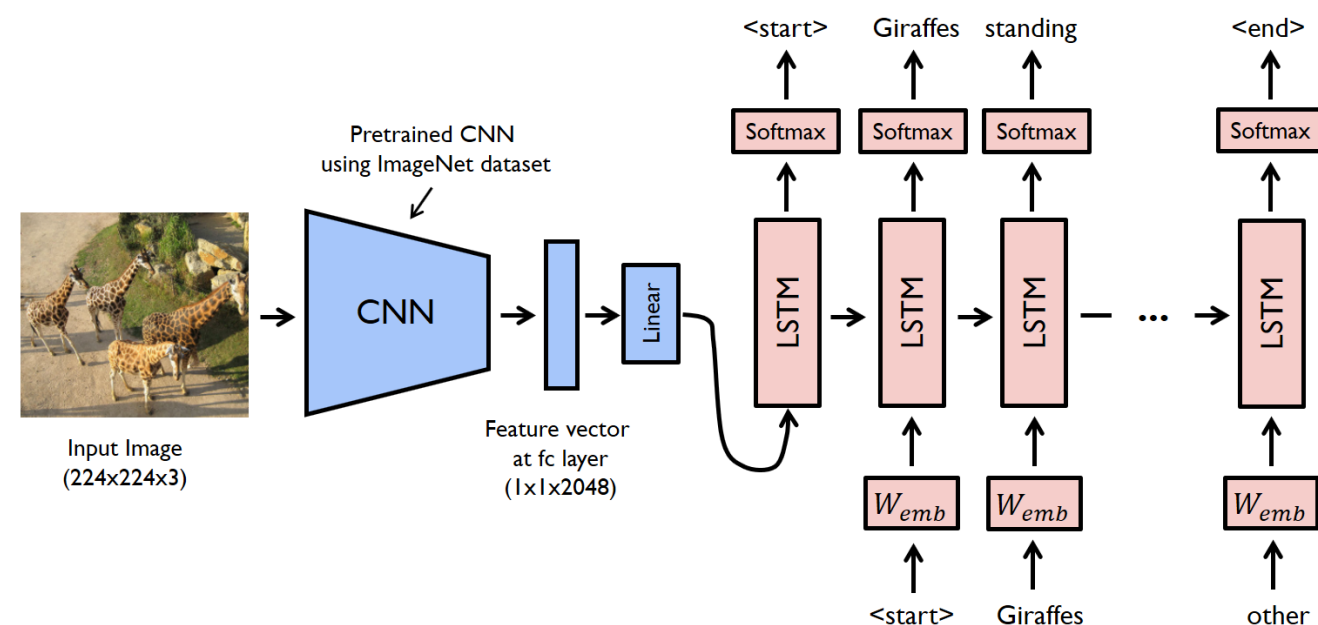


Image Captioning

---

Informal: *I'd say it is punk though.*  
Formal: *However, I do believe it to be punk.*

---

Informal: *Gotta see both sides of the story.*  
Formal: *You have to consider both sides of the story.*

---

Text Style Transfer



# Methodology — Image Captioning

We designed a dedicated prompt for multilingual data annotation. It consists of two steps:

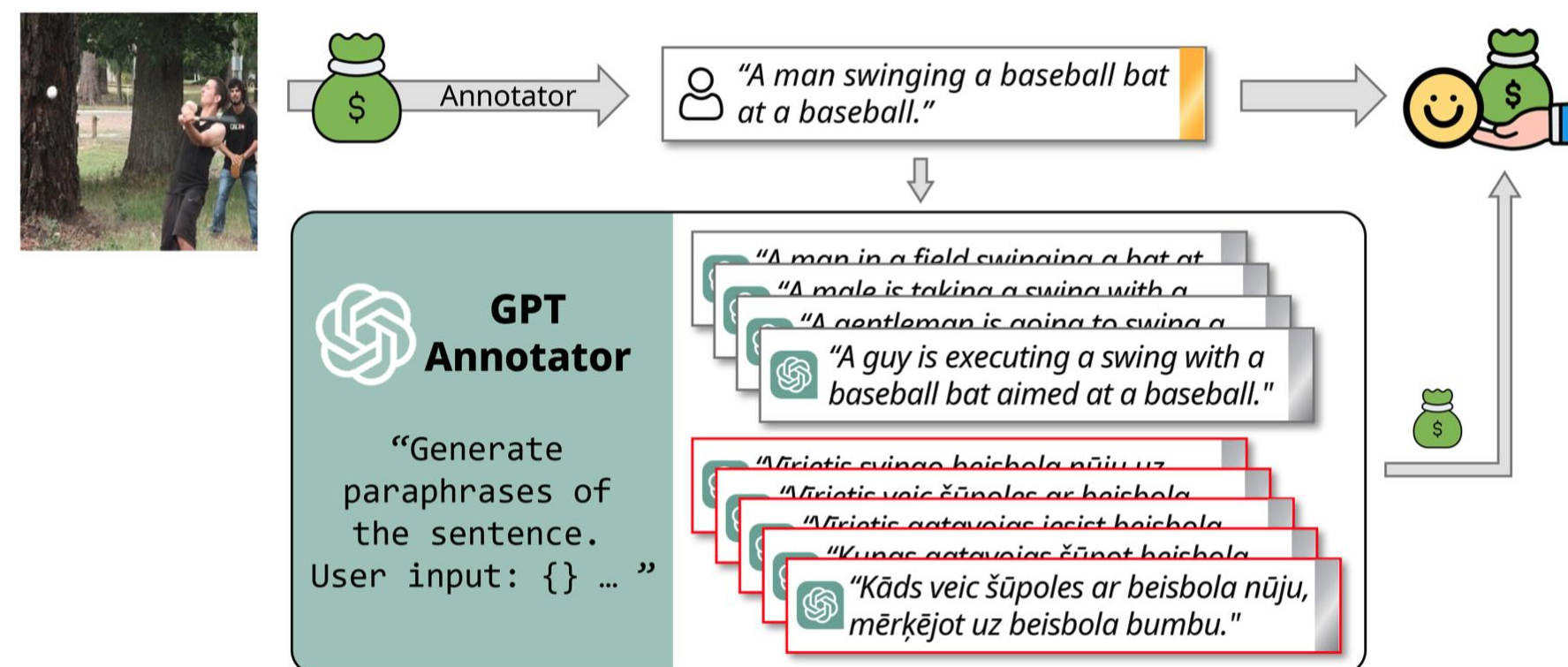
## 1. Paraphrasing

- We generate four paraphrases of given sentence
- This allows to construct an image captioning dataset with only one human annotation, lowering the required amount of human-annotated data

## 2. Translation

- We then translate these sentences into designated language
- This separation makes the annotation procedure more distinct

$$Y_{tgt} = \{y_{s_1}^{tgt}, \dots, y_{s_5}^{tgt}\} = \text{GPT}(P^{tgt}, y_{g_1}^{eng})$$



# Example Prompt — Image Captioning

We instruct GPT to follow the guideline with one-shot example as shown below:

---

System

You are a helpful assistant.

User will ask you to generate paraphrases of a sentence.

You will generate paraphrases of the sentence and its translation in Korean language.

VERY IMPORTANT: You must speak ‘-하다’ form in Korean. You must not use ‘-합니다’ or other forms. 한국어 문장을 번역하여 생성할 때, 반드시 ‘-하다’ 체를 사용하여야 한다. ‘-합니다’, ‘-입니다’ 등의 표현을 절대 사용하지 않는다.

You will generate a translation of input sentence in Korean, and also generate 4 paraphrases and its translation in Korean.

Output sentence should be neutral expression. You should not generate phrases like ‘You will see’ or ‘You will find’.

Output sentence will be complete, natural and fluent.

Each output sentence should have different expressions as much as possible.

You will not generate the same sentence as the input sentence.

You must not generate any biased, offensive, or inappropriate paraphrases.

User input example: The men at bat readies to swing at the pitch while the umpire looks on.

Your output example:

Translation: 타석에 있는 남자들이 심판이 지켜보는 동안 스윙할 준비를 한다.

Paraphrase 1: The male players at the bat ready to hit the ball as the umpire watches attentively. / 심판이 주의 깊게 지켜보는 가운데 배트를 든 남자 선수들이 공을 칠 준비를 하고 있다.

Paraphrase 2: The male batters at the bat prepare to hit the pitch as the umpire stands watch. / 타석에 선 남성 타자들이 심판이 지켜보는 가운데 타구를 칠 준비를 하고 있다.

Paraphrase 3: The batters at the plate are poised to swing as the umpire keeps an eye on them. / 타석에 있는 타자가 심판이 지켜보는 가운데 스윙할 자세를 취한다.

Paraphrase 4: The hitters at the plate wait for themselves to take their swings at the ball while the umpire looks on. / 타석에 선 타자들은 심판이 지켜보는 동안 공을 향해 스윙할 준비를 한다.

You will not say ‘Sure! here’s the output’ or any similar phrases.

You will not say ‘I don’t know’ or any similar phrases.

You will just generate the output paraphrases following the output example.

User

Input: Living room with furniture with garage door at one end.

---

# Methodology — Text Style Transfer

Similarly, we designed a dedicated prompt for text style transfer task

- We generate two set of formal and informal text from one set of human-annotated data, **lowering the required amount of human-annotated data**

---

System

You are a helpful assistant. You are fluent in French and English.

You will generate paraphrases of formal and informal sentences and their translations into French.

Output sentence should be neutral expression.

Output sentence will be complete, natural and fluent.

Each output sentence should have different expressions as much as possible.

You will not generate the same sentence as the input sentence.

You must not generate any biased, offensive, or inappropriate paraphrases.

You will not say 'Sure! here's the output' or any similar phrases.

You will not say 'I don't know' or any similar phrases.

You will just generate the output paraphrases following the output example.

[Input Sentence]

Formal 1: Then kiss her, brother; that works every time.

Informal 1: Then kiss her;) works every time bro!!!!

[Paraphrase]

Formal 2: Subsequently, kiss her, sibling; that method proves effective on each occasion.

Informal 2: So, just give her a smooch, bro! It seriously works every single time ;)

[Translation in French]

Formal 1: Alors embrasse-la, mon frère. Cela fonctionne à chaque fois.

Informal 1: Alors embrasse-la ;) ça marche à chaque fois frérot!!!!

Formal 2: Ensuite, embrasse-la, frère ; cette méthode fonctionne à chaque fois.

Informal 2: Alors, donne-lui un bisou, mec ! Ça marche à tous les coups ;)

User

[Input Sentence]

Formal 1: After that I never bought her another gift.

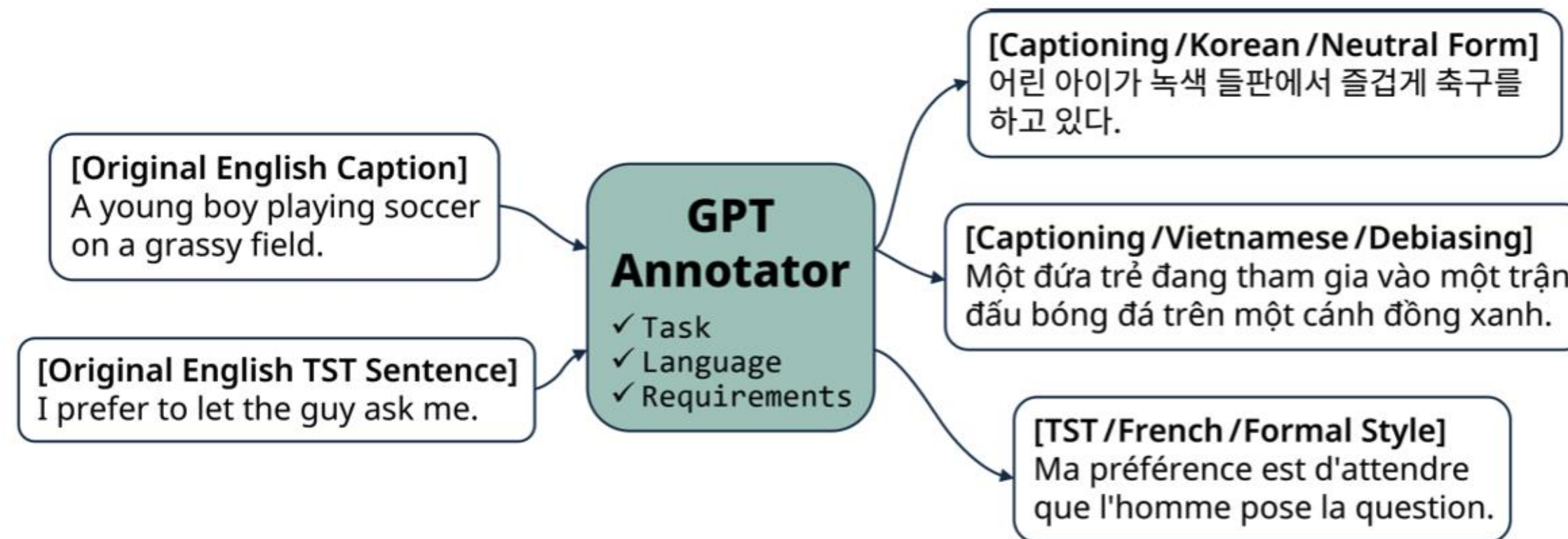
Informal 1: and enver since then i never bought her another gift

---

# Assistant Annotators with Various Benefits

This approach based on prompt design offers several benefits:

- We can easily extend the proposed approach to other tasks and languages
- We can instruct them to consider **special considerations**
  - For instance, we can instruct them to generate sentence in neutral form





# Experimental Result — Cost-Efficiency

We first validated the **cost-efficiency of proposed method**

- We constructed an English image captioning with one gold caption and four silver caption per each image with proposed method
- Based on the cost of the proposed method, we constructed another subset of the dataset with full human annotation within a limited budget
- Additionally, we build another baseline methods such as paraphrase generation<sup>1</sup> and Back-Translation<sup>2</sup>
- The experimental result suggests that the model trained with our method showcases better performance compared to other baselines
  - This indicates that **our method is cost-efficient**

Flickr8k	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Human Annotator w/ Limited Budget	28.96	38.76	17.83	0.7817	-18.379
Synonym Replacement	30.30	38.61	17.61	0.7802	-18.457
Back-Translation	30.02	39.02	17.32	0.7795	-18.413
HRQ-VAE	21.62	29.53	15.83	0.7542	-18.641
GPT Annotator w/ GPT-3.5	33.13	39.98	18.41	0.7892	-18.374

1. Hosking et al., [Hierarchical Sketch Induction for Paraphrase Generation](#), ACL 2022

2. Sennrich et al., [Improving Neural Machine Translation Models with Monolingual Data](#), ACL 2016

# Experimental Result — Multilingual Experiment

We validated the **multilingual annotation ability** of the proposed method

- We compared our method with translation baselines such as NLLB and Google Translator
- Additionally, we also used paraphrase generation model and translating them
- We utilized three image captioning datasets in different languages:
  - Vietnamese, Polish, and Korean
  - For Korean, we performed human evaluation
- Experimental results suggest that **our method is versatile to various languages**

<b>Vietnamese</b>	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Original (Human-Annotated)	48.62	53.82	32.16	0.8309	-14.511
NLLB (Machine-Translated)	31.76	40.49	26.61	0.8114	-14.645
HRQ-VAE + NLLB	21.26	28.64	23.48	0.7720	-15.342
Google Translator	37.22	46.24	26.86	0.8196	-14.534
GPT Annotator w/ GPT-4	41.32	47.83	30.57	0.8235	-14.537

<b>Polish</b>	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Original (Human-Annotated)	8.68	19.38	9.38	0.7405	-18.162
NLLB (Machine-Translated)	4.14	14.46	6.78	0.6466	-18.279
HRQ-VAE + NLLB	3.21	13.15	5.99	0.6495	-18.331
Google Translator	4.64	14.14	6.91	0.6507	-18.244
GPT Annotator w/ GPT-4	5.17	18.90	8.92	0.6962	-18.197

<b>Korean</b>	Precision↑	Recall↑	Fluency↓	THUMB↑
AiHub (Machine-Translated)	4.3	4.09	0.03	4.17
GPT Annotator w/ GPT-4	4.72	4.59	0.02	4.64

# Experimental Result — Text Style Transfer

We extended our experiment to **text style transfer tasks**

- In this experiment, we used French, Brazilian Portuguese, and Italian
- Experimental results suggests that our method is superior than translation baselines
- Especially, our method is particularly superior in terms of the formality of generated sentences
  - This is based on the instruction that guides to generate sentences in designated style

<b>French</b>	BLEU	ROUGE	METEOR	BERTS.	BARTS.	Formality
NLLB (Machine-Translated)	48.59	50.26	31.42	0.8103	-17.596	72.37
Google Translator	51.69	54.02	32.62	0.8076	-17.541	75.38
GPT Annotator w/ GPT-4	54.81	56.83	33.98	0.8175	-17.519	<b>85.12</b>
<b>Brazilian Portuguese</b>	BLEU	ROUGE	METEOR	BERTS.	BARTS.	Formality
NLLB (Machine-Translated)	52.73	55.81	32.44	0.8286	-18.955	68.58
Google Translator	55.98	57.74	34.19	0.8318	-18.938	74.27
GPT Annotator w/ GPT-4	57.94	60.72	35.60	0.8363	-18.864	<b>79.21</b>
<b>Italian</b>	BLEU	ROUGE	METEOR	BERTS.	BARTS.	Formality
NLLB (Machine-Translated)	47.97	49.34	30.12	0.7839	-18.843	68.03
Google Translator	49.13	51.73	30.89	0.7873	-18.805	71.86
GPT Annotator w/ GPT-4	52.34	53.71	32.02	0.7994	-18.702	<b>74.29</b>

# Dataset Construction

Based on the experimental result, **we built image captioning dataset in three languages:**

- Latvian, Estonian, Finnish — these are **low-resource languages**
- We performed additional experiments to validate the quality of constructed datasets
- The result suggests that **our method is superior than baselines**, as shown in previous experiments

Latvian	BLEU	ROUGE	METEOR	BERTS.	BARTS.
NLLB (Machine-Translated)	6.39	17.53	10.13	0.6803	-16.061
HRQ-VAE + NLLB	5.14	16.61	10.21	0.6728	-16.127
Google Translator	8.53	17.09	10.67	0.6848	-16.067
GPT Annotator w/ GPT-4	10.35	18.61	10.79	0.6911	-16.054
Estonian	BLEU	ROUGE	METEOR	BERTS.	BARTS.
NLLB (Machine-Translated)	4.97	13.12	7.89	0.6893	-15.409
HRQ-VAE + NLLB	3.37	7.84	5.87	0.6876	-15.409
Google Translator	6.04	12.51	8.75	0.7008	-15.408
GPT Annotator w/ GPT-4	6.62	13.47	9.22	0.7050	-15.407
Finnish	BLEU	ROUGE	METEOR	BERTS.	BARTS.
NLLB (Machine-Translated)	4.19	10.43	7.74	0.7122	-16.392
HRQ-VAE + NLLB	3.74	10.23	7.06	0.6965	-16.401
Google Translator	4.28	10.84	7.88	0.7128	-16.394
GPT Annotator w/ GPT-4	4.96	12.29	8.64	0.7143	-16.389



# Dataset Construction — Case Analysis

We conducted **case analysis** to compare the quality of generated sentences



## Latvian

Reference: There are four people playing tennis in doubles.

Translator: Divās grupās spēlē četri cilvēki.

*(Four people play in two groups.)*

GPT Annotator: Četri cilvēki spēlē tenisu dubultspēles.

*(A person hits a tennis ball with a tennis racket.)*



## Estonian

Reference: A person swing a tennis racket at a tennis ball.

Translator: Üks inimene käigub tennisepalli peal tennis racket.

*(One person moves a tennis racket on top of a tennis ball.)*

GPT Annotator: Inimene lööb tennis reketiga tennisepalli.

*(A person hits a tennis ball with a tennis racket.)*



## Finnish

Reference: People in uniforms playing baseball in the field

Translator: Joukkueessa pelaavat joukkueessa

*(In the team play in the team)*

GPT Annotator: Ihmiset uniformuissa pelaavat baseballia kentällä.

*(People in uniforms are playing baseball on the field.)*

# Limitation — Extreme Low-resource Language

Despite its various strengths, we found that **GPT-4's ability for translation and data annotation in extremely low-resource languages is still limited**

We used two different languages: Basque and Māori

- **Basque**

- Reference: A black dog and a spotted dog are fighting
- Google Translator: Txakur beltz bat eta txakur orban bat borrokan ari dira
- GPT Annotator: Kolore beltzeko txakur bat eta beste bat orbainekin borrokan ari dira.  
(*A black dog and another with **scars** are fighting.*)

- **Māori**

- Reference: Boys perform dances on poles during the nighttime.
- Google Translator: Ka kanikani nga tama ki runga pou i te po.
- GPT Annotator: Tamariki tāne e mahi ake ana i ngā pou i te po tuturu.  
(*Boys whoi **work up** posts in the real night.*)

# Conclusion

We proposed:

- a method to employ LLMs for data annotation in text generation tasks such as image captioning and text style transfer

We found that:

- LLMs such as GPT-4 are **cost-efficient, multilingual annotator** for text generation tasks
- With our proposed method, we can construct datasets in low-resource languages
- It still has limitations in extremely low-resource languages such as Basque and Maori

We plan to:

- Expanding proposed method to other tasks, especially multimodal task, utilizing image inception ability of GPT-4

**Thank You!**