



University of Stuttgart  
Germany



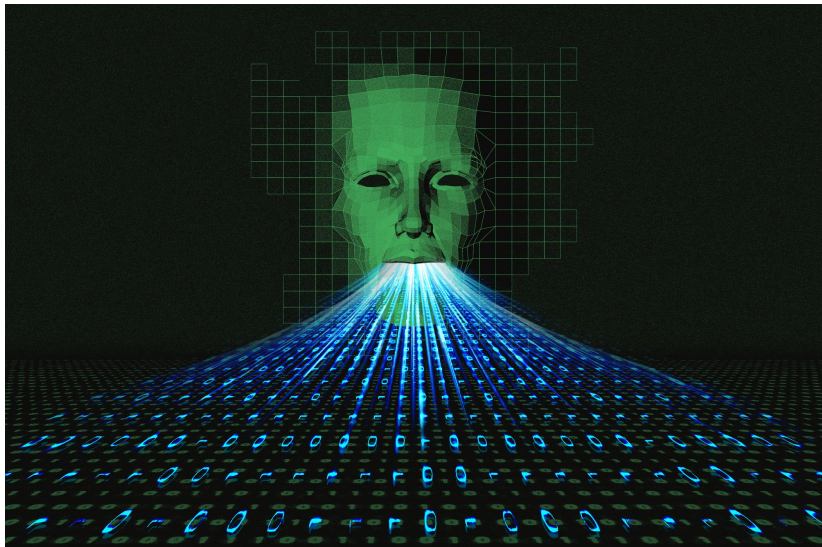
# PhiTag

## An open-source text annotation tool

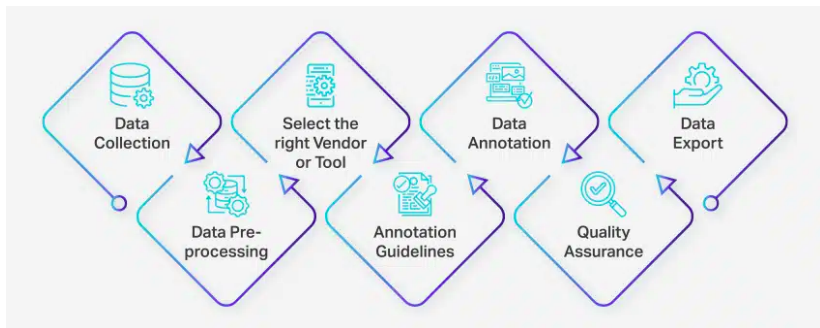
November 8, 2024

Dominik Schlechtweg

## Problem/Opportunity: data-hungry AI models



# Data annotation process



standardize & automate

# Added value

- ▶ **time** and **cost** reduction for
  - ▶ human annotation
  - ▶ computational annotation
  - ▶ model training process

# PhiTag

- ▶ online text annotation system<sup>1</sup>
- ▶ open source<sup>2</sup>
- ▶ various task types
- ▶ human and computational annotation
- ▶ interfaces with Prolific & OpenAI
- ▶ agreement statistics

---

<sup>1</sup><https://phitag.ims.uni-stuttgart.de/>

<sup>2</sup><https://github.com/Garrafao/phitag>

# Task types

- ▶ Text Label
  - ▶ text, tag, label set
  - ▶ graded WSD, entity matching
- ▶ **Text Pair**
  - ▶ text, text, label set
  - ▶ WiC, translation quality
- ▶ Text Rank
  - ▶ text, text, text. . . , rank
  - ▶ question answering
- ▶ Text Free
  - ▶ text, input text
  - ▶ instruction task, topic annotation, summarization
- ▶ Text Choice
  - ▶ text, text, text. . . , choice
  - ▶ preference task, sentiment

# Human annotation

# Interface Prolific



# Computational annotation

## Computational annotation - case study

- ▶ **Main question:** Can we reuse human training data (guideline, tutorial) to train (prompt) LLMs?

# Case study - task type

🔗 [Example-User](#) / [Example-Project-Use-Pair](#) / [ALL](#) / [Annotate](#)

28 out of 28628 completed

☰ Although she often argued with him about their responsibility to the community, Jean's husband didn't like listening to **words** without music and most Wednesdays stayed home, playing board games on his portable computer and watching Court TV.

☰ Good Mrs. Goodwin heard never a **word**, But kept on with her wringing, And though the wind blew most dismally blue, She lightened her care with singing.

1

2

3

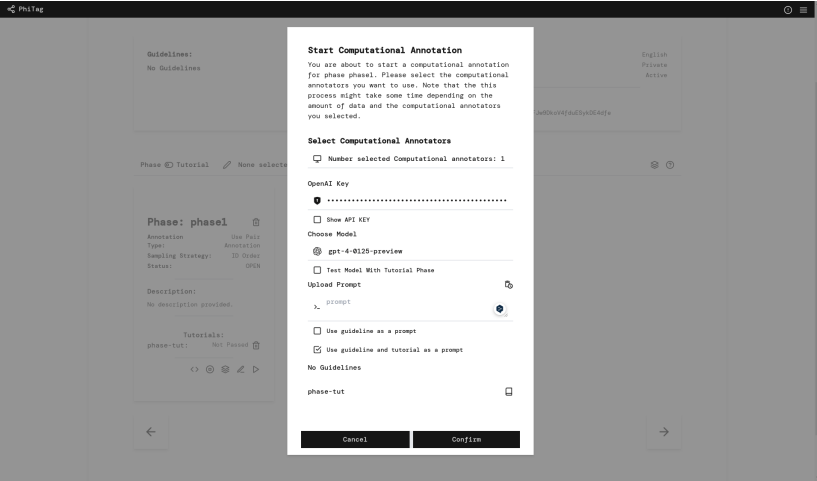
4

-

Comment

🗨 Comment

# Case study - model type



## Case study - prompt types

1. guideline
2. guideline + tutorial
3. customized

## Case study - data

<b>Dev</b>	<b>Train</b>	<b>Test</b>
46	140	744

Test data split statistics using DWUG EN (Schlechtweg et al., 2021)

## Case study - results

Strategy	dev	test
guideline	-0.07	0.03
guideline + tutorial	0.01	0.06
customized	0.74	<b>0.54</b>

Mean prompting results over five trials measured with Krippendorff's  $\alpha$

## More studies

- ▶ topic relatedness (under review)
- ▶ semantic relatedness (ongoing)
- ▶ word sense annotation (Lautenschlager, Hengchen, & Schlechtweg, 2024)



# Next Steps

- ▶ more tasks:
  - ▶ span annotation
  - ▶ refine preference & instruction tasks
  - ▶ what are industry needs?
- ▶ industry partners?
- ▶ funding?

# References I

- Lautenschlager, J., Hengchen, S., & Schlechtweg, D. (2024). *Detection of non-recorded word senses in English and Swedish*. Retrieved from <https://arxiv.org/abs/2403.02285>
- Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., & McGillivray, B. (2021, nov). DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 7079–7091). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.567>

## Image sources

- ▶ p. 2: <https://www.wsj.com/tech/ai/ai-training-data-synthetic-openai-anthropic-9230f8d8>, last accessed May 9, 2024.
- ▶ p. 3: <https://www.habiledata.com/blog/why-data-annotation-is-important-for-machine-learning-ai/>, last accessed May 9, 2024.

## Customized prompt

You are a highly trained text data annotation tool capable of providing subjective responses.

Rate the semantic similarity of the target word in these sentences 1 and 2. Consider only the objects/concepts the word forms refer to: ignore any common etymology and metaphorical similarity! Ignore case! Ignore number (cat/Cats = identical meaning). If target is emoji then rate by its contextual function. Homonyms (like bat the animal vs bat in baseball) count as unrelated. Output numeric rating: 1 is unrelated; 2 is distantly related; 3 is closely related; 4 is identical meaning. Your response should align with a human's succinct judgment.

Sentence 1:[SENTENCE 1]

Sentence 2: [SENTENCE 2]

Target word: [TARGET WORD]

Please provide a judgment as a single integer. For example, if your judgment is Identical, then provide 4. If your judgment is Unrelated, provide 1.

## Guideline prompt

You are a highly trained text data annotation tool capable of providing subjective responses.

[MODIFIED GUIDELINES]

Sentence 1: [SENTENCE 1]

Sentence 2: [SENTENCE 2]

Target word: [TARGET WORD]

Please provide a judgment as a single integer for Sentence 1 and Sentence 2 above. For example, if your judgment is Identical, then provide 4. If your judgment is Unrelated, provide 1.

## Guideline + tutorial prompt

You are a highly trained text data annotation tool capable of providing judgments based on contexts provided to you.

[MODIFIED GUIDELINES]

Here are few sample instances and their corresponding judgements:

Example sentences

Sentence 1: [SENTENCE 1]

Sentence 2: [SENTENCE 2]

Target word: [TARGET WORD]

Please provide a judgment as a single integer for Sentence 1 and Sentence 2 above. For example, if your judgment is Identical, then provide 4. If your judgment is Unrelated, provide 1.