Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents

Vitor Oliveira 🔷 Gabriel Nogueira 🔷 Thiago Faleiros 🔷 Ricardo Marcacini

📒 Agenda

- Dataset
- Problem addressed
- Solution
  - Prompt-based language models
  - Weak supervision
- Experiments
- Results
- Conclusion

🏷️ Dataset

Dataset

- Developed under the KnEDLe project
  (Knowledge Extraction from Documents of Legal content)
- It comprises 'Contract acts' publication from the Federal District Government
- It was extracted from the Official Gazette of the Federal District (Diário Oficial do Distrito Federal - DODF)

Dataset split

- 783 for training
  (instances that successfully fit into the GPT-3's context window of 2049 tokens)
- 379 for validation
- 380 for testing
- Summing 1.542 instances

Annotated entities

Named entities	Entity description
contract_number	Contract identification number
GDF_process	Process number before the Federal District government (GDF)
contractual_parties	Combination of contracting body, contracted entity, and convening entities
contract_object	Object to which the contract refers
contract_date	Contract signature date
contract_value	Estimated contract final value
contract_duration	Contract term of validity
budget_unit	Contract budget union number
work_program	Contract work program number
nature_of_expenditure	Contract nature of expenses number
commitment_note	Contract commitment note

Annotated entities

CONTRACT EXTRACT No. 01/2018,
UNDER THE TERMS OF STANDARD No. 09/2002.

PROCESS: 00367-00001224/2018-26 - PARTIES: DF/RA-XXXI X TRADE FUROS CONSTRUCOES E REFORMAS LTDA - ME. Purpose: hiring of a company to carry out the RENOVATION OF THE GALPAO DA FEIRA DA FERCAL - DF, relevant to the execution of the purpose, under the terms of invitation to bid No. 01/2018-RA-XXXI and its Attachments, ID 13145309, 13145702, 13145702, 13146338 and 13147459, which form part of this letter of invitation. Date of signature: 10/16/2018. Duration: The contract will be valid for 60 (sixty) days, from the date of its publication in the DODF, at the expense of the Administration. Value: R$ 49,620.00 (forty-nine thousand, six hundred and twenty reais), funds from the Federal District budget Budget Law No. 6,060, of 12/29/2017 Budget Allocation: Cod. from U.O.: 59135, Work Program: 15,451,6207,5523,0003, Nature of Expenditure: 449051, Source of Funds: 100; Event 400091; Commitment Note no: 2018NE00107, issued on 10/16/2018; Modality: Global. SIGNATORIES: contractor for the Federal District, Jefferson de Sousa Oliveira, as Regional Administrator; by the Contractor: Paulo Henrique da Silva Oliveira, as Legal Representative.

Contract Number

GDF Process

Contract Date:

Contract Value

Contractual Parties

Contract Object

Contract Duration

Work Program

Commitment Note

Nature of Expenditure

Budget Unit

Fig. 2 "Contract" act labeling example

Source: Oliveira *et al.* (2024)

⚠️ Problem

> "(...) Even though this human participation improves model performance, in many projects, the usual process of reading, searching, identifying, circumscribing, and reviewing can be *costly in terms of time, money, and effort*."
> Oliveira *et al.* (2024)

💡 Solution

- Prompting-based language model
- Weak Supervision

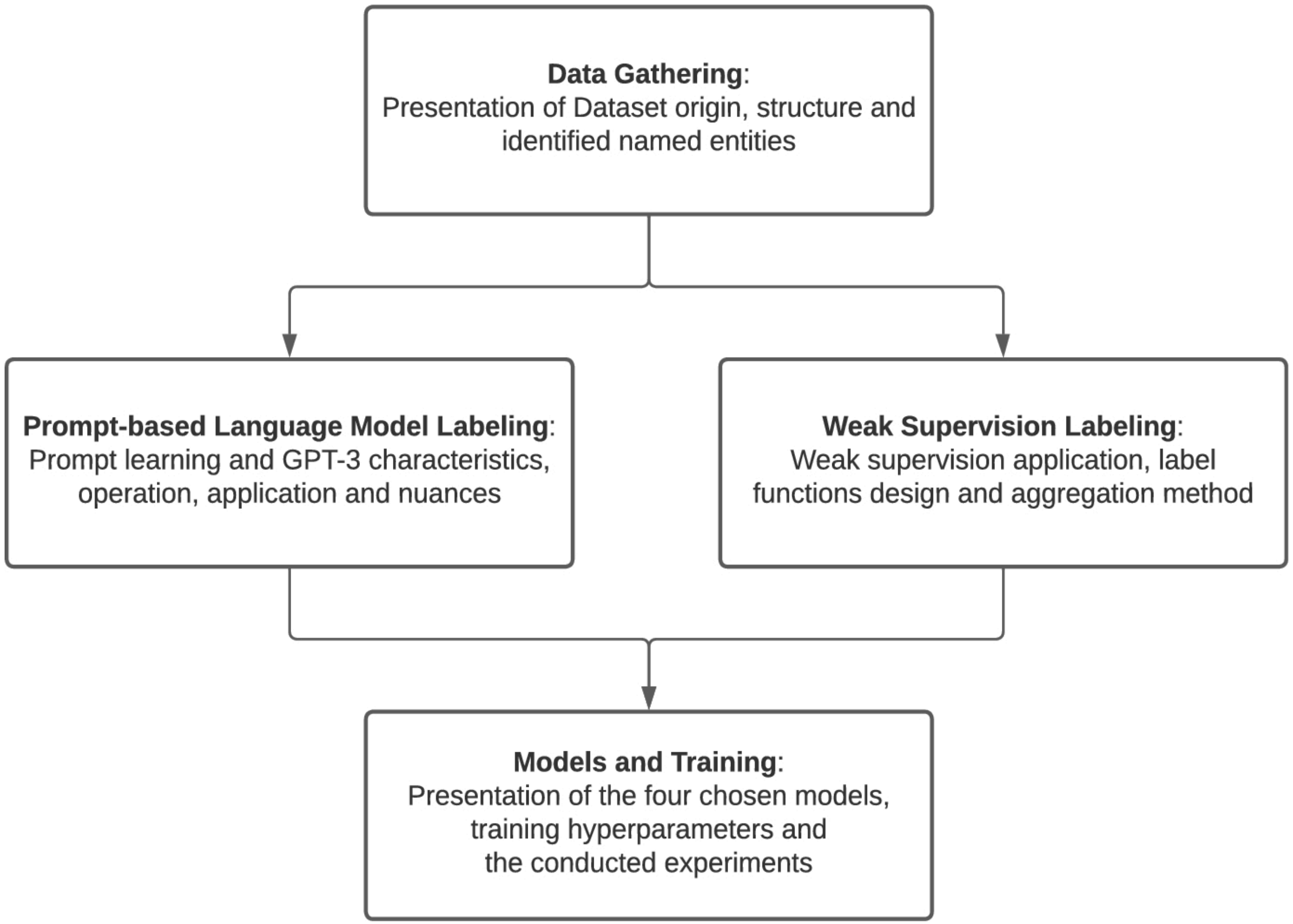


Figure 1 Workflow methodology. Source: Oliveira *et al.* (2024)

🤖 Prompt-based language models

Prompt-based annotation

- It used the GPT-3 Davinci model
- It had a maximum request of about 2049 tokens
- Three dataset instances were handpicked and randomly given as a prompt-example
- Of course, they selected such example that contains all entities labels
- Davinci applies its prediction to exactly one unlabeled act

Prompt instruction

1. GPT-3 is given the prompt below as annotation example.

Prompt input text:

CONTRACT EXTRACT FOR THE ACQUISITION OF GOODS No. 12/2021 Process: 04011-00001803/2021-37. Parties: THE FEDERAL DISTRICT, through the STATE SECRETARIAT FOR WOMEN OF THE FEDERAL DISTRICT, CNPJ nº 15.169.975/0001-15, and the company INDÚSTRIA DE ÁGUA MINERAL IBIÁ LTDA, CNPJ nº 05.655.158/0001-13. Purpose: acquisition of foodstuff material (drinking water) and packaging material (returnable carboy - container) intended for the functioning of this Secretariat of State for Women of the Federal District. BUDGETARY UNIT: 57,101. WORK PROGRAM: 14.122.8211.8517.0163. TYPE OF EXPENSE: 339030. SOURCE OF RESOURCE: 100. INICAL COMMITMENT NOTE: No. 2021NE00159, in the amount of R$ 3,186.00 (three thousand, one hundred and eighty-six reais), issued on 08/24/2021. EVENT: 400091. MODALITY: Estimate. CONTRACT AMOUNT: R$ 11,469.60 (eleven thousand, four hundred and sixty-nine reais and sixty cents). TERM: The contract will be valid for 12 (dose) months, from 09/03/2021 to 09/03/2022. SUBSCRIPTION: 08/24/2021. SIGNATORIES: by Contracting Party: VANDERCY ANTONIA DE CAMARGOS, in the capacity of Executive Secretary; by Contractor: EDUARDO BARROS DE QUEIROZ RODRIGUES, as Legal Representative.

Prompt output text:

contract_number: 12/2021 # gdf_process: 04011-00001803/2021-37 # contracting_organ: STATE SECRETARIAT FOR WOMEN OF THE FEDERAL DISTRICT # contracted_entity: INDÚSTRIA DE ÁGUA MINERAL IBIÁ LTDA # contract_object: acquisition of foodstuffs (drinking water) and material of packaging and packaging (returnable carboy - container) intended for the operation of this Secretariat of State for Women of the Federal District # date_signature_contract: 08/24/2021 # value_contract: 11,469.60 # duration_contract: The contract will be valid for 12 (dose) months, from 09/03/2021 to 09/03/2022 # budget_unit: 57,101 # work_program: 14.122.8211.8517.0163 # expense_nature: 339030 # commitment_note: 2021NE00159

Figure 3. GPT-3 prompt-labeling process example. Source: Oliveira *et al.* (2024)

New instance input

2. GPT-3 is applied in an unlabeled instance.

Unlabeled contract example text:

CONTRACT EXTRACT No. 01/2018,
UNDER THE TERMS OF STANDARD No. 09/2002.

PROCESS: 00367-00001224/2018-26 - PARTIES: DF/RA-XXXI X TRADE FUROS CONSTRUCOES E REFORMAS LTDA - ME. Purpose: hiring of a company to carry out the RENOVATION OF THE GALPAO DA FEIRA DA FERCAL - DF, relevant to the execution of the purpose, under the terms of invitation to bid No. 01/2018-RA-XXXI and its Attachments, ID 13145309, 13145702, 13145702, 13146338 and 13147459, which form part of this letter of invitation. Date of signature: 10/16/2018. Duration: The contract will be valid for 60 (sixty) days, from the date of its publication in the DODF, at the expense of the Administration. Value: R$ 49,620.00 (forty-nine thousand, six hundred and twenty reais), funds from the Federal District budget Budget Law No. 6,060, of 12/29/2017 Budget Allocation: Cod. from U.O.: 59135, Work Program: 15,451,6207,5523,0003, Nature of Expenditure: 449051, Source of Funds: 100; Event 400091; Commitment Note no: 2018NE00107, issued on 10/16/2018; Modality: Global. SIGNATORIES: contractor for the Federal District, Jefferson de Sousa Oliveira, as Regional Administrator; by the Contractor: Paulo Henrique da Silva Oliveira, as Legal Representative.

3. GPT-3 generates a labeled text based in the prompt example.

Generated labeled contract output text:

contract_number: 01/2018 # gdf_process: 00367-00001224/2018-26 # contracting_organ: RA-XXXI # contracted_entity: TRADE FUROS CONSTRUCOES E REFORMAS LTDA - ME # contract_object: contracting of a company to carry out RENOVATION works at the FAIR FAIR SHED FERCAL - DF, relevant to the execution of the object, in accordance with the invitation notice No. 01/2018-RA-XXXI and its Annexes, ID 13145309, 13145702, 13145702, 13146338 AND 13147459, which are part of this invitation letter. # contract_date: 10/16/2018 # contract_value: 49,620.00 # contract_duration: The contract will be valid for 60 (sixty) days, counting from the date of its publication in the DODF, at the expense of the Administration. # budget_unit: 59135 # work_program: 15,451,6207,5523,0003 # nature_of_expenditure: 449051 # commitment_note: 2018NE00107 # event_note: 400091 # modality: global # signatories: contractor for the Federal District, Jefferson de Sousa Oliveira, as Regional Administrator; by the Contractor: Paulo Henrique da Silva Oliveira, as Legal Representative.

Figure 3. GPT-3 prompt-labeling process example. Source: Oliveira *et al.* (2024)

Expected outcome

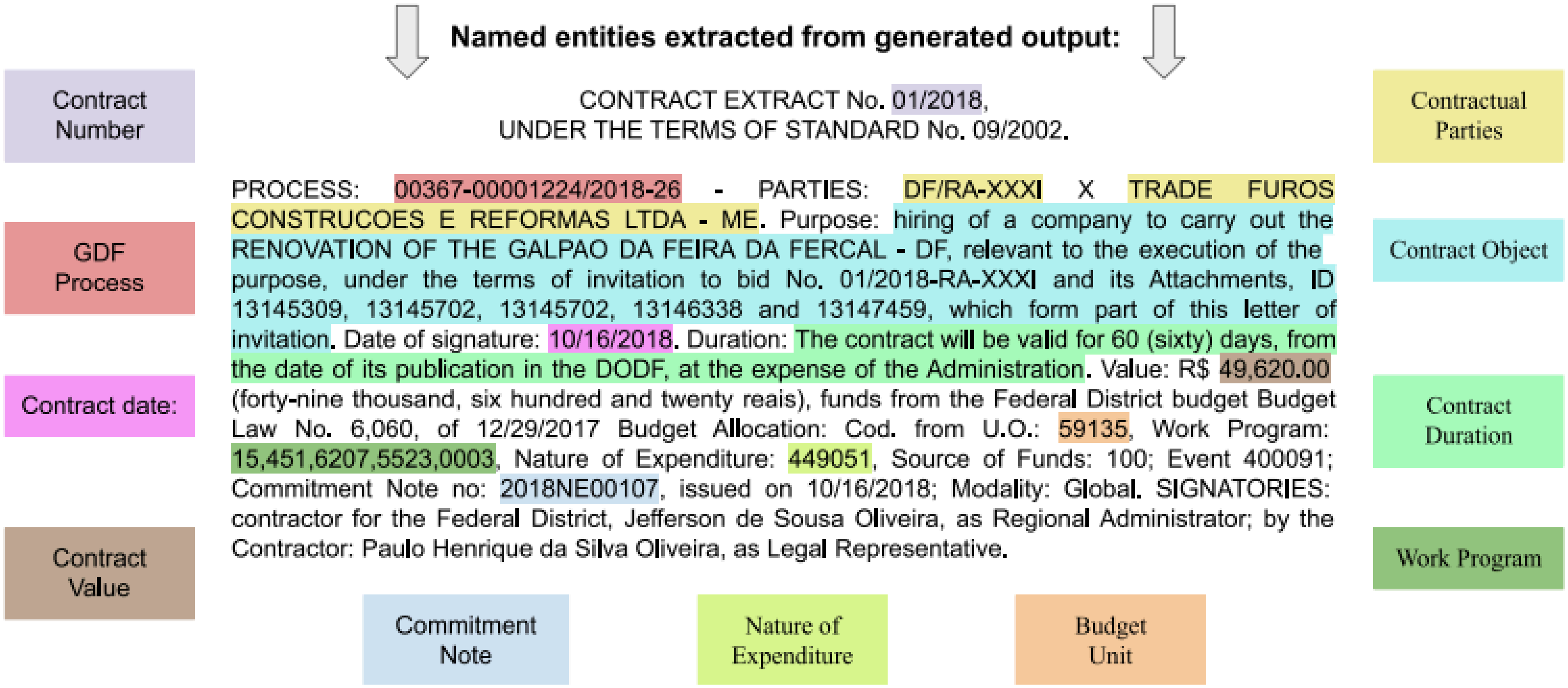


Figure 3. GPT-3 prompt-labeling process example. Source: Oliveira *et al.* (2024)

The annotation cost using GPT-3

- To annotated 783 training instances they had
  - 1.565.108 tokens
  - at a final cost of $ 31.30 dollars

Named Entity Recognition to Enrich Text

D. Carpintero

Oct 19, 2023

Open in Github

`Named Entity Recognition` (NER) is a `Natural Language Processing` task that identifies and classifies named entities (NE) into predefined semantic categories (such as persons, organizations, locations, events, time expressions, and quantities). By converting raw text into structured information, NER makes data more actionable, facilitating tasks like information extraction, data aggregation, analytics, and social media monitoring.

This notebook demonstrates how to carry out NER with **chat completion** and **functions-calling** to enrich a text with links to a knowledge base such as Wikipedia:

Text:

In Germany, in 1440, goldsmith Johannes Gutenberg invented the movable-type printing press. His work led to an information revolution and the unprecedented mass-spread of literature throughout Europe. Modelled on the design of the existing screw presses, a single Renaissance movable-type printing press could produce up to 3,600 pages per workday.

Text enriched with Wikipedia links:

*In **Germany**, in 1440, goldsmith **Johannes Gutenberg** invented the **movable-type printing press**. His work led to an **information revolution** and the unprecedented mass-spread of literature throughout **Europe**. Modelled on the design of the existing screw presses, a single **Renaissance** **movable-type printing press** could produce up to 3,600 pages per workday.*

Inference Costs: The notebook also illustrates how to estimate OpenAI API costs.

1. Setup

1.1 Install/Upgrade Python packages

```
%pip install --upgrade openai --quiet
%pip install --upgrade nlpia2-wikipedia --quiet
%pip install --upgrade tenacity --quiet
```

https://platform.openai.com/docs/guides/gpt/function-calling

Example

```
messages = [
          {"role": "system", "content": system_message(labels=labels)},
          {"role": "assistant", "content": assisstant_message()},
          {"role": "user", "content": user_message(text=text)}
      ]

response = openai.chat.completions.create(
    model="gpt-3.5-turbo-0613",
    messages=messages,
    tools=generate_functions(labels),
    tool_choice={"type": "function", "function" : {"name": "enrich_entities"}},
    temperature=0,
    frequency_penalty=0,
    presence_penalty=0,
)
```

📝 Weak supervision

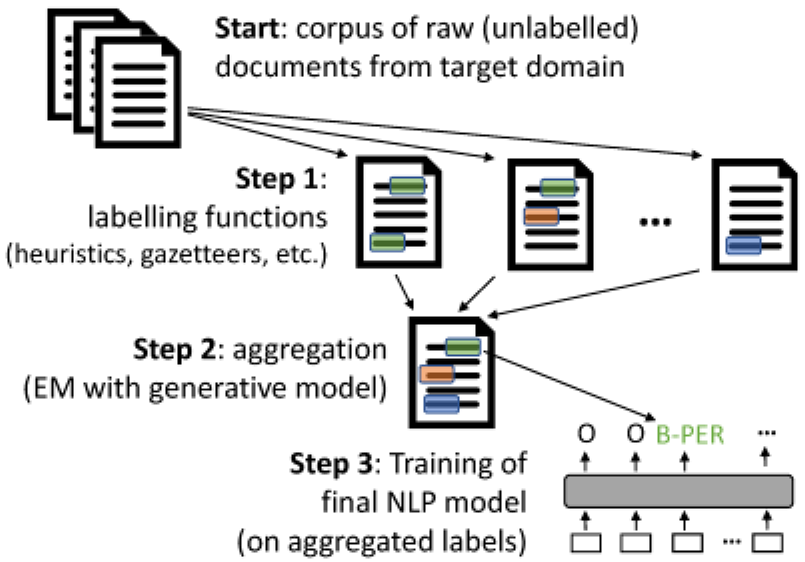
Weak supervision

One can define heuristic rules based on:

- Regular expressions
- Lookup tables or lists
- POS patterns or dependency relations
- Presence or neighbouring words within a given context window
- Using machine learning models trained on related tasks

The paper used the `skweak` library

- Help to implement **Labeling functions** specific heuristics
- It aggregates the resulting labels to obtain a labelled corpus
- It is integrated with spaCy library



General overview of `skweak`. Source: Lison, *et. al* (2021)

Labelling functions

```
import spacy, re
from skweak import heuristics, gazetteers, generative, utils

def money_detector(doc):
  ... # do some stuff

heuristics.FunctionAnnotator("money_detector", money_detector)

NAMES = [...] # list of names

trie = gazetteers.Trie(NAMES) # under the hood it uses trie regex
gazetteers.GazetteerAnnotator("presidents", {"PERSON":trie})
```

Source: https://github.com/NorskRegnesentral/skweak/wiki/Step-1:-Labelling-functions

An alternative would be `snorkel`

snorkel

```python
from snorkel.labeling import labeling_function
from snorkel.labeling import PandasLFApplier

@labeling_function()
def lf_contains_link(x):
  ...
  # do something

lfs = [lf_contains_link, ... ]

applier = PandasLFApplier(lfs=lfs)
L_train = applier.apply(df=df_train)
```

Source: https://www.snorkel.org/use-cases/01-spam-tutorial

Other resources

- Trie regex
  - `trieregex`
  - `trex`
  - `flashtext` 1
- Spacy rule-based matching 2
  - Token matcher
  - Phrase matcher

🧪 Experiments

3.4 Models and training

In relation to the NER models, we chose four pre-trained neural language models. Each model passed through a fine-tuning process (Sun et al 2019), consisting of adding a BI-LSTM (Bidirectional Long Short-Term Memory Network) (Graves and Schmidhuber 2005) layer at its top for sequence labeling. The chosen models are the following:

- BERTimbau (Souza et al 2020): Pre-trained BERT model with a Brazilian Portuguese textual corpus.
- Lener-BR[4]: A fine-tuned BERTimbau model for Brazilian Portuguese legislative texts.
- RoBERTa (Liu et al 2019): An optimized version of the BERT model, developed with support from Facebook researchers.
- DistilBERT-PT[5]: A lighter (distilled) version of BERT, pre-trained with a Brazilian Portuguese textual corpus.

Source: Oliveira *et al.* (2024)

ktrain

```
import ktrain
from ktrain import text as txt

WV_URL='https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.nl.300.vec.gz'

model = txt.sequence_tagger('bilstm-transformer',
                            preproc,
                            transformer_model='wietsedv/bert-base-dutch-cased',
                            wv_path_or_url=WV_URL)

learner = ktrain.get_learner(model, train_data=trn, val_data=val, batch_size=128)

learner.fit(0.01, 1, cycle_len=5, checkpoint_folder='/tmp/saved_weights')
```

Source https://github.com/amaiya/ktrain/tree/master/examples#seqlab

Models

- Hence, they considered 4 embeddings techniques:
  - BERTimbau (Souza et al 2020)
  - LeNER-BR (1)
  - RoBERTa
  - DistilBERT-PT

(1) A fine-tuned BERTimbau model, not from the original Luz de Araújo paper

Datasets variations

- They trained each model on:
  - Human labeled
  - GPT-3 labeled
  - Weak-supervision
  - GPT-3 + Human-labeled (gradual combination of 10%, 20% ... 100%)
  - GPT-3 + Weak supervision (complete combination)

> They considered that combination of Weak-supervision and Human-labeled does no make sense, and it is a bad resource management.

🔎 Results

#1 - Isolated annotation datasets

Model	GPT-3	Weak supervision	Human labeling
NER-LenerBR	0.554	0.676	0.761
NER-BERTimbau	0.543	0.703	0.755
NER-RoBERTa	0.542	0.674	0.707
NER-DistilBERT-PT	0.473	0.664	0.631
Average F1-Scores	0.528	0.679	0.713

Table 3. F1-Score metric and average F1-Score metric of each model in every dataset

Source: Oliveira *et al.* (2024)

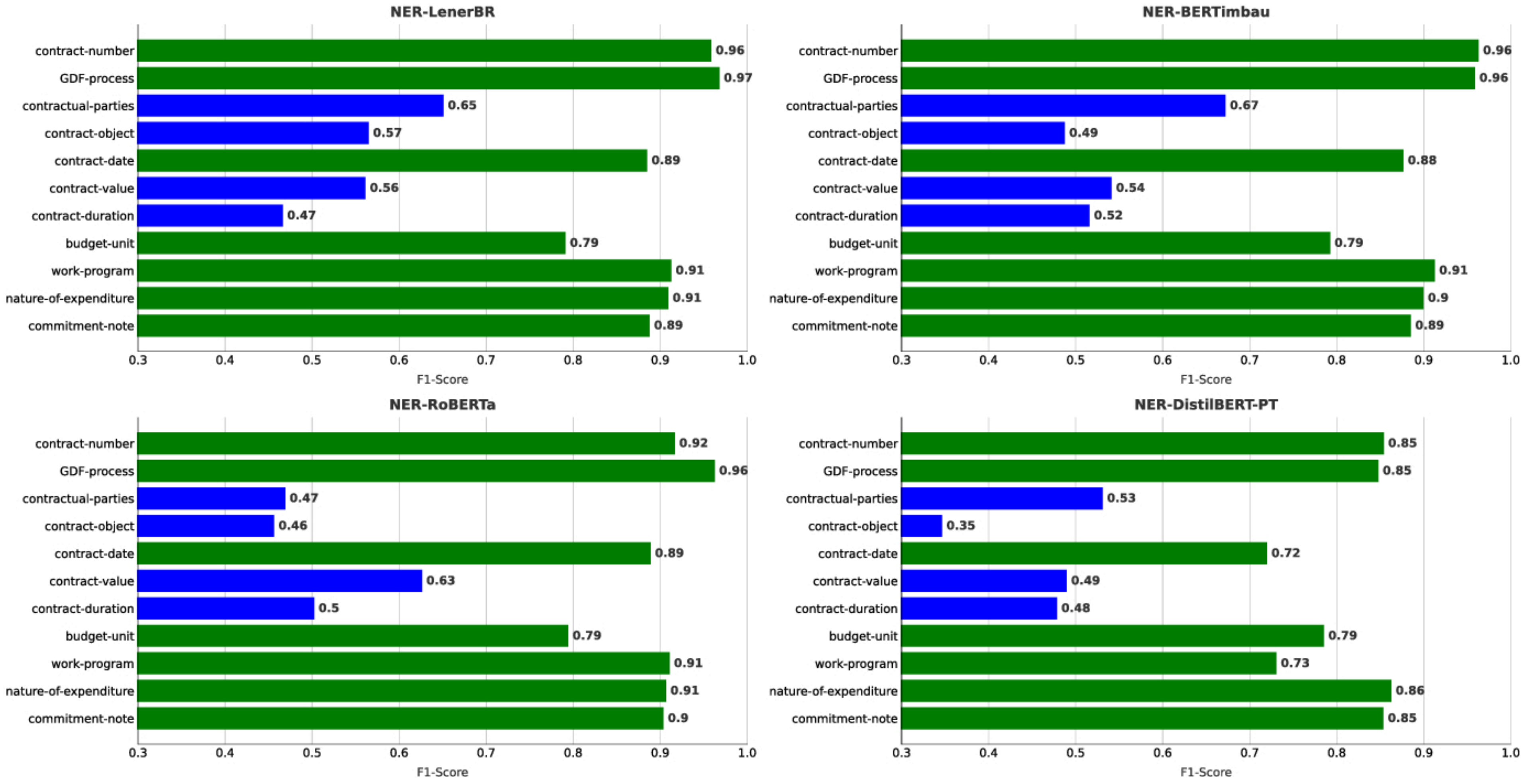
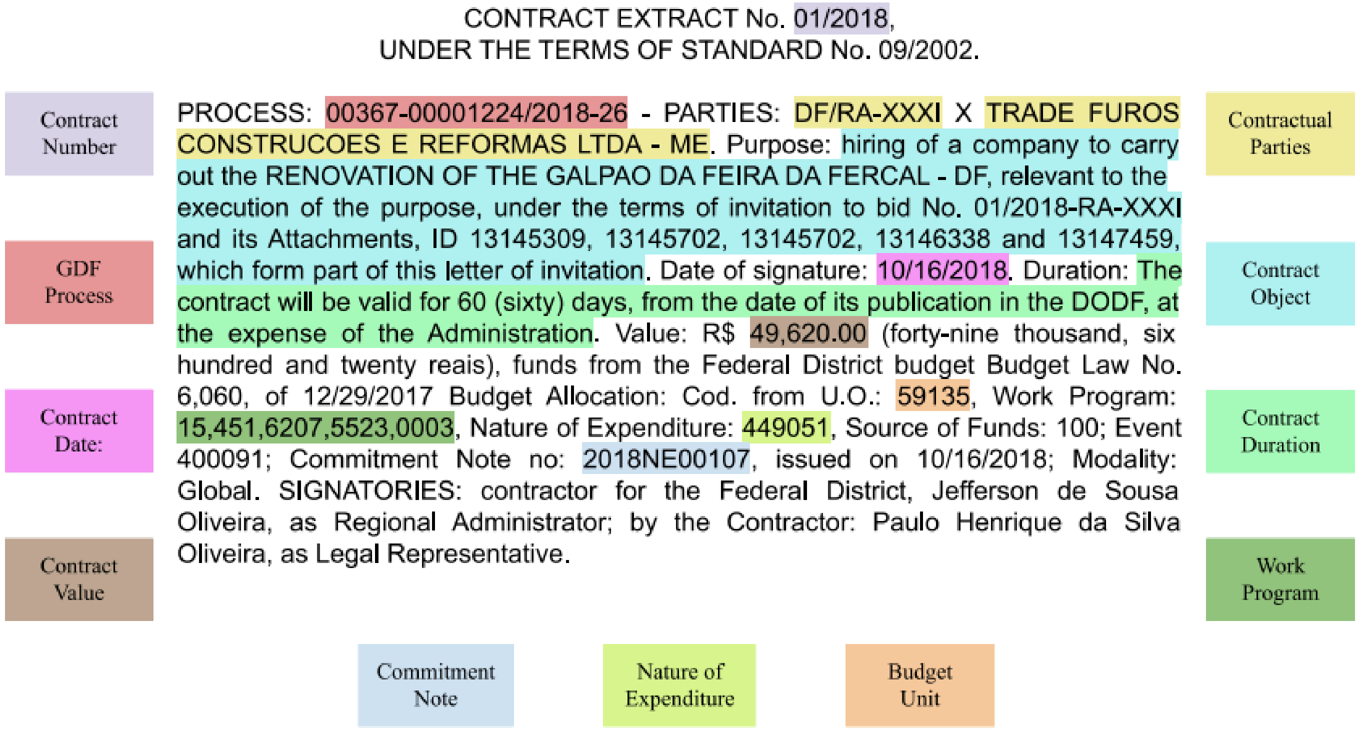


Figure 5. F1-Scores of all named entities for each model trained with the human-labeled dataset. In green are the chosen seven best-performing entities.
Source: Oliveira *et al.* (2024)

Seven best entities

1. contract number
2. GDF process
3. contract value
4. budget unit
5. work program
6. nature of expediture
7. commitment note

What do they have in common?



CONTRACT EXTRACT No. 01/2018,
UNDER THE TERMS OF STANDARD No. 09/2002.

PROCESS: 00367-00001224/2018-26 - PARTIES: DF/RA-XXXI X TRADE FUROS CONSTRUCOES E REFORMAS LTDA - ME. Purpose: hiring of a company to carry out the RENOVATION OF THE GALPAO DA FEIRA DA FERCAL - DF, relevant to the execution of the purpose, under the terms of invitation to bid No. 01/2018-RA-XXXI and its Attachments, ID 13145309, 13145702, 13145702, 13146338 and 13147459, which form part of this letter of invitation. Date of signature: 10/16/2018. Duration: The contract will be valid for 60 (sixty) days, from the date of its publication in the DODF, at the expense of the Administration. Value: R$ 49,620.00 (forty-nine thousand, six hundred and twenty reais), funds from the Federal District budget Budget Law No. 6,060, of 12/29/2017 Budget Allocation: Cod. from U.O.: 59135, Work Program: 15,451,6207,5523,0003, Nature of Expenditure: 449051, Source of Funds: 100; Event 400091; Commitment Note no: 2018NE00107, issued on 10/16/2018; Modality: Global. SIGNATORIES: contractor for the Federal District, Jefferson de Sousa Oliveira, as Regional Administrator; by the Contractor: Paulo Henrique da Silva Oliveira, as Legal Representative.

Fig. 2 "Contract" act labeling example

Source: Oliveira *et al.* (2024)

Model	GPT-3	Weak supervision	Human labeling
NER-LenerBR	0.815	0.878	0.906
NER-BERTimbau	0.776	0.887	0.902
NER-RoBERTa	0.798	0.881	0.899
NER-DistilBERT-PT	0.664	0.847	0.804
Average F1-Scores	0.763	0.873	0.877

Table 4. F1-Score metric considering only the seven best performing named entities.
Source: Oliveira *et al*. (2024)

#2 - Combining percentages of Human and GPT-3 labeling

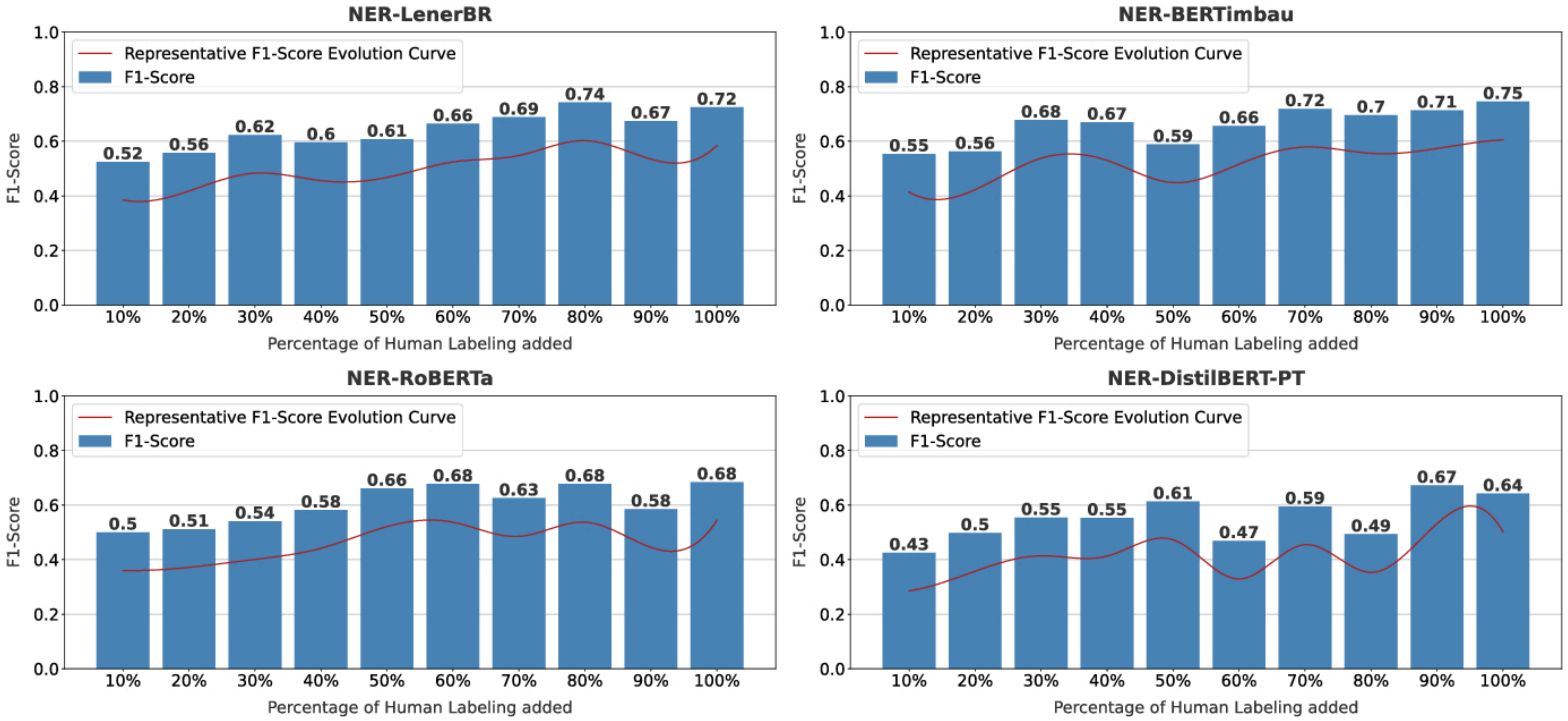


Figure 6. F1-Score over the GPT-3 and Human Labeling combining iterations and all eleven entities. Source: Oliveira *et al.* (2024)

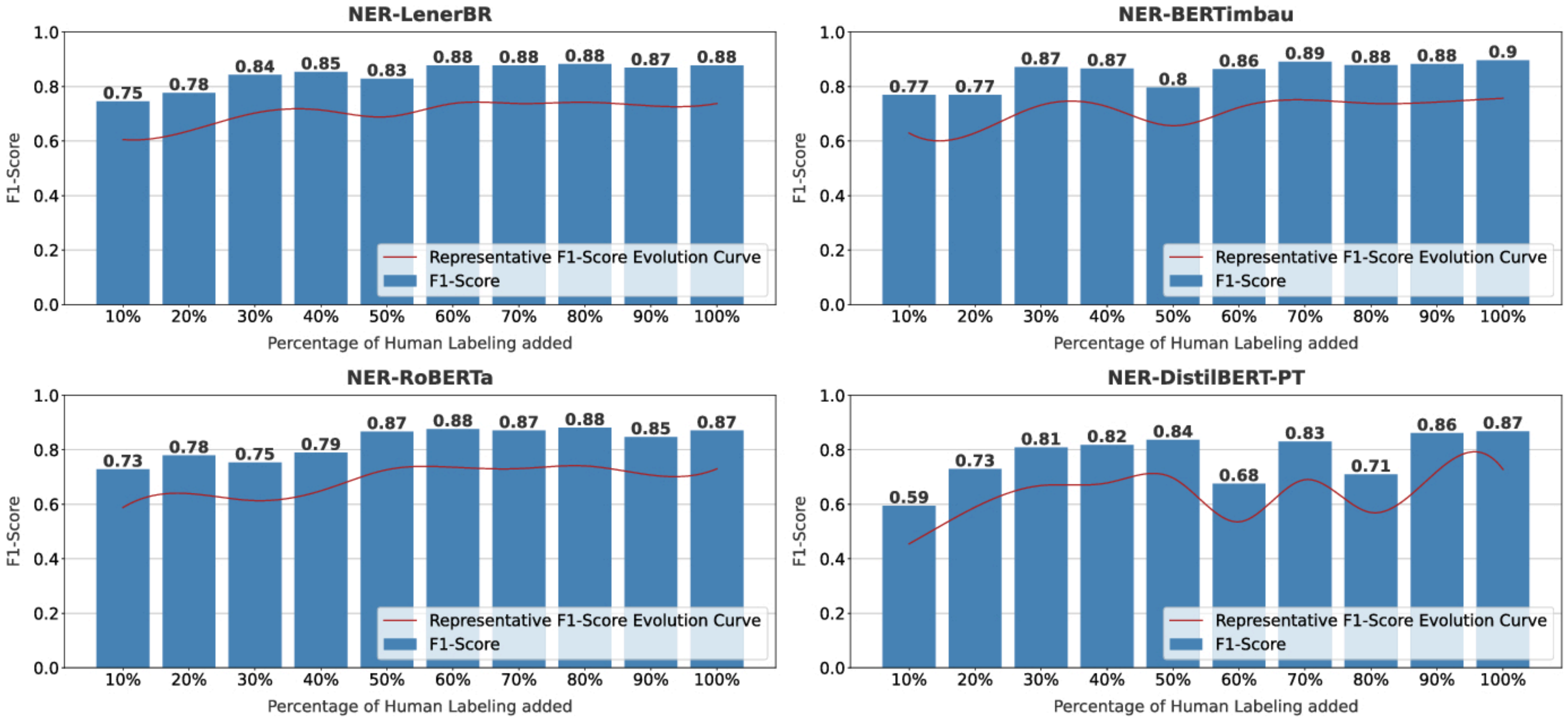


Figure 7. F1-Score over the GPT-3 and Human Labeling combining iterations considering only the seven best-performing entities. Source: Oliveira *et al*. (2024)

#3 - Combination of GPT-3 and Weak-supervision labeling

	GPT-3 and weak supervision	GPT-3 and weak supervision
Model	All eleven entities	Seven best entities
NER-LenerBR	0.709	0.888
NER-BERTimbau	0.686	0.884
NER-RoBERTa	0.558	0.773
NER-DistilBERT-PT	0.632	0.831
Average F1-Scores	0.646	0.844

Table 5. F1 scores values resulting the combination of GPT-3 and Weak supervision datasets. Source: Oliveira *et al.* (2024)

#4 - The preservation score analysis

Preservation score

$$P_{score} = \frac{F1_{tested\_model}}{F1_{human\_label}} \rightarrow \text{baseline}$$

Where $F1_{\cancel{human\_label}}$ correspond to:

- GPT-3
- Weak-supervision
- GPT-3 + Weak-supervision
- GPT-3 + 30% Human

Model	GPT-3	Weak-Sup	GPT Weak-Sup	GPT 30% Human
NER-LenerBR	0.728	0.888	0.931	0.818
NER-BERTimbau	0.719	0.931	0.908	0.897
NER-RoBERTa	0.766	0.953	0.789	0.765
NER-DistilBERT-PT	0.749	1.052	1.001	0.877
Average	0.740	0.956	0.907	0.839

Table 6-A. Preservation score comparison for all eleven entities. Source: Oliveira *et al.* (2024)

Model	GPT-3	Weak-Sup	GPT Weak-Sup	GPT 30% Human
NER-LenerBR	0.899	0.969	0.980	0.930
NER-BERTimbau	0.860	0.983	0.980	0.966
NER-RoBERTa	0.887	0.980	0.859	0.838
NER-DistilBERT-PT	0.825	1.053	1.033	1.006
Average	0.867	0.996	0.963	0.935

Table 6-B. Preservation score for the seven best entities. Source: Oliveira *et al.* (2024)

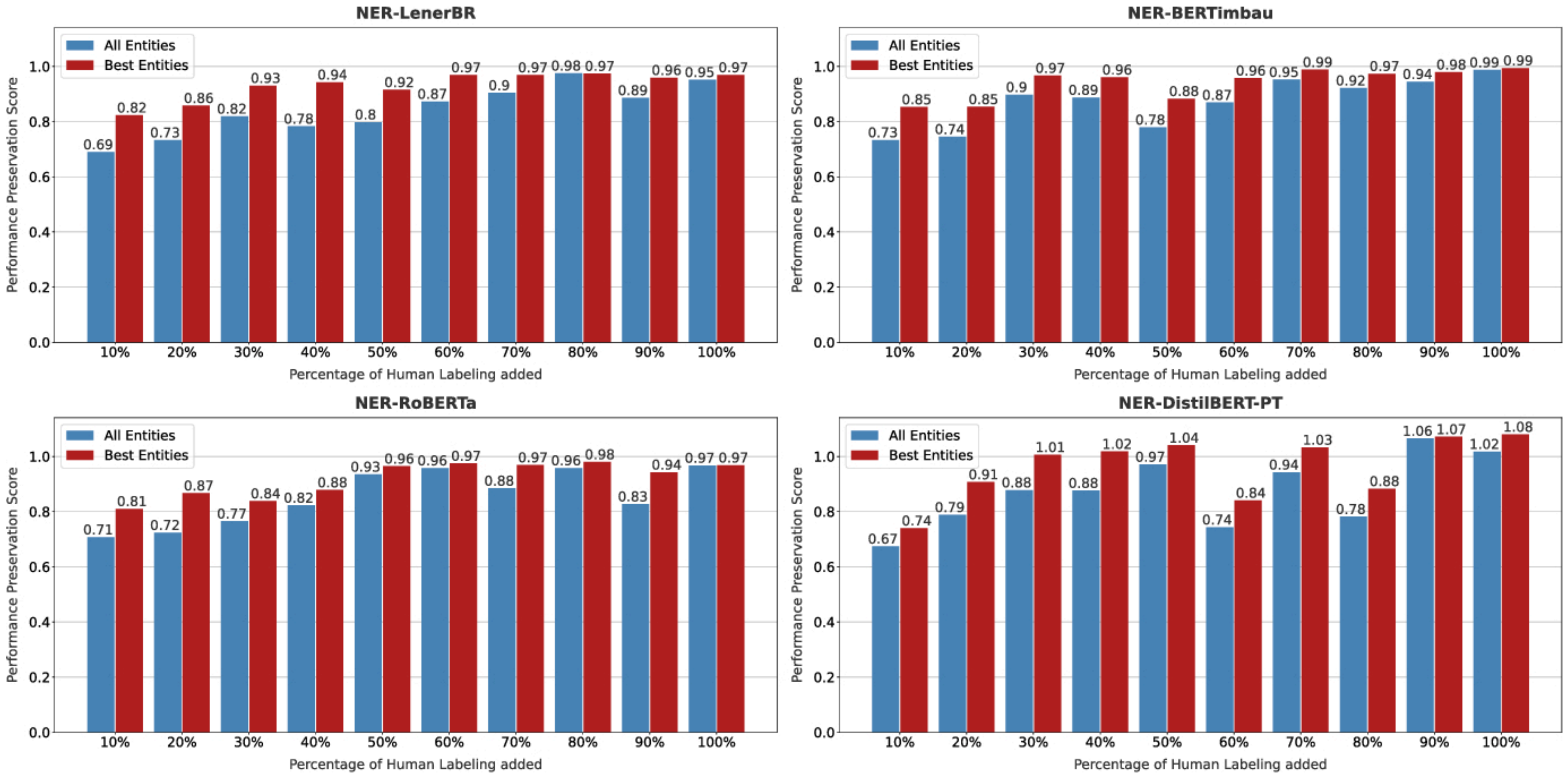


Figure 8. Preservation score for each iteration on the combination of GPT-3 and Human annotation. Source: Oliveira *et al.* (2024)

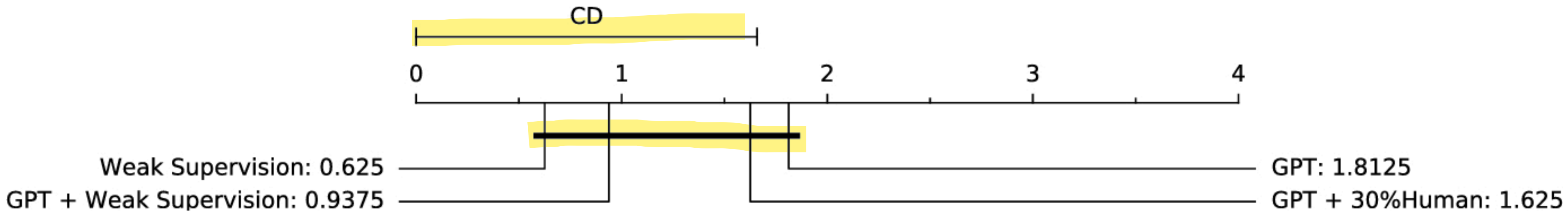


Figure 9. Friedman's test with Nemenyi's post test graphical analysis. Source: Oliveira *et al*. (2024)

🎯 Conclusions

- These strategies can still be a valid approach, even with their lower performance trade-off
- Human labeling is the best strategy considering accuracy and performance
- The statistical test did not show significant differences between the alternative approaches
- Limitations are:
  - There is no precise way to estimate each approach's cost
  - The size of the dataset
  - Did not considered the mistakes among annotators.

My observations

- How the GPT-3 and the weak supervision annotation agrees with the human labeled dataset?
- Is worth to compute standard metrics considering the weak supervision and the prompt-based as a black box model?
- Is it possible to explorer other prompt engineering techniques?

References

- Carpintero, D. **Named Entity Recognition to Enrich Text**. 2023. OpenIA Cookbook. https://cookbook.openai.com/examples/named_entity_recognition_to_enrich_text
- Knowledge Extraction from Documents of Legal content. https://unb-knedle.github.io/
- Lison, P; Barnes, J; Hubin, A. 2021. **skweak: Weak Supervision Made Easy for NLP** Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations
- Ratner, Alex; Varma, P; Hancock B; Ré, C; and others. 2019. **Weak Supervision: A New Programming Paradigm for Machine Learning**. https://ai.stanford.edu/blog/weak-supervision/

📜 References

Models card:

- BERTimbau https://huggingface.co/neuralmind/bert-base-portuguese-cased (Souza et al 2020)
- LeNER-BR https://huggingface.co/pierreguillou/bert-base-cased-pt-lenerbr
- RoBERTa https://huggingface.co/FacebookAI/roberta-base
- DistilBERT-PT https://huggingface.co/adalbertojunior/distilbert-portuguese-cased