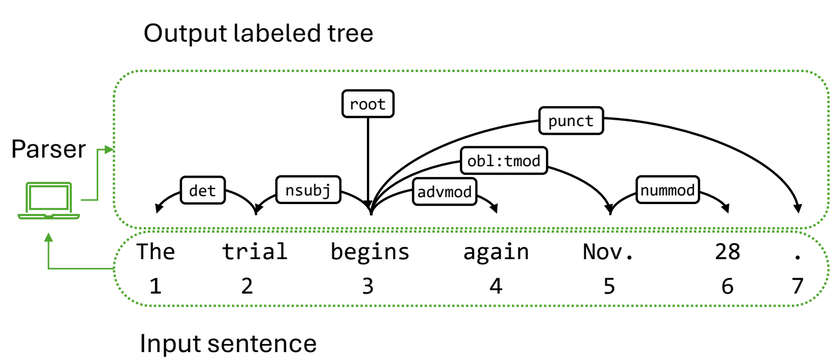
Better Benchmarking LLMs for Zero-Shot Dependency Parsing

Ana Ezquerro, Carlos Gómez-Rodríguez and David Vilares

Zero-Shot Dependency Parsing

- Evaluation of LLMs in a zero-shot setting (prompt).
- Difficult structured-task.



- Prompt with the CoNLL format:

```
In dependency parsing the CoNLL format for the sentence
<The trial begins again Nov. 28 .> is:
ID  FORM    LEMMA  UPOS  XPOS  FEATS  HEAD  DEPREL   DEPS  MISC
1   The     _      _     _     _      2     det      _     _
2   trial   _      _     _     _      3     nsubj    _     _
3   begins  _      _     _     _      0     root     _     _
4   again   _      _     _     _      3     advmod   _     _
5   Nov.    _      _     _     _      3     obl:tm   _     _
                                            od
6   28      _      _     _     _      5     nummod   _     _
7   .       _      _     _     _      3     Punct    _     _
Now return the CoNLL format for the sentence:
<What if Google Morphed into GoogleOS?>
```

LLM vs Naive baseline

```
Sure! This is the CoNLL format for
the sentence <What if Google
Morphed into GoogleOS?>:
1 What _ _ _ 0 nsubj _ _
2 if _ _ _ _ 4 mark _ _ _
3 Google _ _ _ _ 4 nsubj
4 Morphed _ _ _ _ 0 root _ _
5 into _ _ _ _ 6 case _ _
6 GoogleOS _ _ _ _ 8 nmod _ _
7 ? _ _ _ _ 4 punct _ _
Let me know if I can help you
again!
```

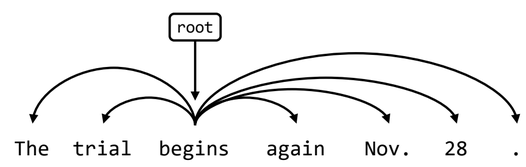
https://github.com/anaezquerro/naipar

Better Benchmarking with Naïve Baselines

Key idea: Do LLMs perform better than uninformed baselines?

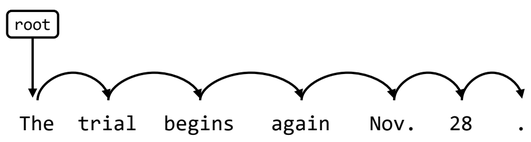
- No information about the input sentence (fully unsupervised).
- Only treebank sampling considers the length of the sentence and a reference treebank.

A: All to root



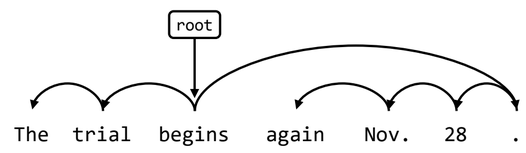
1. Select root at random.
2. All connections to root.

R: Right



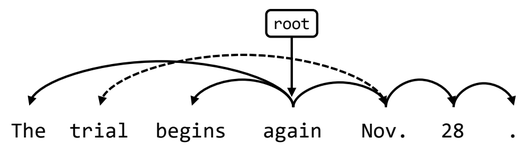
The parent is always the previous word.

L: Left



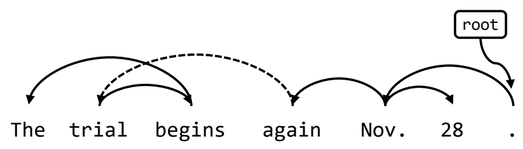
1. Select root at random.
2. The parent is always the next word.
3. For the last word, its parent is the root.

RD: Random (projective)



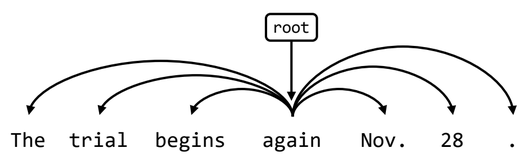
1. Uniformly random tree .
2. (Optional) Force projectivity.

LI: Optimal linear arrangement (projective)



1. Uniformly random tree .
2. Find arrangement to minimize dependency distances.
3. (Optional) Force projectivity.

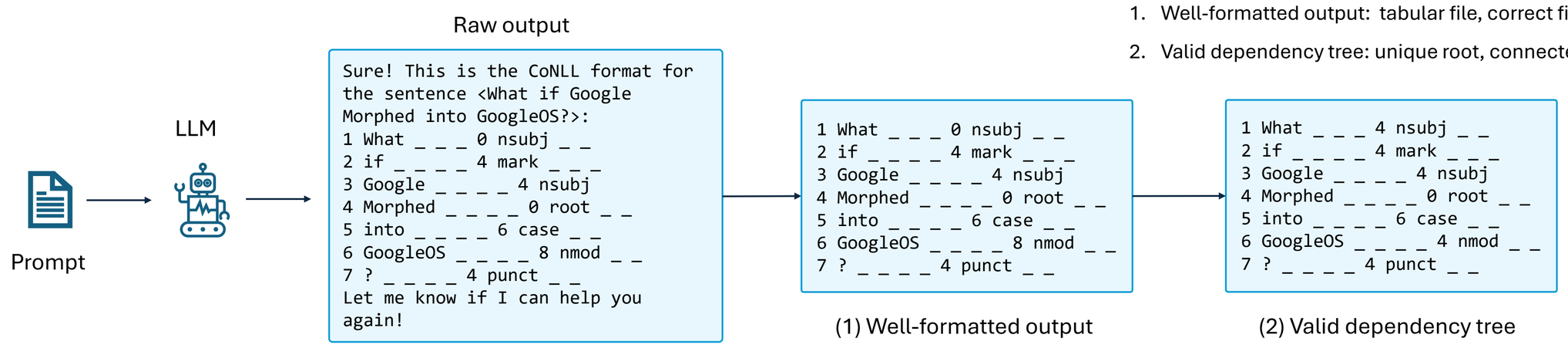
S: Treebank sampling



Sample the tree distribution from a reference treebank

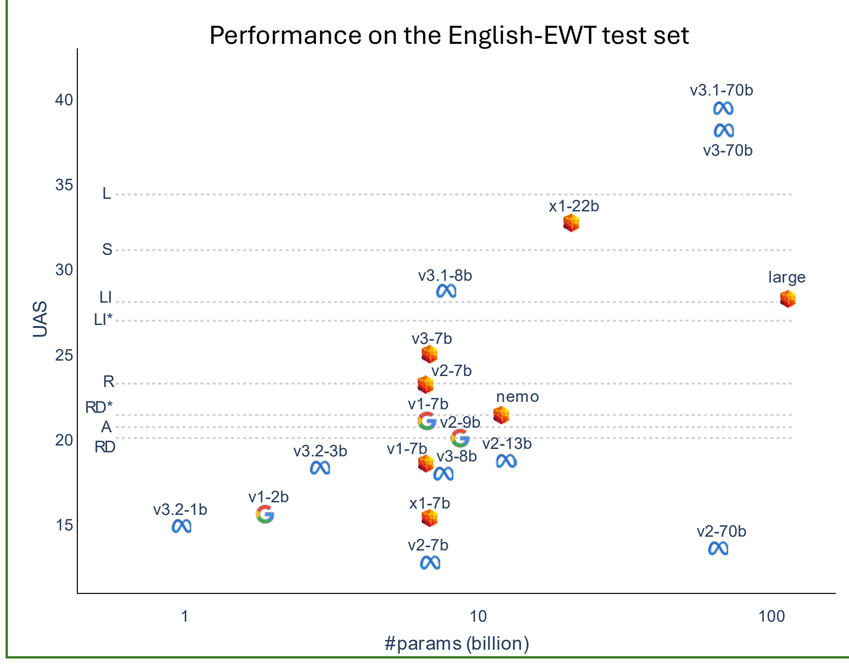
Post-Processing Pipeline

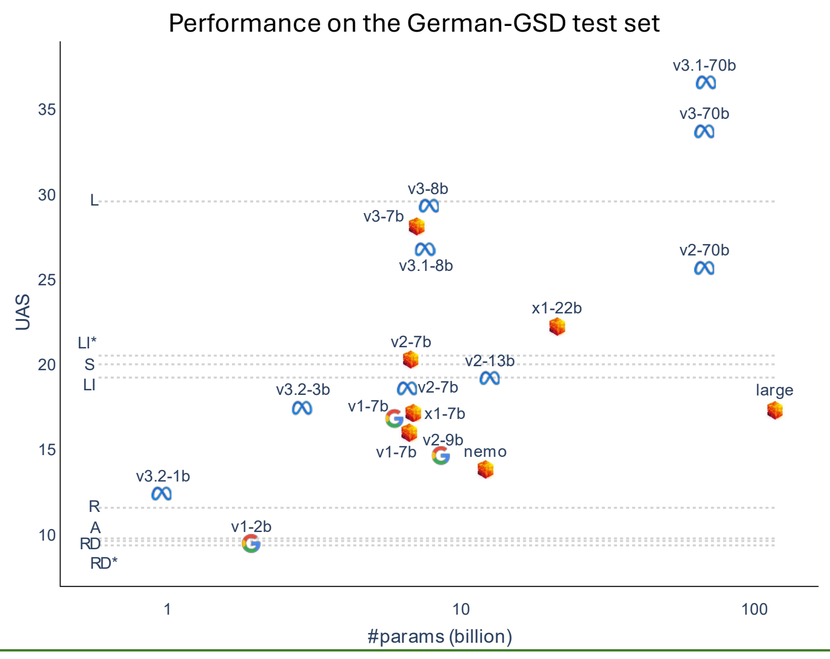
1. Well-formatted output: tabular file, correct fields.
2. Valid dependency tree: unique root, connected component, no cycles.

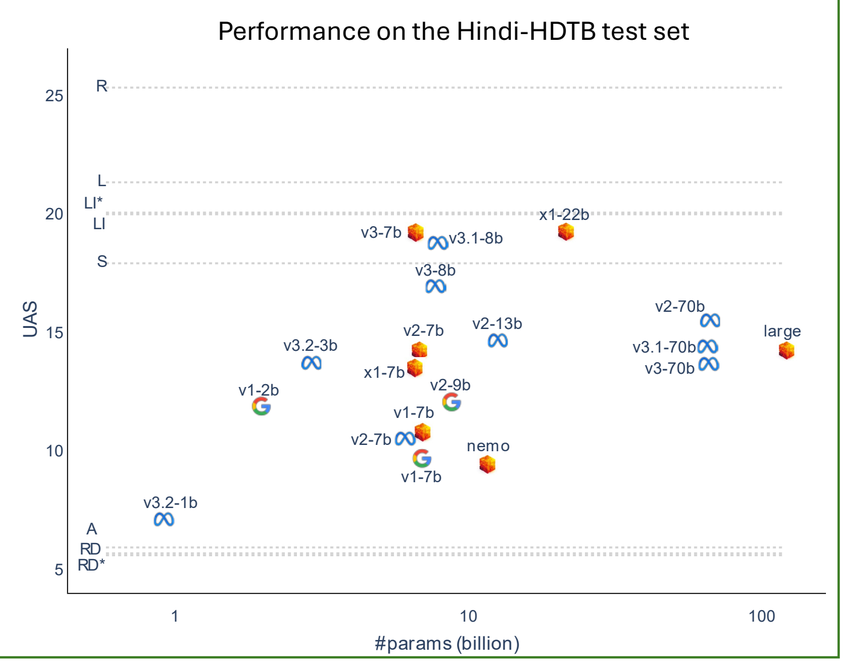


Experiments

- Multilingual benchmark : English, French, German and Hindi from Universal Dependencies.
- Different instruction-based LLMs: Gemma, LLaMA and Mistral.
- Half precision (minimal impact).
- Only the largest models (LLaMA and Mistral) outperform baselines.
- Considerable improvement with the latest versions of LLaMA (v3.2 vs v2).
- Mixtral (22B) offers a trade-off between size and zero-shot performance.







We acknowledge the European Research Council (ERC), which has funded this research under the Horizon Europe research and innovation programme (grant agreement No 101100615). We acknowledge grants SCANNER-UDC (PID2020-113230RB-C21) funded by MICIU/AEI/10.13039/501100011033; GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and ERDF, EU; LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU; and TSI-100925-2023-1 funded by Ministry for Digital Transformation and Civil Service and "NextGenerationEU" PRTR; as well as funding by Xunta de Galicia (ED431C 2024/02), and Centro de Investigación de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

UNIVERSIDADE DA CORUÑA · XUNTA DE GALICIA · citic · LYS · Funded by the European Union · MINISTERIO DE CIENCIA E INNOVACIÓN · UNIÓN EUROPEA · AGENCIA ESTATAL DE INVESTIGACIÓN