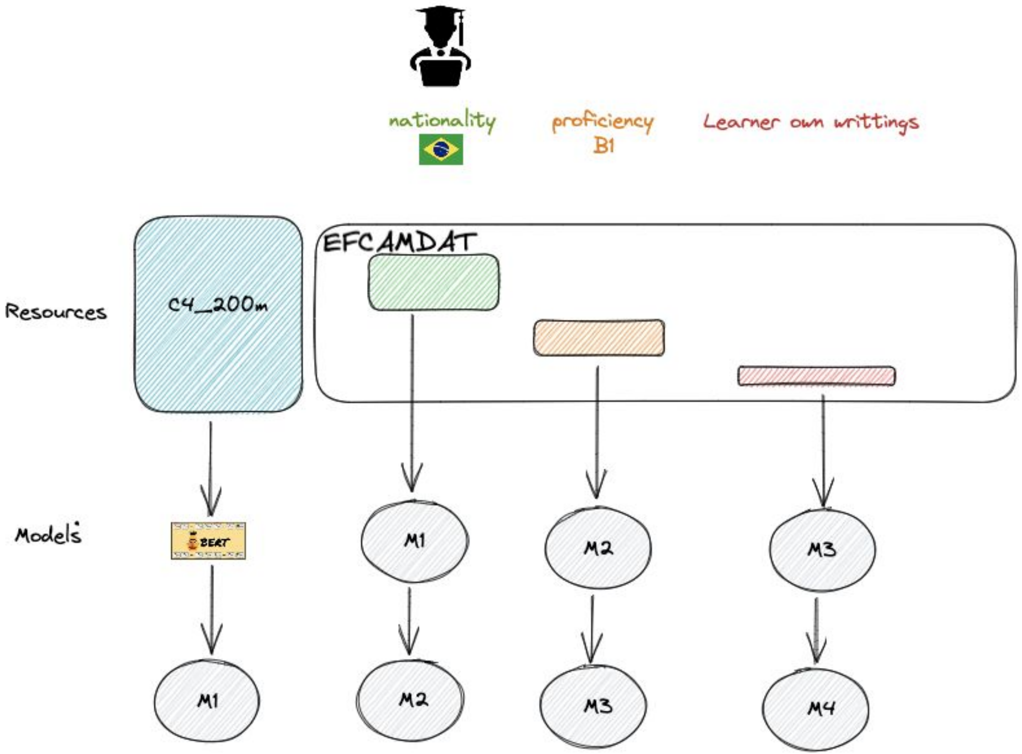
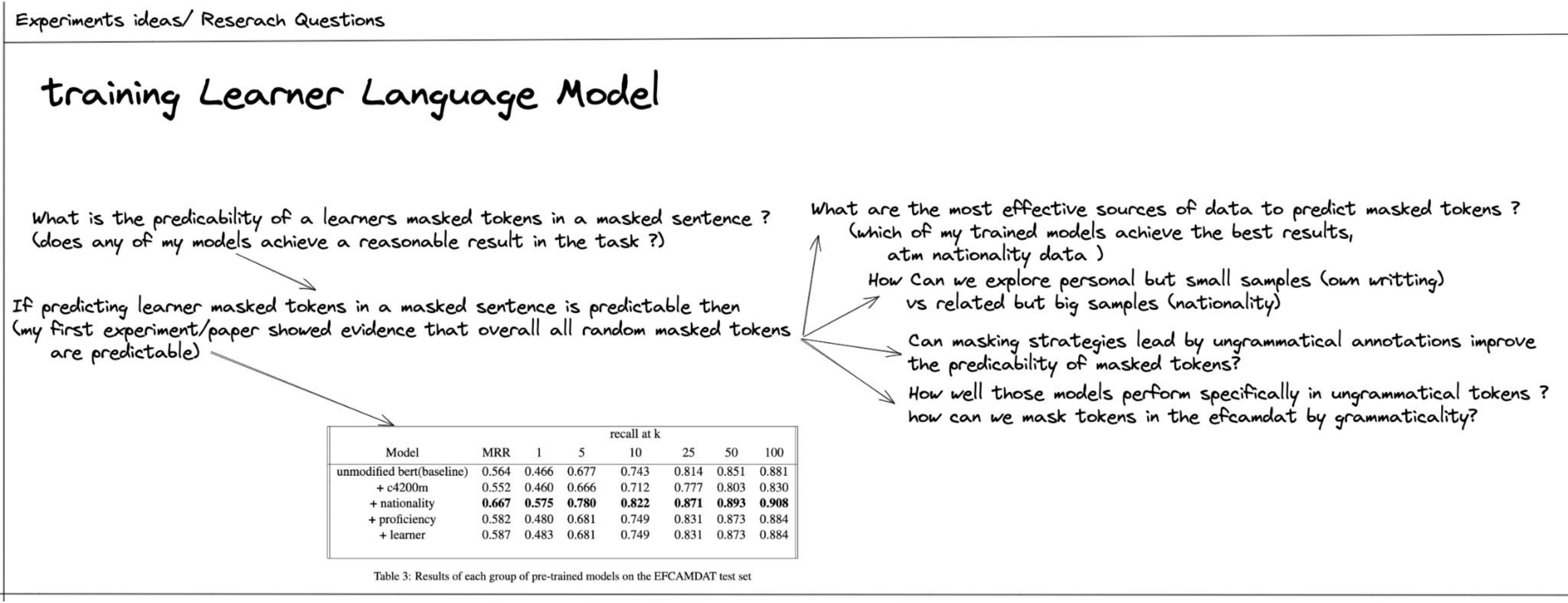
The Task

training Learner Language Model



Some thoughts



Experiments ideas/ Reserach Questions

training Learner Language Model

What is the predicability of a learners masked tokens in a masked sentence ?
(does any of my models achieve a reasonable result in the task ?)

If predicting learner masked tokens in a masked sentence is predictable then
(my first experiment/paper showed evidence that overall all random masked tokens
are predictable)

What are the most effective sources of data to predict masked tokens ?
(which of my trained models achieve the best results,
atm nationality data)

How Can we explore personal but small samples (own writting)
vs related but big samples (nationality)

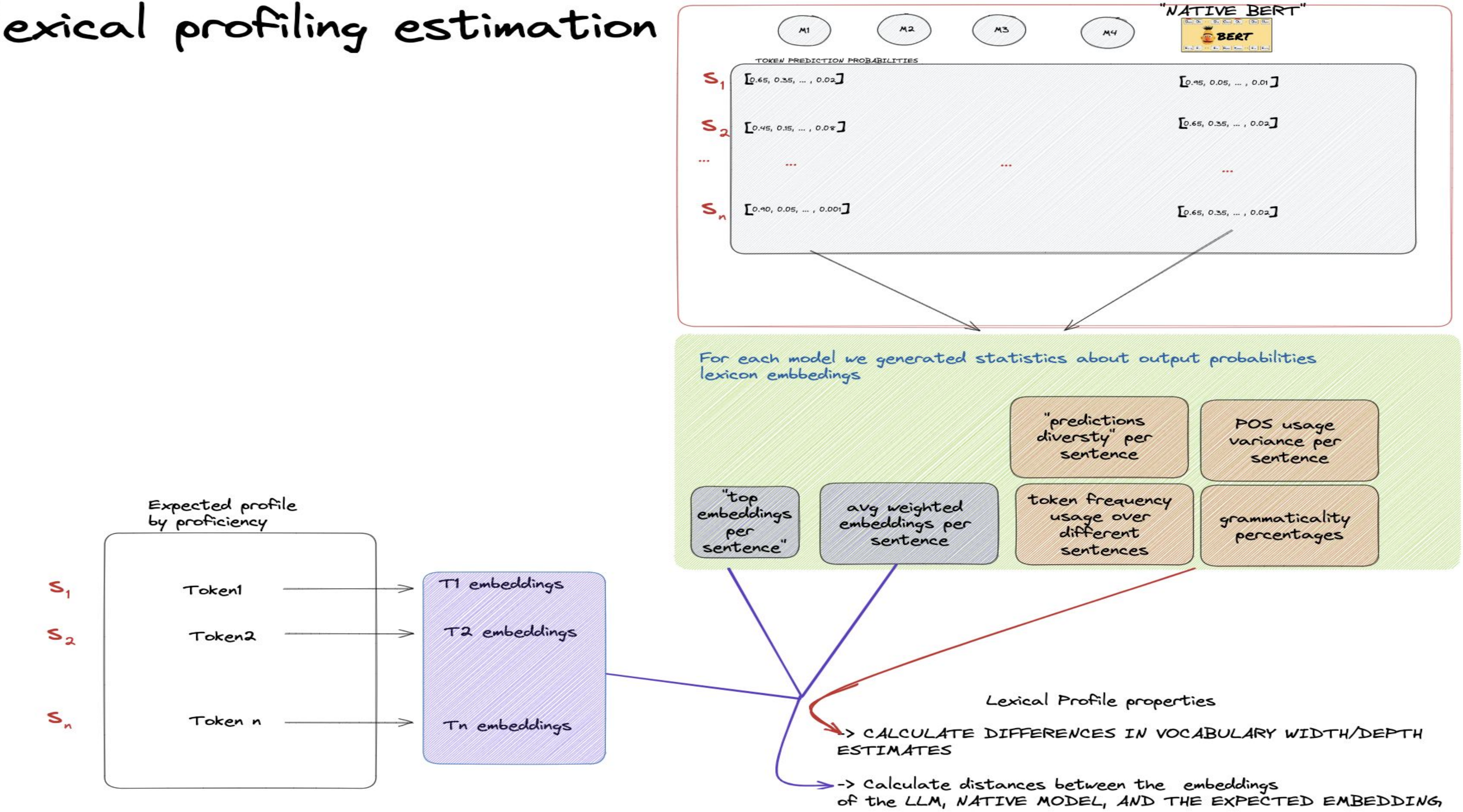
Can masking strategies lead by ungrammatical annotations improve
the predicability of masked tokens?

How well those models perform specifically in ungrammatical tokens ?
how can we mask tokens in the efcamdat by grammaticality?

		recall at k					
Model	MRR	1	5	10	25	50	100
unmodified bert(baseline)	0.564	0.466	0.677	0.743	0.814	0.851	0.881
+ c4200m	0.552	0.460	0.666	0.712	0.777	0.803	0.830
+ nationality	**0.667**	**0.575**	**0.780**	**0.822**	**0.871**	**0.893**	**0.908**
+ proficiency	0.582	0.480	0.681	0.749	0.831	0.873	0.884
+ learner	0.587	0.483	0.681	0.749	0.831	0.873	0.884

Table 3: Results of each group of pre-trained models on the EFCAMDAT test set

lexical profiling estimation


"NATIVE BERT"
M1
M2
M3
M4
BERT
TOKEN PREDICTION PROBABILITIES
S1
[0.65, 0.35, ... , 0.02]
[0.95, 0.05, ... , 0.01]
S2
[0.45, 0.15, ... , 0.08]
[0.65, 0.35, ... , 0.02]
...
...
...
...
Sn
[0.90, 0.05, ... , 0.001]
[0.65, 0.35, ... , 0.02]

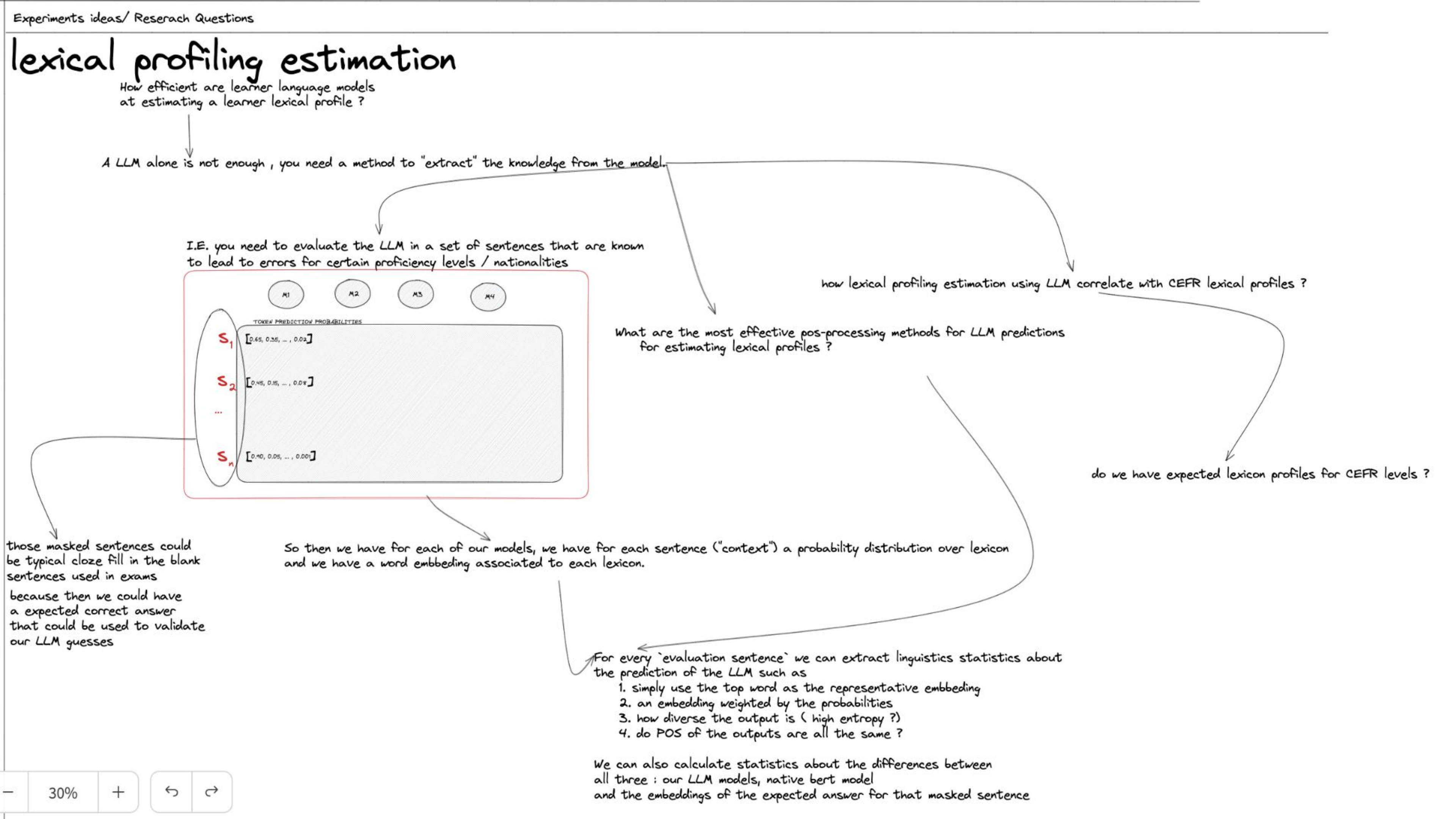
For each model we generated statistics about output probabilities lexicon embbedings

"predictions diversty" per sentence
POS usage variance per sentence
"top embeddings per sentence"
avg weighted embeddings per sentence
token frequency usage over different sentences
grammaticality percentages
Expected profile by proficiency
S1
Token1
T1 embeddings
S2
Token2
T2 embeddings
Sn
Token n
Tn embeddings

Lexical Profile properties

-> CALCULATE DIFFERENCES IN VOCABULARY WIDTH/DEPTH ESTIMATES

-> Calculate distances between the embeddings of the LLM, NATIVE MODEL, AND THE EXPECTED EMBEDDING



Experiments ideas/ Reserach Questions

lexical profiling estimation

How efficient are learner language models
at estimating a learner lexical profile ?

A LLM alone is not enough , you need a method to "extract" the knowledge from the model.

I.E. you need to evaluate the LLM in a set of sentences that are known
to lead to errors for certain proficiency levels / nationalities

those masked sentences could
be typical cloze fill in the blank
sentences used in exams

because then we could have
a expected correct answer
that could be used to validate
our LLM guesses

So then we have for each of our models, we have for each sentence ("context") a probability distribution over lexicon
and we have a word embbeding associated to each lexicon.

What are the most effective pos-processing methods for LLM predictions
for estimating lexical profiles ?

how lexical profiling estimation using LLM correlate with CEFR lexical profiles ?

do we have expected lexicon profiles for CEFR levels ?

For every `evaluation sentence` we can extract linguistics statistics about
the prediction of the LLM such as

1. simply use the top word as the representative embbeding
2. an embedding weighted by the probabilities
3. how diverse the output is (high entropy ?)
4. do POS of the outputs are all the same ?

We can also calculate statistics about the differences between
all three : our LLM models, native bert model
and the embeddings of the expected answer for that masked sentence