A Platform

The [redacted] platform and its sister platform [redacted] are available free of charge. Users agree to the sharing of their input for research purposes. For a screenshot of the [redacted] user interface, see Figure 1.

B Limitations

Dictionary look-up events are rare, sparse, and noisy. While DLU includes more than 8,800 look-up events among 260,000 content word tokens, these features of look-up events inherently limit model performance and some applications. The additionally released chatbot-dialogue dataset is smaller, and therefore its usefulness is limited to evaluation.

Our data is exclusive to English language texts and the first languages of the learners who performed click actions are unevenly distributed (see Table 10). The same is true for CEFR levels. Further personalisation would require more even data distribution.

Due to compute restrictions, we focused on models with comparatively few parameters, although we do include evaluation on LLMs such as LLaMA-3.2-1B. Since we and others (Smădu et al., 2024) found that model size does not appear to predict model performance well, we believe that this restriction poses no major problems. Our focus is on using publicly available models, ensuring replicability.

C Safety and Privacy Considerations

The information in the DLU data poses few risks. While we release information about learner L1 and estimated CEFR-level, personal identification is practically impossible since this information is very broad and the lookup patterns themselves are specific to the platform.

The additional chatbot-dialogue data we release should be handled with greater care, because it includes user input and the chatbot model was not filtered for sensitive content (*reference redacted for peer-review*). As described above (see Section 4), we have manually filtered the dataset and removed critical personal information about the chat participants, e.g. changing first names.

D Dataset Description

For the overall description of the DLU dataset, see Section 3. Further description of CEFR levels and first languages (L1s) across the dataset can be found in tables 6 to 8 and 10 to 12.

	B2	B1	A2	C1	C2	C2+	sum
all	228	242	112	55	17	9	663
train	208	227	108	52	15	6	616
dev	29	44	11	4	1	1	90
test	26	10	11	14	4	3	68

Table 6: Self-reported CEFR levels of users.

	A	B	C	UNK	sum
all	135	324	35	169	663
train	123	302	34	157	616
dev	21	49	6	14	90
test	13	37	5	13	68

Table 7: CEFR levels for users as estimated by essays from [redacted]

D.1 Format of the Data

The data is formatted as a document-level token-classification task. Tokenisation follows the [redacted] pipeline used by [redacted]. For each token a label is provided, with the default label -100 used for non-content word tokens.

Example

Text	Taco	Bell	restaurants	decided	Wednesday	to	remove	…
Labels	0	0	0	0	0	-100	1	…

A 0 label indicates no click, a 1 a click. -100 indicates non-content word POS. A text is a document, i.e. an entire WikiNews article.

E Ensemble Baseline

The classifiers used for the ensemble model are (using sklearn class names):

1. `RandomForestClassifier`
2. `GradientBoostingClassifier`
3. `HistGradientBoostingClassifier`
4. `MLPClassifier`
5. `LogisticRegression`
6. `BaggingClassifier`

They were combined using the sklearn `VotingClassifier` class, which was set to soft

	A	B	C	UNK	sum
all	270	669	116	272	1327
train	229	577	97	240	1143
dev	23	53	8	17	101
test	18	39	11	15	83

Table 8: CEFR levels as estimated by essays from [redacted] across documents by users (i.e. some users and WikiNews articles appear more than once in this table).

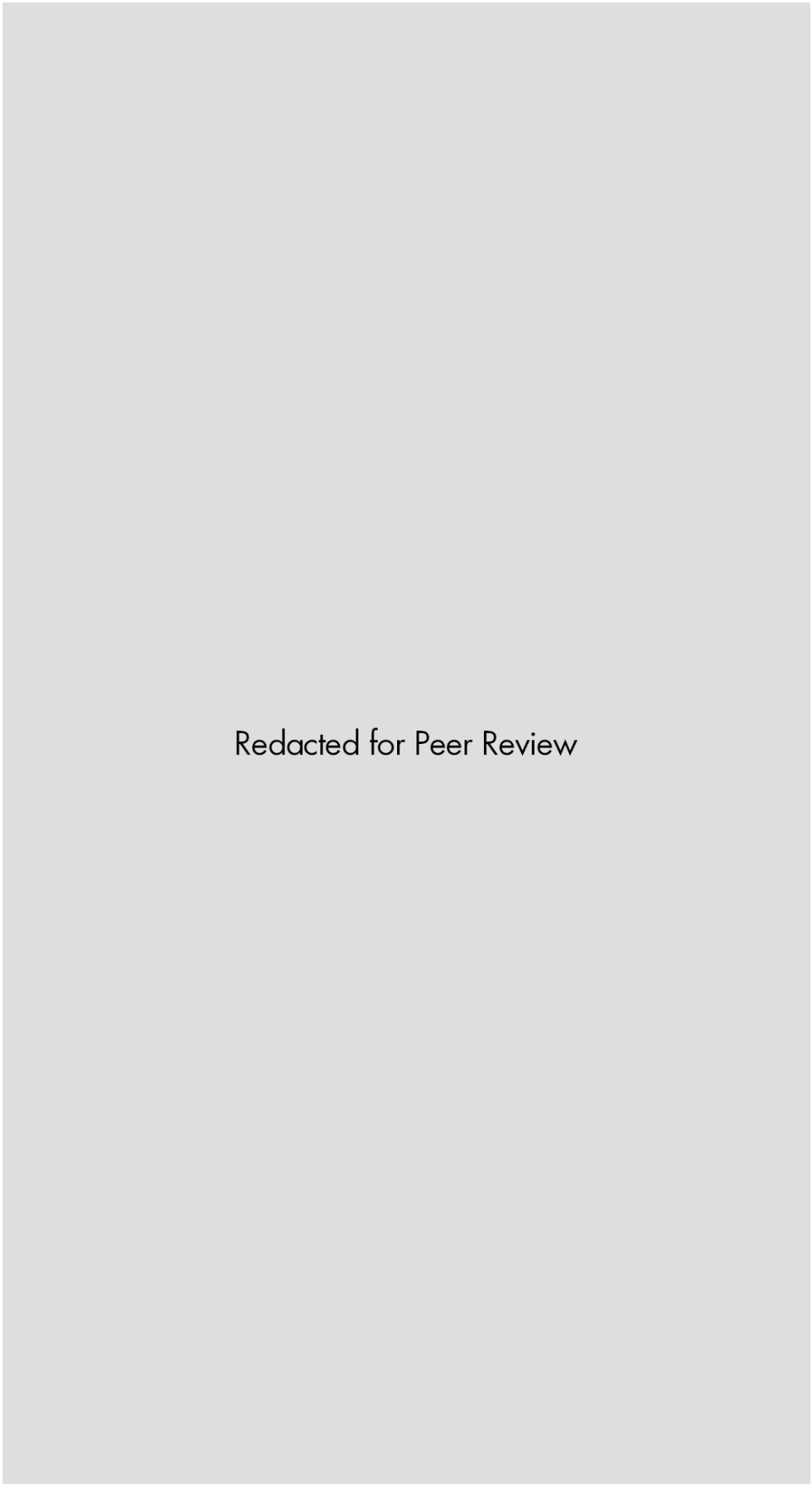

Figure 1: Screenshot of [redacted] platform with information provided by lookup of the word “export”.

	A2	B1	B2	C1	C2	unk	sum
all	2102	2540	1638	522	6	2050	8858
train	1882	2295	1424	343	6	1872	7822
dev	143	139	71	155	0	122	630
test	77	106	143	24	0	56	406

Table 9: Look-up events across CEFR levels as estimated by essays from [redacted].

	ar	bg	ca	cs	de	en	es	fa	fr	hi	hu	id	it	ja	jv	ka	ml	my	ne	pt	ro	ru	sr	ta	te	tr	ur	vi	zh	unk	sum
all	5	1	2	3	2	12	93	2	4	1	1	2	27	1	1	1	1	1	1	14	1	4	1	2	1	27	1	6	7	438	663
train	4	0	2	3	0	10	83	2	4	1	1	2	24	1	0	1	1	1	1	13	1	4	1	1	1	23	1	6	7	417	616
dev	1	0	0	0	1	1	15	0	1	0	0	0	6	1	1	0	0	1	0	2	0	0	0	0	0	4	0	1	3	52	90
test	2	1	0	0	1	3	16	1	0	0	1	0	2	0	0	0	0	1	0	4	0	0	0	1	0	3	0	0	0	32	68

Table 10: Users per L1. For experiments, less frequent languages are merged into the unknown category (unk),

voting. No systematic hyperparameter tuning was required.

The used features were:

- The frequency baseline score as described in Section 6.
- Relative position of the token in the text, defined as the proportion of seen tokens for the first 1000 tokens.
- Proportion of look-up events by user on splits used for training.
- CEFR-level as estimated by essays submitted by the user.
- Count of definitions for the word in the *Cambridge Advanced Learner's Dictionary*.
- Whether the word type occurred before in the text.
- Proportion of people who did not know the word type as retrieved from the ratings by Brysbaert et al. (2014).

For missing values, the average was used. To address label imbalance we upsampled positive cases to achieve a proportion of 1-to-1. For the additionally added positively labelled data, we added small Gaussian noise to the frequency score, proportion of look-up event by user, the relative position.

	ar	en	es	it	pt	tr	vi	zh	unk	sum
all	12	19	169	70	29	48	10	15	955	1327
train	8	14	135	62	23	40	9	12	840	1143
dev	1	2	16	6	2	5	1	3	65	101
test	3	3	18	2	4	3	0	0	50	83

Table 11: L1s across documents seen by users (i.e. some users and articles appear multiple times in this table).

	ar	en	es	it	pt	tr	vi	zh	unk	sum
all	5	12	93	27	14	27	6	7	472	663
train	4	10	83	24	13	23	6	7	446	616
dev	1	1	15	6	2	4	1	3	57	90
test	2	3	16	2	4	3	0	0	38	68

Table 12: L1s across users – less frequent languages merged into unknown (unk). This merging process is used for our transformer models.

split	chats	clicks	con. tokens
D-chat	25	5	10027
D-read	26	67	33130

Table 13: Description of data and splits, including the number of content tokens for chatbot dialogues.

F Neural Models

The models used are described in Table 14. We used the LLaMA 3.1-8B, rather than a LLaMA 3.2 version, because it was closer to the size of the Gemma model.

model	hf-name	approach
Longformer	allenai/longformer-base-4096	finetuning
LLaMA 3.2	meta-llama/Llama-3.2-1B	finetuning
LLaMA Instruct	unsloth/Meta-Llama-3.1-8B-Instruct	prompting
Gemma	unsloth/gemma-2-9b-it	prompting

Table 14: Details of models used, including name on huggingface hub and experimental approach.

F.1 Hyperparameters

The datasets for the different tasks strongly differ in input length. Both the SEP and DLU dataset operate on the document-level, but while DLU consists of WikiNews texts, the SEP consists of student essays. The 2018 CWI dataset (Yimam et al., 2017) is on the sentence level, i.e. the inputs are much shorter than for the other datasets. To work with these different datasets, we found it necessary to change the hyperparameter space, in particular the space for the training batch size.

The hyperparameter spaces as well as the selected hyperparameters are described in tables 15 to 17. For each combination of model and loss function, we run 20 trials without pruning, where the searches were performed with Optuna. Additional settings for Optuna, such as using the log space are noted in the table. The target metric for maximization was the AUC.

G Prompting

We use two prompt templates, one for zero-shot and one for few-shot inference. Both prompts in-

	Space	Info	Longformer (ROC*)	Longformer (BCE)	LLaMA (ROC*)	LLaMA (BCE)
Epochs	[1, 30]		25	14	30	14
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	3.6×10^{-6}	6.7×10^{-5}	3.7×10^{-5}	2.4×10^{-4}
Pos. Weight	[0.8, 30]	BCE only	-	0.81	-	29
γ	[0.05, 0.75]	ROC* only	0.59	-	0.05	-
Sample Size	[300, 10000]	ROC*, step size=100	6600	-	300	-
Batch Size (p.D.)	[4, 14]	step size = 2	12	8	4	12

Table 15: Hyperparameter space and selected hyperparameters for DLU prediction models. We report the per device batch size. The number of devices was always set to 4.

	Space	Info	Longformer (ROC*)	Longformer (BCE)	LLaMA (ROC*)	LLaMA (BCE)
Models finetuned only on CWI						
Epochs	[1, 30]		8	11	22	11
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	7.0×10^{-5}	4.6×10^{-5}	1.1×10^{-4}	2.3×10^{-5}
Pos. Weight	[0.8, 30]	BCE only	-	29.9	-	26.5
γ	[0.05, 0.75]	ROC* only	0.69	-	0.45	-
Sample size	[300, 10000]	ROC*, step size=100	3400	-	4200	-
Batch size (p.D.)	[8, 80]	step size = 2	48	10	50	72
Models finetuned on DLU and then on CWI						
Epochs	[1, 30]		10	21	29	8
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	1.7×10^{-4}	7.4×10^{-5}	8.0×10^{-5}	9.2×10^{-5}
Pos. Weight	[0.8, 30]	BCE only	-	12.72	-	23.54
γ	[0.05, 0.75]	ROC* only	0.09	-	0.52	-
Sample size	[300, 10000]	ROC*, step size=100	900	-	4100	-
Batch size (p.D)	[8, 80]	step size = 2	36	30	52	18

Table 16: Hyperparameter space and selected hyperparameters for CWI prediction models. We report the per device batch size. The number of devices was always set to 4.

	Space	Info	Longformer (ROC*)	Longformer (BCE)	LLaMA (ROC*)	LLaMA (BCE)
Models finetuned only on SEP task						
Epochs	[1, 30]		24	10	10	6
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	3.1×10^{-5}	1.0×10^{-5}	8.6×10^{-6}	2.3×10^{-5}
Pos. Weight	[0.8, 30]	BCE only	-	15.08	-	16.90
γ	[0.05, 0.75]	ROC* only	0.34	-	0.65	-
Sample size	[300, 10000]	ROC*, step size=100	2600	-	9100	-
Batch size (p.D.)	[4, 44]	step size = 2	36	34	38	18
Models finetuned on DLU and then on SEP task						
Epochs	[1, 30]		28	26	7	1
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	2.1×10^{-8}	1.5×10^{-9}	1.6×10^{-8}	2.0×10^{-6}
Pos. Weight	[0.8, 30]	BCE only	-	22.96	-	1.33
γ	[0.05, 0.75]	ROC* only	0.18	-	0.62	-
Sample size	[300, 10000]	ROC*, step size=100	4300	-	9900	-
Batch size (p.D)	[4, 44]	step size = 2	40	16	20	32

Table 17: Hyperparameter space and selected hyperparameters for SEP prediction models. We report the per device batch size. The number of devices was always set to 4.

struct the LLM to consider a paragraph of text and the learner's English CEFR level. The the models are asked to predict which words the learner is likely unfamiliar with, andreturn these words in a JSON format. The zero-shot prompt directly provides the task instructions and desired output format, while the few-shot prompt includes three illustrative examples of different learners' word choices in different paragraphs of text.

G.1 Prompts

```
CLICK_DATA_APPROXIMATION_PROMPT = {'system': """ # Task Introduction You are an AI assistant now doing a language test. You will receive a paragraph of text. you will need to predict based on your user's English level what words the user might click on(The user will click on the words he or she is not familiar with.

–

# About the user's english level A1: Can write personal information (e.g. likes and dislikes, family, pets) using simple words, phrases and sentences.

A2: Can write a series of simple phrases and sentences, linked with words like 'and', 'but' and 'because'.

B1: Can write straightforward texts about familiar topics or simple information and ideas. Can link sentences into a connected text.

B2: Can write clear, detailed texts on different subjects. Can use information and arguments from other sources in their writing.

C1: Can write clear, well-structured, detailed texts on complex subjects, showing the important issues, giving examples and writing a conclusion if appropriate. Can use the correct style of writing relevant to the target reader.

C2: Can write clear, smoothly flowing, complex
```

```
texts in an appropriate and effective style. Can use a logical structure which helps the reader find the main points.
  –
  # Expected Output Your answers should be formatted in JSON format with following keys and values: 1. output_tokens: a list of tokens that you predict the user will click on, each token should appear only once 2. reason: a short string explaining your prediction of the tokens
  NOTE: please make sure the output tokens are unique. each token in the list should appear only once """, 'user': """
  # task detail
  The user's english level is:
  {cefr_level}
  The paragraph you need to predict on:
  {paragraph_text}
  The tokens in the paragraph:
  {tokens}
  Respond only with valid JSON.
  –
  """ }
  CLICK_DATA_APPROXIMATION_FEWSHOT_PROMPT = {'system': """ # Task Introduction You are an AI assistant now doing a language test. You will receive a paragraph of text. you will need to predict based on your user's English level what words the user might click on(The user will click on the words he or she is not familiar with.
  –
  # About the user's english level
  A1: Can write personal information (e.g. likes and dislikes, family, pets) using simple words, phrases and sentences.
  A2: Can write a series of simple phrases and sentences, linked with words like 'and', 'but' and 'because'.
  B1: Can write straightforward texts about familiar topics or simple information and ideas. Can link sentences into a connected text.
  B2: Can write clear, detailed texts on different subjects. Can use information and arguments from other sources in their writing.
  C1: Can write clear, well-structured, detailed texts on complex subjects, showing the important issues, giving examples and writing a conclusion if appropriate. Can use the correct style of writing relevant to the target reader.
  C2: Can write clear, smoothly flowing, complex texts in an appropriate and effective style. Can use a logical structure which helps the reader find the main points.
  –
  # Expected Output Your answers should be formatted in JSON format with following keys and values: 1. output_tokens: a list of tokens that you predict the user will click on, each token should appear only once
  2. reason: a short string explaining your prediction of the tokens
  NOTE: please make sure the output tokens are unique. each token in the list should appear only once
  –
  # Examples Here are some examples from user of the same english level as the one you are goingto mimic.
  ## Example1:
  {example1}
  ## Example2:
  {example2}
  ## Example3:
  {example3}
  """, 'user': """
  # task detail
  The user's english level is:
  {cefr_level}
  The paragraph you need to predict on:
  {paragraph_text}
  The tokens in the paragraph:
  {tokens}
  Respond only with valid JSON.
  –
  """ }
```

H Significance Tests

We perform a two-sided permutation test using SciPy (Virtanen et al., 2020). We set `permutation_type='samples'` and `random_state='1848'`. The number of permutations is left at the default 9999. The test statistics and associated p-values can be found in tables tables 18 to 20.

The Bonferroni-correct p-value is 0.0027. We rounded the digits of the threshold using the floor, as this makes the significance test more restrictive.

	Metric	Statistic	p-Value
Longformer	AUC	2.9×10^{-2}	2.4×10^{-1}
LLaMA	AUC	6.3×10^{-2}	4.0×10^{-4}

Table 18: Significance tests for DLU. The tests concern whether using the ROC* vs. the BEC loss changes the AUC.

	Metric	Loss	Statistic	p-Value
Longformer	AUC	bce	4.8×10^{-4}	8.5×10^{-1}
Longformer	F1	bce	7.3×10^{-3}	4.7×10^{-2}
Longformer	AUC	roc	7.3×10^{-4}	9.0×10^{-1}
Longformer	F1	roc	2.6×10^{-1}	2.0×10^{-4}
LLaMA	AUC	bce	4.2×10^{-4}	9.4×10^{-1}
LLaMA	F1	bce	1.7×10^{-2}	4.0×10^{-4}
LLaMA	AUC	roc	5.5×10^{-3}	7.7×10^{-2}
LLaMA	F1	roc	1.2×10^{-3}	7.9×10^{-1}

Table 19: Significance tests for CWI task, testing whether models finetuned on DLU first perform differently on F_1 or AUC.

I Processing of CWI

The CWI dataset we used (Yimam et al., 2017, 2018) provides one data row for each labelled word, even if these words occur in the same sentences. To reduce training time and make the processing more similar to DLU, we treated these words as occuring together during training. For evaluation, we again made one prediction per input, as in the original

	Metric	Loss	Statistic	p-Value
LLaMA	AUC	bce	1.0×10^{-2}	3.9×10^{-2}
LLaMA	F1	bce	1.3×10^{-2}	5.5×10^{-2}
LLaMA	AUC	roc	1.8×10^{-4}	5.1×10^{-1}
LLaMA	F1	roc	1.2×10^{-3}	7.6×10^{-2}
Longformer	AUC	bce	1.9×10^{-3}	3.9×10^{-2}
Longformer	F1	bce	4.0×10^{-4}	8.9×10^{-1}
Longformer	AUC	roc	8.4×10^{-4}	2.0×10^{-1}
Longformer	F1	roc	8.1×10^{-4}	8.0×10^{-1}

Table 20: Significance tests for SEP task, testing whether models finetuned on DLU first perform differently on F_1 or AUC.

CWI dataset for comparability. This might have affected our performance negatively, explaining some of the difference to the results reported by Smădu et al. (2024).

J Further Discussion of Results

Using an adaptive threshold for the F_2 (aF_2) consistently improves the performance of the baseline further, which is not always the case for the transformer models. This suggests that the decision threshold for transformer models is context dependent and cannot be transferred between splits. Furthermore, it shows that the simple frequency baseline can be further improved with simple.

As a result of the different effect of the adaptive threshold, the highest F_2 value (23.4%) by a transformer model (Longformer ROC*) is higher than the aF_2 (21%) of the frequency baseline, even though the baseline achieves the highest aF_2.

K Additional Results

In Section 7 we report results on the DLU train split, but as we release only the dev split with this paper, we report the results on this split in Table 22. The training method was the same as for the results on the test split.

The results might be affected by the same documents being repeated in the evaluation split (dev or test) because more than one user interacted with it. To investigate this effect, we also evaluated on these splits after removing all but one randomly selected version of each document, i.e. the look-up data for one random user per document. The results are shown in tables 24 and 25. The adaptive threshold for the aF_2 is the same as for the original evaluation.

		A				B				C				unk				All				D-read			
		F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC
Gemma-Inst.	zeroshot	10.3	15.1	-	56.5	14.1	18.0	-	57.6	**13.4**	20.9	-	64.1	9.1	14.3	-	57.3	12.2	17.0	-	57.7	2.2	4.6	-	57.4
	fewshot	10.2	16.1	-	57.4	12.5	17.8	-	57.5	12.8	**21.9**	-	67.2	10.4	16.4	-	59.1	11.7	17.6	-	58.4	-	-	-	-
LLaMA-Inst.	zeroshot	8.7	16.4	-	58.6	7.8	12.8	-	53.0	5.3	10.0	-	56.6	6.9	13.0	-	57.2	7.6	13.5	-	55.0	1.0	2.4	-	56.1
	fewshot	8.3	15.1	-	56.7	7.6	12.4	-	52.7	4.5	8.9	-	55.5	3.8	7.1	-	49.7	6.7	11.7	-	53.2	-	-	-	-
LLaMA	ROC*	0.0	0.0	8.0	79.3	0.0	0.0	5.4	66.7	0.0	0.0	0.0	70.2	0.0	0.0	7.5	70.2	0.0	0.0	5.9	69.9	0.0	0.0	1.4	76.2
	BCE	8.8	17.0	14.3	66.3	11.2	18.9	17.2	62.1	6.2	11.3	9.6	63.8	6.2	12.6	13.0	65.5	9.2	16.7	15.4	63.5	**2.9**	**5.8**	4.2	77.7
Longformer	BCE	0.0	0.0	18.9	73.3	0.0	0.0	15.8	69.8	0.0	0.0	9.9	67.2	0.0	0.0	15.1	73.7	0.0	0.0	15.9	70.8	0.0	0.0	**8.6**	77.0
	ROC*	12.9	23.1	19.3	78.5	16.1	26.1	21.6	72.0	4.9	9.5	14.9	66.8	12.4	21.1	19.2	76.6	13.8	23.4	20.3	73.7	1.8	3.8	3.2	83.5
Baseline	freq.	8.7	18.9	24.7	75.8	9.6	20.6	23.1	71.4	4.2	9.9	10.8	72.3	5.7	12.9	16.6	72.2	8.1	17.7	21.0	72.5	0.9	2.2	3.3	**84.9**
	ens.	**20.5**	**30.2**	**33.5**	**85.6**	**17.4**	**26.2**	**27.9**	**76.3**	11.9	18.4	**19.9**	**82.0**	**13.9**	**24.0**	**22.0**	**80.8**	**17.0**	**26.1**	**27.4**	**79.3**	-	-	-	-

Table 21: Prediction results on the DLU test split, but for the prompting model, we take all occurrences of a word listed by the prompted model to be looked-up. (Results on non-prompting models are unchanged.)

		A				B				C				unk				All			
		F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC
Gemma-Inst.	zeroshot	11.7	14.0	-	54.7	9.9	13.3	-	55.7	12.0	9.5	-	52.2	13.0	15.2	-	55.2	11.2	13.1	-	54.4
	fewshot	10.8	12.5	-	53.9	9.4	12.4	-	55.1	9.7	7.5	-	51.5	12.9	18.0	-	56.3	10.6	12.8	-	54.1
LLaMA-Inst.	zeroshot	8.9	9.4	-	52.6	9.0	14.3	-	56.6	15.1	16.1	-	51.7	6.1	8.5	-	49.7	9.4	12.6	-	53.4
	fewshot	11.2	15.6	-	55.1	6.1	10.4	-	53.1	12.9	12.9	-	51.0	9.1	13.2	-	52.6	8.4	12.4	-	52.7
LLaMA	BCE	15.4	**25.4**	20.9	69.0	7.8	13.4	10.1	58.2	15.3	14.6	10.9	62.2	13.8	24.2	21.7	67.4	11.8	18.9	15.5	62.1
	ROC*	0.0	0.0	13.2	71.6	0.0	0.0	7.1	64.8	0.0	0.0	1.6	51.2	0.0	0.0	9.9	68.5	0.0	0.0	7.9	63.3
Longformer	BCE	0.0	0.0	22.0	**73.3**	0.0	0.0	16.1	**71.1**	0.0	0.0	9.3	56.8	0.0	0.0	17.8	72.9	0.0	0.0	16.3	68.3
	ROC*	**17.0**	25.4	18.0	71.8	10.2	19.1	16.6	69.5	15.0	17.9	10.0	51.5	15.3	23.8	19.9	71.7	12.8	21.0	16.2	65.6
Baseline	freq.	9.8	20.6	22.4	63.2	6.5	14.6	17.0	68.3	**22.9**	**39.7**	**37.7**	62.1	11.4	23.8	27.2	69.8	9.7	20.6	22.7	65.7
	ens.	14.7	23.3	**23.5**	68.7	**11.3**	**20.1**	**18.8**	69.2	22.5	24.0	31.2	**65.1**	**21.7**	**33.0**	**31.7**	**76.9**	**15.2**	**23.8**	**23.9**	**69.0**

Table 22: Prediction results on the DLU dev split. "aF2" stands for F2 with a adaptive threshold, as discussed in Section 5.

		A				B				C				unk				All			
		F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC
Gemma-Inst.	zeroshot	11.9	14.7	-	55.0	8.8	12.1	-	54.9	15.0	12.3	-	53.3	13.5	17.3	-	56.1	11.3	13.8	-	54.6
	fewshot	11.2	14.8	-	54.8	8.5	11.9	-	54.7	12.7	10.2	-	52.4	12.4	19.8	-	57.3	10.7	14.3	-	54.6
LLaMA-Inst.	zeroshot	10.1	12.4	-	53.6	6.9	12.5	-	55.1	21.4	26.8	-	55.4	6.9	11.1	-	49.5	9.4	14.8	-	54.1
	fewshot	10.3	16.0	-	54.9	4.9	9.3	-	51.1	20.1	23.8	-	54.6	8.7	14.5	-	52.0	8.3	13.8	-	52.7
LLaMA	BCE	15.4	**25.4**	20.9	69.0	7.8	13.4	10.1	58.2	15.3	14.6	10.9	62.2	13.8	24.2	21.7	67.4	11.8	18.9	15.5	62.1
	ROC*	0.0	0.0	13.2	71.6	0.0	0.0	7.1	64.8	0.0	0.0	1.6	51.2	0.0	0.0	9.9	68.5	0.0	0.0	7.9	63.3
Longformer	BCE	0.0	0.0	22.0	**73.3**	0.0	0.0	16.1	**71.1**	0.0	0.0	9.3	56.8	0.0	0.0	17.8	72.9	0.0	0.0	16.3	68.3
	ROC*	**17.0**	25.4	18.0	71.8	10.2	19.1	16.6	69.5	15.0	17.9	10.0	51.5	15.3	23.8	19.9	71.7	12.8	21.0	16.2	65.6
Baseline	freq.	9.8	20.6	22.4	63.2	6.5	14.6	17.0	68.3	**22.9**	**39.7**	**37.7**	62.1	11.4	23.8	27.2	69.8	9.7	20.6	22.7	65.7
	ens.	14.7	23.3	**23.5**	68.7	**11.3**	**20.1**	**18.8**	69.2	22.5	24.0	31.2	**65.1**	**21.7**	**33.0**	**31.7**	**76.9**	**15.2**	**23.8**	**23.9**	**69.0**

Table 23: Prediction results on the DLU dev split, but for the prompting model, we take all occurrences of a word listed by the prompted model to be looked-up. (Results on non-prompting models are unchanged.)

		A				B				C				unk				All			
		F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC
Gemma-Inst.	zeroshot	9.6	11.6	-	52.7	15.8	20.3	-	59.2	0.0	0.0	-	47.2	1.7	2.9	-	48.0	9.3	13.0	-	55.1
	fewshot	20.3	22.0	-	59.1	14.4	17.9	-	57.7	0.0	0.0	-	47.8	6.8	11.7	-	56.4	12.1	16.4	-	57.4
LLaMA-Inst.	zeroshot	11.0	16.9	-	54.7	10.5	18.5	-	58.8	3.0	6.4	-	56.5	4.4	8.4	-	53.1	8.6	15.2	-	57.1
	fewshot	12.4	16.0	-	55.1	6.7	10.3	-	51.8	**6.8**	**14.1**	-	69.8	0.0	0.0	-	43.4	5.9	9.5	-	51.9
LLaMA	ROC**	8.3	7.2	7.2	78.0	17.5	14.1	14.1	**74.8**	0.0	0.0	0.0	**73.9**	4.9	5.3	5.3	61.4	11.5	10.0	10.0	73.8
	BCE	13.5	18.0	18.0	59.9	18.6	25.5	25.5	66.3	0.0	0.0	0.0	30.7	**10.8**	**19.8**	**19.8**	**78.6**	14.3	21.3	21.3	66.5
Longformer	ROC**	18.6	22.7	22.7	**78.9**	**19.9**	**28.8**	**28.8**	74.3	4.2	8.3	**8.3**	64.1	4.8	7.2	7.2	70.9	**15.1**	**22.0**	**22.0**	**74.6**
	BCE	**26.9**	**28.7**	**28.7**	76.6	13.7	15.1	15.1	69.0	0.0	0.0	0.0	71.6	6.8	8.8	8.8	74.2	13.8	16.3	16.3	71.4

Table 24: Prediction results on test split when for each document only one user was randomly selected.

		A				B				C				unk				All			
		F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC	F_1	F_2	aF_2	AUC
Gemma-Inst.	zeroshot	7.5	11.2	-	53.7	4.1	5.7	-	50.8	**26.2**	**21.7**	-	**58.5**	0.0	0.0	-	46.6	8.8	11.2	-	53.7
	fewshot	5.1	7.6	-	51.0	6.2	9.7	-	53.5	0.0	0.0	-	47.7	0.0	0.0	-	47.3	4.4	6.0	-	50.0
LLaMA-Inst.	zeroshot	8.3	14.3	-	56.2	6.3	12.1	-	56.0	14.8	19.4	-	53.2	2.6	6.1	-	59.0	6.9	12.8	-	53.3
	fewshot	4.8	8.4	-	50.1	4.9	9.3	-	52.4	6.7	5.5	-	50.6	3.0	6.4	-	54.8	4.8	8.1	-	49.6
LLaMA	ROC**	5.7	5.6	5.6	64.7	6.1	5.6	5.6	67.6	0.0	0.0	0.0	50.6	**16.7**	**20.8**	**20.8**	76.2	5.0	4.3	4.3	61.0
	BCE	4.5	7.0	7.0	61.8	4.4	6.0	6.0	53.8	7.5	5.7	5.7	54.2	8.0	16.1	16.1	**83.7**	5.5	7.1	7.1	52.1
Longformer	ROC**	**17.4**	**20.0**	**20.0**	71.3	13.3	20.3	20.3	68.5	13.7	15.6	**15.6**	54.5	10.8	20.4	20.4	70.6	13.7	**18.7**	**18.7**	64.9
	BCE	13.0	15.0	15.0	**71.7**	**18.0**	**22.2**	**22.2**	**69.6**	12.8	12.4	12.4	55.0	15.4	20.0	20.0	72.3	**15.3**	17.2	17.2	**66.6**

Table 25: Prediction results on dev split when for each document only one user was randomly selected.