DLU: Dictionary Look-Up Data and Prediction

David Strohmaier[1], Gladys Tyen[1]*, Hongyi Gu[2], Diane Nicholls[3], Zheng Yuan[4, 1], Paula Buttery[1]

[1]ALTA Institute, [2]NetMind.AI
[3]Cambridge University Press & Assessment
[4]The University of Sheffield
Correspondence: david.strohmaier@cl.cam.ac.uk

Abstract

Knowing which words language learners struggle with is crucial for developing personalised education technologies. In this paper, we advocate for the novel task of "dictionary look-up prediction" as a means for evaluating the complexity of words in reading tasks. We release the *Dictionary Look-Up development* dataset (DLU-dev) and the *Dialogue Dictionary Look-Up* dataset (D-DLU), which is based on chatbot dialogues. We demonstrate that dictionary look-up is a challenging task for LLMs (results are presented for LLaMA, Gemma, and Longformer models). We explore finetuning with the ROC* loss function as a more appropriate loss for this task than the commonly used Binary Cross Entropy (BCE). We show that a feature-based model outperforms the LLMs. Finally, we investigate the transfer between DLU and the related tasks of Complex Word Identification (CWI) and Semantic Error Prediction (SEP), establishing new state-of-the-art results for SEP.

1 Introduction

When a learner is reading a text, they may encounter unfamiliar words. When this happens, a learner can choose to seek further information about the word, such as definitions or examples of use. The event of *looking-up* a word is an indication that the word was difficult for the learner in its textual context. By noting the look-up events of many learners, we can discover the relative *contextual lexical complexity* of words for different ability groups; and by collecting look-up data en masse, we can build empirical models of the same. Such models can be used to:

1. Improve readability of texts for specific learner groups;
2. Predict reading competence of learners; or
3. Generate proficiency calibrated test items (e.g. cloze tests).

As a means of evaluating models of *contextual lexical complexity* built from look-up data, we introduce the task of *dictionary look-up prediction*; that is, the task of predicting which words a learner will look up when reading a word in context.

We also introduce the first widely available *Dictionary Look-Up* dataset (DLU). The source of the DLU data is the publicly accessible *Read&Improve* platform,[1] where second language (L2) learners engage in a reading-summarisation task. This dataset captures the words that learners looked up whilst reading a passage of text. Look-ups are recorded within their original context, and metadata regarding the learner is also recorded (specifically, their first language L1, and their estimated language proficiency on the CEFR[2] scale).

With this paper, we release a development portion of this data, DLU-dev, consisting of over 16,000 content word tokens and 630 clicks. The full dataset consists of >260,000 content word tokens and >8,800 lookup events, and is expected to be released for an open participation shared task.

To evaluate the generalisability of *contextual lexical complexity* models built from the DLU dataset, we present a second dataset, the *Dialogue Dictionary Look-Up* dataset (D-DLU). This dataset is sourced from a language learning system that allows learners to look up words in generated chatbot responses (Tyen et al., 2024).

Compared to existing complexity-oriented datasets, such as complex word identification and eye-tracking data (e.g. Paetzold and Specia, 2016; Yimam et al., 2017; Shardlow, 2013; Shardlow et al., 2020; Berzak et al., 2022), our datasets have the following benefits:

[*] Now at Google DeepMind.

[1] https://readandimprove.englishlanguageitutoring.com/

[2] The Common European Framework of Reference for Languages defines levels of language competence (CoE, 2020).

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 481–501
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

1. **High external validity:** It provides behavioural patterns of L2 learners engaged in a naturalistic language learning task.
2. **Document-level context:** It captures lookup events that are interdependent across the length of the document.
3. **Learner meta-data:** It provides estimated CEFR levels for all learners and first language (L1) for ∼33% of learners.

Dictionary Look-up Prediction poses significant challenges for NLP models. Dictionary lookups are sparse events that vary widely based on context and individual knowledge, and are thus extremely difficult to predict. In response to these challenges, we argue that F_2 and ROC-AUC are appropriate evaluation metrics that reflect how useful a look-up prediction model would be for personalised learning applications.

Formulating Look-up Prediction as a sequence-to-sequence task, we evaluate fine-tuned LLaMA, Gemma, and Longformer models (Touvron et al., 2023; Gemma Team et al., 2024; Beltagy et al., 2020). During fine-tuning, in addition to a standard cross-entropy loss function, we also investigate the ROC* loss function that more directly targets the ROC-AUC (Yan et al., 2003). We conjecture that this is a more appropriate metric (see Section 5), and we find evidence that, in certain conditions, this ROC* function increases performance.

The goal of our research is to assess the suitability of different models for evaluating and aiding learners of English. Our contributions are as follows:

1. We release to the research community **two datasets**: a) DLU-dev, a dataset of >16,000 content word tokens and 630 clicks; and b) D-DLU, a set of 51 chatbot dialogues containing 43,000 content word tokens and 72 clicks.
2. We propose a new NLP task, **Dictionary Look-up Prediction**, and present a number of results for a sequence-to-sequence approach to this task.
3. We are the **first to apply ROC* loss to an NLP task**, and find that for certain cases it seems to outperform BCE loss.[3]
4. We argue that **fine-tuned LLMs are unable to satisfactorily capture contextual lexical complexity**. Not only do fine-tuned LLMs fail to outperform a feature-based ensemble model, but they also fail to generalise to other related tasks, such as Complex Word Identification (CWI) (see Section 2).

We release our data at https://englishlanguageitutoring.com/.

2 Related Work

The literature on word complexity includes contributions from not only NLP but also psycholinguistics and education (e.g. Bulté and Housen, 2012). This section focuses on data contributions.

CWI and LCP: Complex Word Identification (CWI) and Lexical Complexity Prediction (LCP) are two tasks in which the complexity of a word is predicted, either in the form of a binary label (CWI) or a continuous value (LCP). Both CWI and LCP have been explored in a series of shared tasks and other publications (Paetzold and Specia, 2016; Yimam et al., 2017; Shardlow, 2013; Shardlow et al., 2020, 2021; Gooding and Kochmar, 2018; Zaharia et al., 2022). Neural approaches are prevalent, but contemporary LLMs (such as GPT-4o, OpenAI et al., 2024) exhibit little to no benefit over smaller transformer models, such as $\text{RoBERTa}_{\text{large}}$ (see Smădu et al., 2024).

Existing CWI/LCP datasets have a number of shortcomings: They rely on annotators, which are often proficient L1 speakers of the language (but see Yimam et al., 2017, for the use of L2 speakers). Furthermore, the datasets typically operate on the word-in-sentence level; but a word that is difficult at the beginning of a document might be easy towards the end, when more context has been established. Thus, CWI and LCP datasets are unlikely to reflect the specific challenges L2 learners face when engaging in natural reading exercises.

Among the CWI and LCP datasets, the one for the 2018 shared task (Yimam et al., 2017, 2018) is especially interesting, as one of its sources was WikiNews (in addition to other news sources and Wikipedia), which forms also the basis for DLU. We use this dataset for further evaluation in Section 8.

Eye-Tracking Datasets: Eye-tracking is another way of approximating perceived word difficulty. Similar to the DLU dataset, eye-tracking datasets are often (but not always) created using reading tasks (Cop et al., 2017; Luke and Christianson, 2018; Hollenstein et al., 2018, 2020; Schmidtke et al., 2021; Berzak et al., 2022).

Compared to DLU, however, eye-tracking datasets are typically less naturalistic because the

[3] The performance difference is not statistically significant.

participants are not engaged in the task for the sake of learning, as is the case for our data, but are recruited for the experiment.

Additionally, eye-tracking datasets do not capture definition-seeking behaviour, but rather, a wide variety of cognitive processes. Definition-seeking is a behaviour aimed distinctly at lexical information, while the causes of eye movement are diverse and therefore harder to interpret.

Word Lists: Researchers and education specialists have created various word lists graded for difficulty (Negishi et al., 2013; Capel, 2015; Volodina et al., 2016; Flor et al., 2024). For example, the *English Vocabulary Profile* (Capel, 2015) provides CEFR levels for many definitions (CoE, 2020).

Compared to complexity-graded word lists, our data is specific to words in context.[4] Such contextualisation is useful when selecting e.g. reading tasks at an appropriate level for a learner.

Word lists cover only a limited vocabulary and struggle when new senses for a word are introduced. Furthermore, existing word lists describe a generic level of difficulty, and do not reflect e.g. how different L1s influence how challenging words in an L2 are. As our dataset provides L1 information when available, it helps address this gap.

Semantic Error Prediction: Since it is based on a reading and summarisation task, our dataset is specifically focused on complexity in comprehension. A comparable *production-side* dataset is the Semantic Error Prediction dataset by Strohmaier and Buttery (2024), which provides information on which content words learners fail to produce when writing essays. Like DLU, the SEP dataset is also based upon behavioural data from L2 learners engaged in a naturalistic learning task. That being said, we can expect differences to exist between production and comprehension, because in the case of production, learners have a (rough) meaning in their mind and have to retrieve correct word forms, while in the case of comprehension, the form is given by the text and learners have to access the correct meaning for it (cf. Jiang, 2000). We use the SEP dataset for evaluation in Section 8.

3 Description of the DLU Dataset

This section describes DLU's main features.

3.1 Data Source

We use the *Read&Improve* (R&I) platform as our data source. This platform allows L2 learners to engage in the task of reading and summarising an article to improve their English (see Figure 2 in Appendix A for a screenshot of the platform interface). Upon submission of the summary, they receive automated feedback. During reading, learners can click on words to retrieve definitions and examples.

The texts used for this task are taken from WikiNews (available under a Creative Commons license). Different users might be presented with the same article to summarise.

3.2 Scope of DLU

Our data shows which content words in a seen document have been clicked on to retrieve dictionary information. That is, for each content word token of a document, the data specifies whether it has been clicked on or not by the user. Tokens are considered content words if they have been tagged as adjectives, adverbs, nouns, or verbs by the RASP pipeline used by R&I (Briscoe et al., 2006).

To ensure that the learner has seen all tokens, only documents for which they have successfully submitted a summary are included. We also exclude data from users who have clicked less than five times in total, as this might indicate that they are unfamiliar with the functionality of clicking words to look up their dictionary information, or that they are so proficient as to never require dictionary information.

3.3 Data Selection

Overall, our dataset includes more than 260,000 seen content word tokens, with more than 8,800 clicks on these tokens. We split the DLU dataset into three parts by document: a train split, a dev split, and a test split, where the dev and test splits contain slightly more than 10% of unique documents. More information on the size of the dataset and its splits can be found in Table 1.

The dev-split (DLU-dev) is released with this paper, while the train and test splits are reserved for a future shared task open to public participation. All data will be released upon completion of the shared task.

3.4 User Information

Users likely differ in their look-up patterns both idiosyncratically and systematically based on their:

[4] For an application of word lists to contextualised uses, see Aleksandrova and Pouliot (2023).

split	tot. docs	uniq. docs	users	clicks	con. tokens
all	1327	221	663	8858	266011
train	1143	176	616	7822	235786
dev	101	21	90	630	16084
test	83	24	68	406	14141

Table 1: Description of data and splits, including the number of content tokens. Multiple users might see the same document, therefore the number of total documents can diverge from that of unique documents.

	A2	B1	B2	C1	C2	UNK	sum
all	135	198	126	34	1	169	663
train	123	185	117	33	1	157	616
dev	21	32	17	6	0	14	90
test	13	22	15	5	0	13	68

Table 2: Essay-based estimation of user CEFR levels.

- first language (L1)
- language ability as estimated CEFR level

For the wide range of L1s in DLU-dev, see Table 10 in the appendix. The language with most users is Spanish (93) followed by Italian and Turkish (both 17). For some languages (e.g. Serbian, Hindi), data is only available for a single user.

Our datasets include two estimates of the learner CEFR level. One estimate is based on submissions to the associated essay writing platform *Write&Improve* (W&I)[5] and described in Table 2, while the other relies on self-reports. While the self-reports have full coverage (see Table 6), the essay data are more comparable across users, as it is based on the same automatic grading system. We therefore only use essay-based estimates of CEFR-levels in our experiments, even though both are included in our data release.

While the automatic scores are likely imperfect, we believe that they provide a reasonable approximation of the learner proficiency because they correlate with look-up propensity (see Figure 1). With only one exception, learners at higher levels tend to look up a smaller proportion of word tokens.

3.5 Noise and Uncertainty

Look-up events are affected by many idiosyncratic features, not all of which are captured by our dataset. Notably, how often a learner has previously encountered a word will strongly affect whether they look it up.[6] As a result, our dataset leads to high aleatoric uncertainty for the models trained on it (Hüllermeier and Waegeman, 2021).

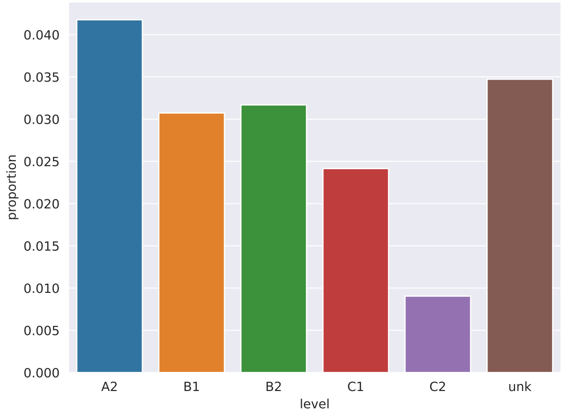


Figure 1: Proportion of content words that were looked up, for each CEFR-level as estimated using W&I essays.

For many applications, however, perfect prediction of look-ups are not required. Rather, the main goal is to separate words that are difficult enough to require a definition, from words that are easier. With this separation, text readability can be improved and vocabulary test items can be created.

4 Description of Chatbot Dataset

The chatbot dialogue dataset (D-DLU) is derived from an earlier experiment by Tyen et al. (2024) using BlenderBot v1 (Roller et al., 2021). It consists of two types of dialogue data:

1. A reading condition (D-read), where participants read self-chats between only the bot.
2. A chat condition (D-chat), in which participants chat with the bot.

We filtered this dataset manually to remove chats containing unsafe texts such as insults or inappropriate topics, and instances in which the chatbot behaved erratically, e.g. when the chatbot defined similar words repeatedly. As a result, we ended up with a set of 51 chats from the original 80.

A closer look at the data distribution (Table 13) reveals that the remaining 25 dialogues in D-chat only contained 5 look-up events. This number is too low for informative evaluation. While we release both portions of the dataset, we recommend only using the D-read split for lookup-prediction.

[5] `https://writeandimprove.com/`. For W&I essay data, see Nicholls et al. (2024).

[6] Research on language acquisition suggests that many encounters are needed for even passive knowledge of vocabulary. See Hargis et al. (1988) for some estimates of required repetitions. See Peters and Webb (2018) for how encounters during TV consumption lead to vocabulary acquisition.

5 Evaluation Metrics for DLU

Considering data sparsity, noise, and intended application areas, we argue that F_2 and ROC-AUC are most appropriate evaluation metrics for DLU.

5.1 F_2 and adaptive F_2 Metric

Unlike F_1, F_2 prioritises recall over precision:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision}\cdot\text{Recall}}{\beta^2\cdot\text{Precision}+\text{Recall}},\ \beta = 2$$

That is, we accept more false positives to avoid false negatives. In our case, false positives are predictions of look-up events where no such event has occurred. However, we expect that learners do not click on every difficult word, as browsing dictionary information disrupts the flow of reading.

Furthermore, for many use cases it is acceptable to wrongly predict that a few words are difficult. For example, if the goal is to select words for a cloze test, then adding a small portion of comparatively easy gap words to the test will not create a problem.

Look-up events (and therefore positive labels) are rare, which might render a decision threshold of 0.5 too strict. We are instead looking for a metric that is robust to the high sparsity of DLU and provides more general information about whether the models are able to separate words that are likely to be looked up from those that are unlikely to be looked up.

Therefore, we explore using an adaptive threshold for the F_2 metric. That is, we estimate which value in the inclusive range 0.01 – 0.99 (step size 0.01) leads to the highest F_2 score on the training data and then use the same threshold on the evaluation data to calculate an adaptive F_2 (aF_2).

5.2 ROC-AUC Metric

Compared to the F_2 metric, the ROC-AUC provides more direct information on whether the models distinguish easy from difficult words. The area under the curve provides this information, because it can be interpreted as the probability that the model ranks a randomly chosen positive example higher than a randomly chosen negative example (Fawcett, 2006, p. 868).

For many applications, using the raw scores rather than the binary classification is useful. Consider the case of creating a number of test items for the most difficult words in a text. In such a case, we are not committed to any particular threshold at which a score indicates that the token would be looked up in a dictionary. In contrast to the F-metrics, the AUC does not rely on any such threshold. Thus, we believe that the AUC metric is well suited for our purposes. In Section 6.2, we describe the ROC* loss function, which targets the AUC.

6 DLU Models

This section describes 3 types of DLU models.

6.1 Feature-based Baselines

Frequency Baseline First, we provide a baseline based on word frequencies using the `wordfreq` package (Speer, 2022).[7] We use the Zipf frequency estimate provided by `wordfreq`, which consists of a value within the 0 to 10 range.[8] We rescale this value to be between 0 and 1. All together, the score for a token is calculated as follows:

$$\text{score}(word) = 1 - \frac{\log_{10}(\text{proportion}(word))+9}{10}$$

Ensemble Baseline We also explore a more complicated feature-based ensemble model using scikit-learn (Pedregosa et al., 2011), consisting of six classifiers combined in a soft voting ensemble. The features of this model include the frequency baseline and a variety of features pertaining to the word lemma and the word token's position in the text. The exact features are listed in the appendix Appendix E. Due to reliance on the W&I-estimated CEFR-level of the learner, we do not report results of this baseline on D-read, which does not have this feature.

6.2 Finetuned Models

We fine-tune two models on the DLU train split, both from Huggingface transformers (Wolf et al. 2020; see also Appendix F):

1. Longformer (Beltagy et al., 2020), which we choose over other BERT-like models because we operate on the document-level;
2. LLaMA 3.2 (Touvron et al., 2023), for which we choose the 1B parameter version due to compute considerations.

Loss Functions We explore two loss functions. As is common for binary classification tasks, we use the Binary Cross Entropy (BCE) loss as the basis

[7] `https://github.com/rspeer/wordfreq`, which is based on the ExquisiteCorpus (`https://github.com/LuminosoInsight/exquisite-corpus`).

[8] While this calculation can reach 10, due to the distribution of words, the effective range tends to be between 0 and 8. The package also uses 0 as the default value of words not found in the word list, even though 0 does not correspond to zero occurrences due to the Zipfian transformation.

for our first loss function. To adjust for the label imbalance, we use a weight for positive cases (w_p). Thus our BCE loss takes the following form:

$$l_{\text{BCE}} = -\left(w_p y \log\left(\sigma(x)\right) + (1-y)\log\left(1-\sigma(x)\right)\right)$$

We treat the weight as a hyperparameter to be decided through search, but the search space is biased towards higher values as positive cases are under-represented (see Section F.1).

The second loss function, called ROC*, targets the ROC-AUC directly. This function was developed by Yan et al. (2003) and is based on the equivalence between the ROC and the Wilcoxon-Mann-Whitney statistic. We explore this loss-function because we take the correct ranking of words as measured by ROC-AUC to be an excellent metric reflecting probable use-cases (see Section 5.2).

Let **N** be the set of scores for non-clicked content word tokens and **C** the set of scores for clicked content word tokens. As the loss function compares between pairs of these two sets, it is useful to introduce their product: $\mathbf{P} = \mathbf{N} \times \mathbf{C}$

The loss takes the following form (batching is ignored here for illustration):

$$l_{\text{ROC*}} = \frac{1}{|\mathbf{P}|}\sum_{(x,y)\in\mathbf{P}} \begin{cases} (x-y+\gamma)^2 & : x+\gamma > y \\ 0 & : \text{otherwise} \end{cases}$$

where $\gamma > 0$ is a hyperparameter ensuring that a sufficiently large distance exists between clicked and non-clicked cases.[9] Thus, we allow for mini-batch training by storing previously seen scores for content word tokens and sampling them for comparison against scores calculated in the mini-batch. The size of the samples is treated as a hyperparameter.

Hyperparameter Search We perform a 20-trial hyperparameter search using Optuna (Akiba et al., 2019) maximizing ROC-AUC, training on the train split and evaluating on the dev split. The selected hyperparameters are in Section F.1.

Data Processing To account for L1 and CEFR level, we add special tokens for them to the model and append them at the start of each document. While the model will see the same document multiple times with different look-up patterns during training, these will often differ in either indicated L1 or CEFR level. For adding the special token, we merge CEFR levels such that B1 and B2 are represented as B, and so forth (see Table 7). This addresses the problem of having relatively few cases for some CEFR levels. Under-represented L1s are merged into the "unknown" category.

We evaluate also on the chatbot dataset split *D-read*. However, the length of 7 chats in particular pose a problem as the Longformer model we use is limited to 4096 subtokens. To circumvent this, we split longer dialogues after reaching this threshold, which might affect performance on D-read.

Significance Tests We perform permutation significance tests to see if ROC* trained models achieve higher AUC compared to BCE trained models. With 0.05 as the starting p-value, the Bonferroni-corrected threshold for this paper is 0.0027. To avoid further lowering of the threshold, we only perform tests for the aggregated DLU test split.

6.3 Prompting Models

For comparison, we also prompted LLMs on the dictionary look-up task, specifically the instruction-tuned versions of Gemma and LLaMA (Gemma Team et al., 2024; Touvron et al., 2023) (Appendix F). We use both zeroshot and fewshot prompting, as described in Appendix G, except for D-read where we only use zeroshot prompting due to the challenging document length.

Our prompts return complex words from the text. To address cases in which word types occur more than once in the text, we explored two approaches: 1) Predicting only the first occurrence to be looked-up and 2) predicting look-ups for all occurrences. We focus on the first option, as learners usually only need to look up a word once, and we report the results of the second approach in the appendix. The overall picture is not affected by this choice.

As our prompting models output binary results, there is no changeable threshold for an adaptive F_2. Similarly, the AUC is less useful as the score for each word token is 0 or 1.

7 DLU Results

The results of the baselines as well as the transformer models on the DLU test split can be found in Table 3. We report the results for the coarse-grained CEFR-levels (A, B, C, and unknown) separately. Because we release only the dev split of DLU with this paper, we also report the results on this split in the appendix (see Table 22).

Due to the sparsity of look-up events and because the hyperparameter search targets the AUC, some model settings lead to F_1 and F_2 values of 0. The aF_2 consistently takes a value above 0, but

[9] Our implementation follows the public ROC* repository (`https://github.com/iridiumblue/ROC*`). However, similar to Yan et al. (2003), we keep γ as a hyperparameter, instead of deriving it.

		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	10.2	14.3	-	55.9	14.2	17.2	-	57.1	**15.0**	**22.3**	-	64.4	11.1	16.1	-	58.3	12.8	16.8	-	57.5	2.2	4.6	-	57.4
	fewshot	9.7	14.3	-	55.9	13.4	17.2	-	57.1	13.8	22.3	-	65.8	12.3	18.2	-	60.0	12.4	17.1	-	57.8	-	-	-	-
LLaMA-Inst.	zeroshot	10.1	17.8	-	59.6	8.3	12.4	-	53.4	6.5	11.6	-	57.8	9.5	16.5	-	60.1	8.8	14.2	-	55.8	1.0	2.4	-	56.1
	fewshot	10.3	16.7	-	58.0	9.0	12.5	-	53.8	6.5	11.7	-	57.9	6.0	10.0	-	53.6	8.5	12.9	-	54.8	-	-	-	-
LLaMA	ROC*	0.0	0.0	7.2	75.7	0.0	0.0	8.2	65.9	0.0	0.0	0.0	62.1	0.0	0.0	11.2	68.1	0.0	0.0	7.9	67.8	0.0	0.0	3.0	76.7
	BCE	11.0	22.2	21.9	72.7	12.4	21.8	19.4	64.9	5.6	11.8	5.8	63.0	8.0	16.7	19.7	69.4	10.4	20.0	18.9	66.2	2.7	5.6	**5.2**	77.1
Longformer	ROC*	14.4	25.6	18.6	78.5	16.0	25.5	21.7	72.4	5.7	11.0	13.4	64.9	12.1	20.7	18.7	77.6	14.0	23.6	20.0	73.9	**2.7**	**5.8**	2.8	83.4
	BCE	0.0	0.0	12.3	70.9	0.0	0.0	13.2	70.7	0.0	0.0	3.4	60.5	0.0	0.0	8.4	76.6	0.0	0.0	11.6	71.1	0.0	0.0	3.9	74.6
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.	**22.2**	**32.8**	**31.6**	**85.9**	**17.3**	**26.1**	**28.0**	**76.2**	12.3	18.8	**19.6**	**81.9**	**14.2**	**24.3**	**23.6**	**80.4**	**17.3**	**26.5**	**27.4**	**79.2**	-	-	-	-

Table 3: Results on the DLU test split. "aF_2" stands for F_2 with a adaptive threshold, as discussed in Section 5.

because the threshold is estimated only on the training data and only for the entire dataset (i.e. not separately for each CEFR level), the aF_2 is sometimes lower than the F_2. The impact of the adaptive threshold is discussed in Appendix J and does not bear on the general conclusions.

The best finetuned models outperformed the prompt-based models with the exception of the C split of data. This split contains data from learners at the C1 and C2 CEFR-levels, who rarely look up words (24 in the test split, see Table 9).

The results show that the frequency baseline is strong, often outperforming other models. The ensemble baseline is even stronger, outperforming all other models convincingly with only minor exception on the C split. Some of the differences are substantial, e.g. the ensemble baseline achieves an 79.2% AUC on the overall DLU test split, with the next best model reaching only 73.9%.

Among the transformer models, the Longformer ROC* model performs best on the test split. We note, however, that these results do not directly translate to the dev split of DLU (see Table 22), suggesting some overfitting. No difference in AUC scores between ROC* and BCE model is statistically significant, although the ROC* versions consistently perform better.

The results on the D-read split in D-DLU described in Section 4 are also included in Table 3. As is to be expected for a different data source with a different distribution, performance is lower. The highest F_2 (5.8%) and aF_2 (5.2%) are achieved by the ROC*-Longformer and BCE-LLaMA model respectively. The frequency baseline achieves the highest AUC (84.9%), follow by the ROC*-Longformer (83.4%).

As described in Section 3, multiple users might interact with the same document, leading to different look-up events. To account for any effects this might have, we also evaluated on a filtered version of our dataset so that each document was unique. See Appendix K, Table 25, and Table 24 for the results, which show the same overall picture.

8 Evaluation on CWI/SEP

To investigate the degree to which our DLU dataset captures word difficulty information that is specific to the dataset's construction, including the underlying reading task, we ask the following question: Can models transfer DLU knowledge to other tasks that also attempt to track word complexity? We address this question by performing additional experiments using the CWI and SEP datasets.

We chose the CWI task because it is the most widely explored binary task targeting word complexity. However, CWI datasets are usually not as sparse, often do not provide information on the document-level, and frequently rely on annotators proficient in the language, rather than learners.

We also evaluate on the SEP dataset because it not only targets word complexity, but provides highly sparse binary learner behaviour data for longer contexts[10]; as is the case for DLU prediction. Furthermore, investigating the transfer to SEP addresses the question of whether learners struggle to produce the words that they find difficult enough to look up during a reading task. We can thus provide evidence for how similar comprehension difficulty and production difficulty are.

8.1 Experimental Setup

We train all DLU-finetuned models an additional time on the CWI and SEP datasets by Yimam et al. (2017) and Strohmaier and Buttery (2024). For comparison, we finetune the base models on the

[10] The SEP dataset is standardly formatted to chunks of one or more paragraphs.

CWI and the SEP task without using DLU data, and provide the frequency baseline. The experimental procedure follows the same pattern as described in Section 6, i.e. an initial hyperparameter search followed by evaluation on the dataset.

Significance Tests We use permutation significance tests to see if the models finetuned on both DLU and CWI or SEP perform better than models only finetuned on the latter. We perform these tests for the F_1 and AUC metrics because the F_1 was used in previous work and the AUC was targeted by the hyperparameter search.[11]

8.2 CWI/SEP Results

We present the CWI results split by data source (News, Wikipedia, and WikiNews) and in aggregate. For the SEP dataset, we offer the same split by CEFR level as for DLU.

The CWI results (see Table 4) suggest that the BCE-Longformer architecture is best suited for this task when considering F-scores and AUC. The DLU-finetuned version of the BCE-Longformer model produces the highest AUC (85.6%) and the base model the highest F_1 (78.5%), but only the comparison of the F_1 is statistically significant. The only other result significant at the 0.0027 threshold was the difference between the F_1 of the base ROC*-LLaMA (74.3%) and the DLU-finetuned model (71.1%), which favours the base model.

The SEP results (reported in Table 5) clearly suggest a strongest model on the F-score metrics: the BCE-Longformer model finetuned only on SEP. Among the 6 significant results (see Table 20), only the difference between the F_1-scores of the DLU-finetuned BCE-LLaMA (7.3%) and the CWI-only version (4.2%) points towards positive transfer, the rest pointing in the opposite direction. The overall best F_1 (11.9%) and AUC (71.0%) slightly outperform the numbers (11% and 69.8%) previously reported (Strohmaier and Buttery, 2024).

9 Discussion

Our results show that **look-up prediction is a challenging, but addressable task**. Finetuned transformer models outperform a frequency baseline, but fail to beat a feature-based ensemble model.

Similar to the CWI results reported by Smădu et al. (2024), we find that **model size is not the decisive variable**: among the finetuned models, the ROC*-Longformer model outperformed the LLaMA models, even though the latter has considerably more parameters (∼149M vs 1B). Similarly, the prompting models were considerably larger than the finetuned models and yet performed worse.

The ROC* loss which we explored following Yan et al. (2003) performed well on DLU for its target metric, the ROC-AUC. Among the finetuned neural models, the highest AUC is always produced by a model using ROC* loss. Thus, we suggest that **the ROC* loss function is of value for tasks in which the AUC is the correct metric**. That being said, neither on the DLU-dev split nor the aggregated CWI data is the highest AUC achieved by a ROC* model. The improvements, thus, appear dependent on the specific data distribution.

The simple frequency baseline proved competitive on all considered tasks. Even more impressive was the performance of the feature-based ensemble model on DLU, which showed a leading performance. Even compared to LLMs, **feature-based baselines remain very competitive in the field of word complexity**. On the combined data of DLU, the highest scores on all four considered metrics were achieved by the ensemble baseline. In the case of the English Wikipedia split of the CWI data (Yimam et al., 2017), the F_1 of the simple frequency baseline (73.9%) outperformed every one of the 14 few-shot prompting models reported by Smădu et al. (2024), where the best one only achieved 70.6%.

These strong baseline results and irrelevance of model size suggests to us that **modelling difficulty in L2 vocabulary acquisition is not solved by existing NLP methods**. We believe that further personalisation is required to move forward, and we see DLU as a major step in this direction.

Furthermore, models will have to be more specifically adapted to the high variance between learners. Providing information about proficiency level and first language as special tokens proved insufficient. To account for the variance between learners, it might be necessary to adapt the training procedure or architecture details of the model even further to information about the learner. With more personalised DLU data, it might, for example, be possible to train layers specialised for certain L1s, CEFR levels, or other background data.

Looking at the CWI and SEP experiments, the significant results do not support that knowledge from finetuning on DLU is transferred to other

[11] We do not run tests for other metrics as this would increase the number of significance tests, decreasing the Bonferroni-corrected threshold further.

		N					W					WN					All				
		Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC
f.-LLaMA	ROC*	76.7	73.7	72.4	81.7	84.2	68.8	69.1	64.2	75.0	78.7	72.0	68.2	65.9	76.7	79.4	73.7	71.1	68.6	78.7	81.7
	BCE	73.6	75.4	83.1	82.9	82.3	69.7	75.4	78.9	76.5	73.2	68.8	70.5	75.7	74.6	74.9	71.4	74.0	80.0	79.0	78.2
f.-Longformer	ROC*	**79.7**	78.7	80.8	86.6	87.7	72.4	74.9	73.4	84.6	80.7	76.4	74.9	75.3	82.0	84.0	77.2	76.8	77.5	84.8	85.3
	BCE	72.3	76.2	**87.6**	**88.2**	**87.9**	68.8	77.8	**87.7**	**87.2**	81.0	67.3	73.0	**84.5**	**84.1**	84.3	70.1	75.6	**86.7**	**86.8**	**85.6**
LLaMA	ROC*	77.4	77.0	80.4	81.9	84.2	69.4	71.8	69.9	73.3	76.5	72.4	71.7	73.6	75.0	78.3	74.3	74.3	76.0	77.9	81.0
	BCE	73.9	76.3	85.0	78.6	83.5	68.9	76.3	82.9	70.1	74.3	66.9	70.9	79.3	68.4	76.9	70.8	74.7	82.9	73.7	79.9
Longformer	ROC*	79.4	79.1	82.8	85.9	87.3	73.2	75.8	74.6	84.2	**81.5**	**77.2**	76.2	77.4	83.1	**84.4**	**77.4**	77.5	79.3	84.7	85.3
	BCE	77.8	**79.1**	86.7	85.4	87.6	**74.6**	**79.4**	83.3	78.0	81.1	74.9	**76.6**	83.0	80.3	83.6	76.3	**78.5**	84.8	82.2	85.1
Baseline	freq.	62.2	65.6	73.0	80.7	67.5	66.1	73.9	80.0	86.8	67.3	61.7	67.8	77.7	81.4	66.9	62.9	68.2	76.0	82.3	67.6

Table 4: Prediction results on the 2018 CWI dataset (Yimam et al., 2017). Models with the prefix "f.-" for "finetuned" have first been finetuned on DLU. "aF_2" stands for F_2 with a adaptive threshold, see Section 5.

		A					B					C					N					All				
		Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC	Acc.	F_1	F_2	aF_2	AUC
f.-LLaMA	ROC*	96.3	0.7	0.4	15.3	58.1	97.1	0.0	0.0	15.6	63.9	98.2	1.2	0.8	9.2	59.1	99.4	0.0	0.0	2.9	53.7	97.8	0.5	0.3	11.3	59.4
	BCE	88.7	9.2	12.5	13.3	62.2	90.8	10.5	14.6	16.0	65.7	91.9	5.7	9.2	8.7	63.8	93.6	1.2	2.5	2.0	53.4	91.4	7.3	11.1	11.5	63.8
f.-Longformer	ROC*	**96.5**	0.0	0.0	7.4	**71.1**	**97.2**	0.0	0.0	8.9	**73.4**	**98.4**	0.0	0.0	7.7	**69.7**	**99.4**	0.0	0.0	0.0	58.5	**98.0**	0.0	0.0	7.4	**71.0**
	BCE	**96.5**	0.0	0.0	15.2	51.1	**97.2**	0.0	0.0	12.4	54.2	**98.4**	0.0	0.0	7.5	53.0	**99.4**	0.0	0.0	2.7	54.9	**98.0**	0.0	0.0	9.4	52.0
LLaMA	ROC*	92.0	8.4	9.6	9.7	59.8	93.4	10.4	12.3	12.1	65.7	95.1	5.6	7.3	6.8	63.1	96.5	1.1	1.9	2.0	53.4	94.4	7.4	9.2	9.1	63.4
	BCE	95.3	3.2	2.5	1.8	63.3	96.3	7.4	6.0	3.4	66.2	97.7	1.9	1.6	0.8	63.0	98.9	0.0	0.0	0.0	52.6	97.1	4.2	3.4	2.1	64.7
Longformer	ROC*	93.3	11.6	12.3	2.2	67.8	95.2	**16.3**	16.6	7.2	72.0	96.4	**11.0**	**12.6**	2.5	67.8	97.7	**2.6**	**3.7**	0.0	**61.0**	95.8	11.8	13.0	3.9	70.1
	BCE	90.0	**13.9**	**18.4**	**19.8**	67.8	92.8	14.2	**17.9**	**19.6**	71.1	95.0	9.4	12.5	**14.9**	68.4	97.6	0.8	1.2	1.7	55.6	94.1	**11.9**	**15.6**	**17.0**	70.5
Baseline	freq.	61.3	7.6	15.2	14.4	54.1	59.6	6.5	13.6	12.9	56.3	55.6	3.7	8.3	8.5	54.5	50.4	1.2	2.9	**3.2**	54.7	56.5	4.4	9.7	9.5	53.3

Table 5: Results on the Semantic Error Prediction (SEP) dataset (Strohmaier and Buttery, 2024). Models with the prefix "f.-" have first been finetuned on DLU. "aF_2" stands for F_2 with a adaptive threshold, see Section 5.

tasks. Based on this observation, we conjecture that **different approaches that all ostensibly concern word complexity, in fact track different phenomena**. In particular, for the CWI task the significant results point in the direction of negative transfer between DLU and CWI. We take this to show that the data distribution diverge too strongly.

Part of the distributional differences are, without doubt, the sparse nature of DLU and the shorter length of the CWI texts. Another difference, however, is that the CWI data we used was derived from *proficient* speakers of English rather than *learners*. Because DLU directly records dictionary usage during a naturalistic learning task, it has higher external validity. Hence, we speculate that the CWI data do not sufficiently reflect which words L2 learners of English struggle with.

In contrast to CWI, we found at least one significant result on the aggregated SEP dataset pointing in the direction of improvement, although with five other significant result pointing in the opposite direction. Like DLU, SEP is derived from learner behaviour in a naturalistic task. However, DLU targets difficulties in comprehending words, while SEP targets production difficulties, which may explain the differences.

10 Conclusion

We introduce the dictionary look-up task, which provides insight into word complexity for the purpose of supporting personalised learning technologies. We release the *Dictionary Look-Up development* (DLU-dev) dataset. Additionally, we release a look-up dataset of chatbot dialogues (D-DLU) for evaluation. We provide results from zero- and few-shot prompting as well as fine-tuning.

Investigating the transfer from DLU to other tasks such as complex word identification (CWI) and semantic error prediction (SEP), we find that DLU and CWI appear to track different phenomena. For SEP, we set new state-of-the-art results, but find mixed to negative results on transfer.

The leading performance of a feature-based model on DLU strongly suggests that further research is required to adequately incorporate information about individual learners and their lexical acquisition into neural models of word complexity. The release of DLU-dev is a major step toward achieving this goal. We release our data at `https://englishlanguageitutoring.com/`.

Acknowledgments

We thank Andrew Caines, Øistein Andersen, Paul Ricketts, Jon Phillips, Scott Thomas, and other members of ALTA and CUP&A for their support. Furthermore, we thank the anonymous reviewers for their comments.

This research is supported by Cambridge University Press & Assessment.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.

Desislava Aleksandrova and Vincent Pouliot. 2023. Cefr-based contextual lexical complexity classifier in english and french. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, page 518–527, Toronto, Canada. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. (arXiv:2004.05150). ArXiv:2004.05150 [cs].

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, 6:41–50.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, page 77–80, Sydney, Australia. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Bram Bulté and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, Language Learning & Language Teaching, pages 21–46. John Benjamins Publishing Company.

Annette Capel. 2015. The english vocabulary profile. *English profile in practice*, 5(1):9–27.

Council of Europe CoE. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Companion Volume*. Council of Europe Publishing, Strasbourg.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Michael Flor, Steven Holtzman, Paul Deane, and Isaac Bejar. 2024. Mapping of american english vocabulary by grade levels. *ITL - International Journal of Applied Linguistics*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas

Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, page 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Charles H. Hargis, Marge Terhaar-Yonkers, Patricia Couch Williams, and Mellissa Testerman Reed. 1988. Repetition Requirements for Word Recognition. *Journal of Reading*, 31(4):320–327.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5(11):180291.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, page 138–146, Marseille, France. European Language Resources Association.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Nan Jiang. 2000. Lexical representation and development in a second language. *Applied Linguistics*, 21(1):47–77.

Steven G. Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Masashi Negishi, Tomoko Takada, and Yukio Tono. 2013. A progress report on the development of the cefr-j. In *Exploring language frameworks: Proceedings of the ALTE kraków conference*, page 135–163. Citation Key: negishi2013progress.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. *The Write & Improve Corpus 2024: Error-Annotated and CEFR-Labelled Essays by Learners of English*.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai

Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Elke Peters and Stuart Webb. 2018. Incidental Vocabulary Acquisition Through Viewing L2 Television and Factors That Affect Learning. *Studies in Second Language Acquisition*, 40(3):551–577.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Daniel Schmidtke, Julie Van Dyke, and Victor Kuperman. 2021. Complex: An eye-movement database of compound word reading in english. *Behavior Research Methods*, 53:59–77.

Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

David Strohmaier and Paula Buttery. 2024. Semantic error prediction: Estimating word production complexity. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, 13:209–225.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Gladys Tyen, Andrew Caines, and Paula Buttery. 2024. Llm Chatbots as a Language Practice Tool: A User Study. In *Swedish Language Technology Conference and NLP4CALL*, pages 235–247.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016. Swellex: Second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, page 76–84, Umeå, Sweden. LiU Electronic Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 38–45, Online. Association for Computational Linguistics.

Lian Yan, Robert Dodier, Michael C. Mozer, and Richard Wolniewicz. 2003. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 848–855, Washington, DC, USA. AAAI Press.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

George-Eduard Zaharia, Răzvan-Alexandru Smădu, Dumitru Cercel, and Mihai Dascalu. 2022. Domain adaptation in multilingual and multi-domain monolingual settings for complex word identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 70–80, Dublin, Ireland. Association for Computational Linguistics.

A Platform

The *Read&Improve* (R&I) platform and its sister platform *Write&Improve* are available free of charge. Users agree to the sharing of their input for research purposes. For a screenshot of the R&I user interface, see Figure 2.

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 (Tyen et al., 2024). 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 W&I.

	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 W&I across documents by users (i.e. some users and WikiNews articles appear more than once in this table).

D.1 Format of the Data

The data is formatted as a document-level token-classification task. Tokenisation follows the RASP pipeline used by R&I (Briscoe et al., 2006) For each token a label is provided, with the default label -100 used for non-content word tokens.

Example

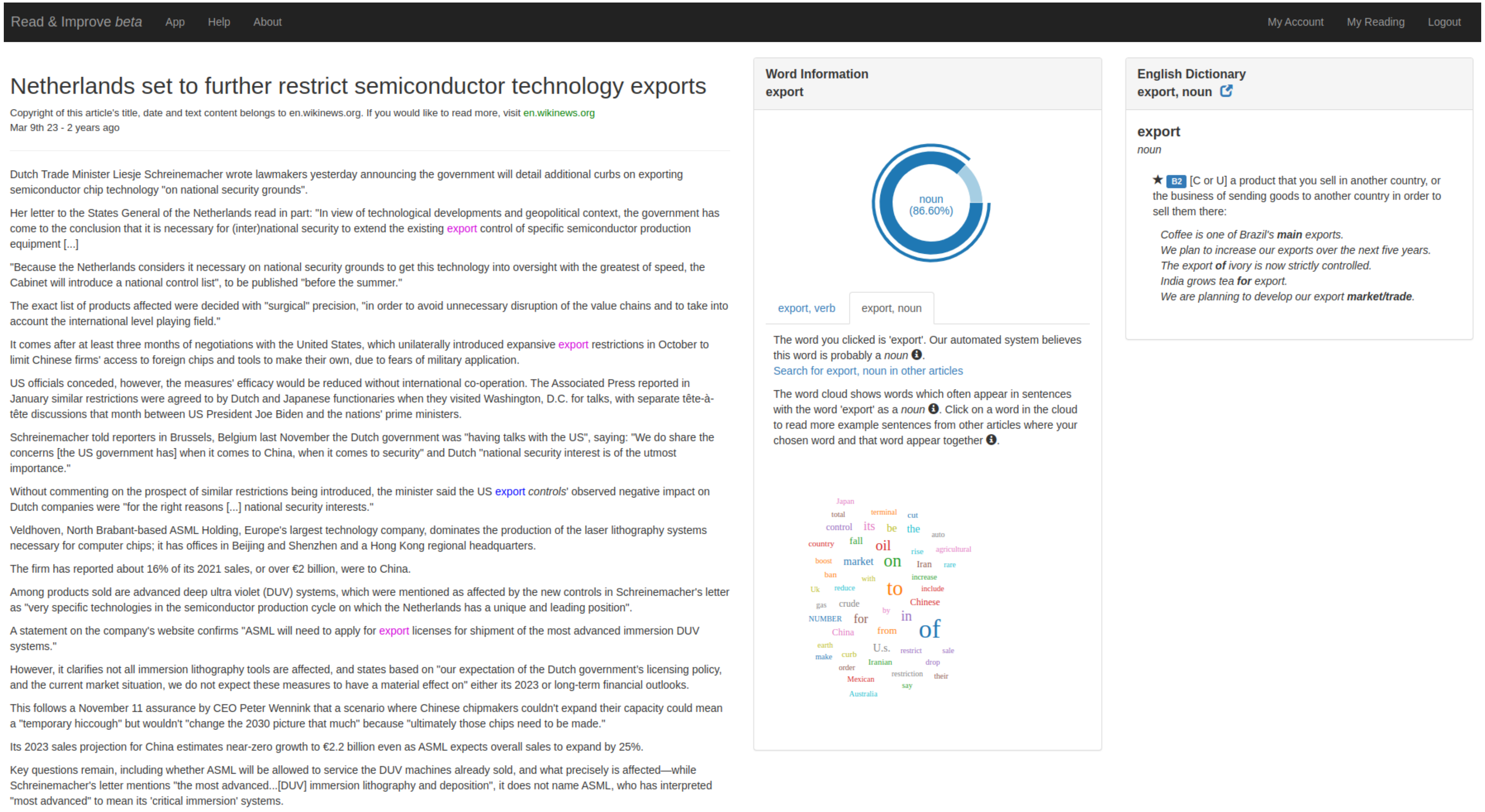
Text	Taco	Bell	restaurants	decided	Wednesday	to	remove	…
Labels	0	0	0	0	0	-100	1	…

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`



Netherlands set to further restrict semiconductor technology exports

Copyright of this article's title, date and text content belongs to en.wikinews.org. If you would like to read more, visit en.wikinews.org
Mar 9th 23 - 2 years ago

Dutch Trade Minister Liesje Schreinemacher wrote lawmakers yesterday announcing the government will detail additional curbs on exporting semiconductor chip technology "on national security grounds".

Her letter to the States General of the Netherlands read in part: "In view of technological developments and geopolitical context, the government has come to the conclusion that it is necessary for (inter)national security to extend the existing export control of specific semiconductor production equipment [...]

"Because the Netherlands considers it necessary on national security grounds to get this technology into oversight with the greatest of speed, the Cabinet will introduce a national control list", to be published "before the summer."

The exact list of products affected were decided with "surgical" precision, "in order to avoid unnecessary disruption of the value chains and to take into account the international level playing field."

It comes after at least three months of negotiations with the United States, which unilaterally introduced expansive export restrictions in October to limit Chinese firms' access to foreign chips and tools to make their own, due to fears of military application.

US officials conceded, however, the measures' efficacy would be reduced without international co-operation. The Associated Press reported in January similar restrictions were agreed to by Dutch and Japanese functionaries when they visited Washington, D.C. for talks, with separate tête-à-tête discussions that month between US President Joe Biden and the nations' prime ministers.

Schreinemacher told reporters in Brussels, Belgium last November the Dutch government was "having talks with the US", saying: "We do share the concerns [the US government has] when it comes to China, when it comes to security" and Dutch "national security interest is of the utmost importance."

Without commenting on the prospect of similar restrictions being introduced, the minister said the US export *controls*' observed negative impact on Dutch companies were "for the right reasons [...] national security interests."

Veldhoven, North Brabant-based ASML Holding, Europe's largest technology company, dominates the production of the laser lithography systems necessary for computer chips; it has offices in Beijing and Shenzhen and a Hong Kong regional headquarters.

The firm has reported about 16% of its 2021 sales, or over €2 billion, were to China.

Among products sold are advanced deep ultra violet (DUV) systems, which were mentioned as affected by the new controls in Schreinemacher's letter as "very specific technologies in the semiconductor production cycle on which the Netherlands has a unique and leading position".

A statement on the company's website confirms "ASML will need to apply for export licenses for shipment of the most advanced immersion DUV systems."

However, it clarifies not all immersion lithography tools are affected, and states based on "our expectation of the Dutch government's licensing policy, and the current market situation, we do not expect these measures to have a material effect on" either its 2023 or long-term financial outlooks.

This follows a November 11 assurance by CEO Peter Wennink that a scenario where Chinese chipmakers couldn't expand their capacity could mean a "temporary hiccough" but wouldn't "change the 2030 picture that much" because "ultimately those chips need to be made."

Its 2023 sales projection for China estimates near-zero growth to €2.2 billion even as ASML expects overall sales to expand by 25%.

Key questions remain, including whether ASML will be allowed to service the DUV machines already sold, and what precisely is affected—while Schreinemacher's letter mentions "the most advanced...[DUV] immersion lithography and deposition", it does not name ASML, who has interpreted "most advanced" to mean its 'critical immersion' systems.

Word Information
export

The word you clicked is 'export'. Our automated system believes this word is probably a *noun* ℹ.
Search for export, noun in other articles

The word cloud shows words which often appear in sentences with the word 'export' as a *noun* ℹ. Click on a word in the cloud to read more example sentences from other articles where your chosen word and that word appear together ℹ.

English Dictionary
export, noun

export
noun

★ B2 [C or U] a product that you sell in another country, or the business of sending goods to another country in order to sell them there:

*Coffee is one of Brazil's **main** exports.*
We plan to increase our exports over the next five years.
*The export **of** ivory is now strictly controlled.*
*India grows tea **for** export.*
*We are planning to develop our export **market/trade**.*

Figure 2: Screenshot of *Read&Improve* 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 W&I.

	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).

They were combined using the sklearn `VotingClassifier` class, which was set to soft voting. No systematic hyperparameter tuning was required, instead we compared a variety of settings and combinations manually on the dev-set (to avoid over-fitting) and then applied the best to the test set.

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, calculated from the training split.
- Length of word in characters.
- CEFR-level as estimated by essays submitted by the user.
- Count of definitions for the word in the *Cambridge Advanced Learner's Dictionary*.
- 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 data longer than sentences, but while DLU consists of WikiNews texts, the SEP consists of student essays split into chunks of one or more paragraphs. 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 instruct the LLM to consider a paragraph of text and the learner's English CEFR level. The models are asked to predict which words the learner is likely unfamiliar with, and return 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.
  –
  """ }
```

	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]		27	10	26	25
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	3.6×10^{-6}	5.4×10^{-5}	7.2×10^{-5}	5.2×10^{-5}
Pos. Weight	[0.8, 30]	BCE only	-	23.78	-	15.46
γ	[0.05, 0.75]	ROC* only	0.66	-	0.23	-
Sample size	[300, 10000]	ROC*, step size=100	3300	-	3800	-
Batch size (p.D)	[8, 80]	step size = 2	8	42	16	30

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]		17	2	8	3
Learning Rate	$[10^{-9}, 10^{-2}]$	log space	9.0×10^{-5}	1.8×10^{-4}	3.1×10^{-6}	2.9×10^{-4}
Pos. Weight	[0.8, 30]	BCE only	-	17.99	-	12.33
γ	[0.05, 0.75]	ROC* only	0.05	-	0.55	-
Sample size	[300, 10000]	ROC*, step size=100	300	-	4200	-
Batch size (p.D)	[4, 44]	step size = 2	16	30	40	20

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.

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	compare	2.8×10^{-2}	2.6×10^{-1}
LLaMA	AUC	compare	1.5×10^{-2}	5.2×10^{-1}

Table 18: Significance tests for DLU. The tests concern whether using the ROC* vs. the BEC loss changes the 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 CWI dataset for comparability. This might have affected our performance negatively, explaining some of the difference to the results reported by

	Metric	Loss	Statistic	p-Value
Longformer	AUC	roc	7.8×10^{-5}	9.6×10^{-1}
Longformer	F1	roc	7.5×10^{-3}	6.0×10^{-2}
Longformer	AUC	bce	4.9×10^{-3}	1.8×10^{-1}
Longformer	F1	bce	2.8×10^{-2}	2.0×10^{-4}
LLaMA	AUC	roc	7.2×10^{-3}	1.5×10^{-1}
LLaMA	F1	roc	3.2×10^{-2}	2.0×10^{-4}
LLaMA	AUC	bce	1.6×10^{-2}	8.0×10^{-3}
LLaMA	F1	bce	6.4×10^{-3}	2.1×10^{-1}

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

	Metric	Loss	Statistic	p-Value
Longformer	AUC	roc	9.2×10^{-3}	2.7×10^{-1}
Longformer	F1	roc	1.1×10^{-1}	2.0×10^{-4}
Longformer	AUC	bce	1.8×10^{-1}	2.0×10^{-4}
Longformer	F1	bce	1.1×10^{-1}	2.0×10^{-4}
LLaMA	AUC	roc	3.9×10^{-2}	4.0×10^{-4}
LLaMA	F1	roc	6.8×10^{-2}	2.0×10^{-4}
LLaMA	AUC	bce	8.1×10^{-3}	6.3×10^{-1}
LLaMA	F1	bce	3.1×10^{-2}	2.2×10^{-3}

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

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	7.2	75.7	0.0	0.0	8.2	65.9	0.0	0.0	0.0	62.1	0.0	0.0	11.2	68.1	0.0	0.0	7.9	67.8	0.0	0.0	3.0	76.7
	BCE	11.0	22.2	21.9	72.7	12.4	21.8	19.4	64.9	5.6	11.8	5.8	63.0	8.0	16.7	19.7	69.4	10.4	20.0	18.9	66.2	2.7	5.6	**5.2**	77.1
Longformer	ROC*	14.4	25.6	18.6	78.5	16.0	25.5	21.7	72.4	5.7	11.0	13.4	64.9	12.1	20.7	18.7	77.6	14.0	23.6	20.0	73.9	**2.7**	**5.8**	2.8	83.4
	BCE	0.0	0.0	12.3	70.9	0.0	0.0	13.2	70.7	0.0	0.0	3.4	60.5	0.0	0.0	8.4	76.6	0.0	0.0	11.6	71.1	0.0	0.0	3.9	74.6
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.	**22.2**	**32.8**	**31.6**	**85.9**	**17.3**	**26.1**	**28.0**	**76.2**	12.3	18.8	**19.6**	**81.9**	**14.2**	**24.3**	**23.6**	**80.4**	**17.3**	**26.5**	**27.4**	**79.2**	-	-	-	-

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	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
	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
Longformer	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
	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
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.	14.6	23.3	**24.0**	69.0	**11.3**	**20.1**	**19.3**	69.3	22.2	23.9	31.0	**64.9**	**21.6**	**32.9**	**31.3**	**76.7**	**15.0**	**23.8**	**24.2**	**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	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
	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
Longformer	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
	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
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.	14.6	23.3	**24.0**	69.0	**11.3**	**20.1**	**19.3**	69.3	22.2	23.9	31.0	**64.9**	**21.6**	**32.9**	**31.3**	**76.7**	**15.0**	**23.8**	**24.2**	**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	16.1	19.8	-	57.7	14.5	20.6	-	59.7	0.0	0.0	-	46.8	5.3	9.0	-	53.8	11.4	16.6	-	57.7
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	9.7	15.7	-	53.1	8.5	12.9	-	53.9	3.3	6.8	-	57.0	0.0	0.0	-	42.5	6.1	10.2	-	52.3
LLaMA	ROC*	4.9	3.8	3.8	72.2	21.7	17.9	17.9	70.2	0.0	0.0	0.0	**76.2**	**13.3**	15.2	15.2	58.9	14.9	13.1	13.1	69.5
	BCE	17.1	24.2	24.2	69.5	17.6	24.2	24.2	67.0	0.0	0.0	0.0	58.8	11.9	**22.0**	**22.0**	**74.2**	14.4	22.1	22.1	68.4
Longformer	ROC*	**22.5**	**26.5**	**26.5**	**78.8**	**22.0**	**29.2**	**29.2**	**75.4**	**5.3**	**10.0**	**10.0**	62.8	7.2	10.9	10.9	73.2	**17.3**	**23.7**	**23.7**	**75.5**
	BCE	16.1	16.4	16.4	73.0	12.6	12.5	12.5	72.2	0.0	0.0	0.0	72.7	7.0	9.0	9.0	73.6	11.1	12.2	12.2	72.7

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	6.7	10.4	-	52.9	7.5	10.9	-	54.5	19.2	17.9	-	55.9	0.0	0.0	-	47.7	9.2	12.3	-	54.3
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	8.3	15.8	-	58.2	3.7	6.8	-	49.6	22.7	**35.3**	-	**63.1**	0.0	0.0	-	41.3	9.3	16.8	-	57.9
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.