

L1 Influence in L2 Language Models: A Human-centric Approach

Laura Barbenel^{1,2*} Lily Goulder^{1,2†} Aoife O’Driscoll^{1,2†} Suchir Salhan^{1,2}
Catherine Arnett³ Andrew Caines^{1,2} Paula Buttery^{1,2}

¹ ALTA Institute ² Department of Computer Science & Technology, University of Cambridge
³ EleutherAI

Correspondence: lgb35@cam.ac.uk

Abstract

Language learners typically exhibit first language (L1) influence in their written second language (L2) production. We investigate whether similar patterns emerge in L2 language models (L2LMs), which are typically assessed on task-based benchmarks rather than on language use. We evaluate the use of Native Language Identification (NLI) as a method for detecting whether L2LMs exhibit human-like L1 influence. Using existing learner corpora and our novel L2 English dataset, we identify the conditions that yield the highest NLI accuracy, and show that text length but not proficiency affects performance. We then apply NLI to L2LM-generated text under various instruction-tuning and prompting conditions. We find that instruction tuning on human learner essays yields high NLI accuracy (~90%) and is necessary for detectable L1 influence. Whilst NLI accuracy is similar for L2LM and human essays, human evaluation shows that LM-generated L1 influence remains distinguishable from human writing.

1 Introduction

Multilingual language models (LMs) are trained with shared parameters across languages, facilitating crosslingual generalisation (Conneau et al., 2020). Yet bilingual and multilingual models often underperform compared to monolingual systems (Zhou and Matuselych, 2025; Xu et al., 2025; Salhan et al., 2026). However, these findings are based on evaluations that define linguistic competence primarily in terms of benchmark accuracy, such as controlled grammaticality judgement tasks (e.g., BLiMP; Warstadt et al., 2020), which quantify task performance but do not capture an LM’s linguistic behaviour (Hagendorff et al., 2023), or whether its crosslingual generalisations resemble patterns

observed in human second language acquisition (SLA). Consequently, it remains unclear whether bilingual LMs exhibit human-like behaviour in second language (L2) production.

In human SLA, a learner’s first language (L1) systematically influences written L2 production (Swan and Smith, 2001; Gonzalez-Torres and Mayo, 2025; Shatz, 2017), with errors emerging when L1 and L2 patterns diverge (Selinker, 1969). We investigate whether second language LMs (L2LMs)—models designed to simulate SLA (Aoyama and Schneider, 2024)—can generate human-like L2 text. Human-likeness is defined as the presence of L1-influenced patterns observed in human L2 writing. Our study focuses on seven L1s (Spanish, French, German, Polish, Turkish, Arabic, Mandarin Chinese) with English as the L2.

We evaluate L2LMs using Native Language Identification (NLI), a method for detecting a writer’s L1 from L2 text (Tetreault et al., 2013); this application allows us to assess whether L2LM outputs exhibit human-like L1 influence. To our knowledge, this is the first time that NLI has been applied to LM-generated text. First, we identify the conditions under which NLI is most accurate on human-written text. Based on our findings, we use the best-performing NLI method to evaluate L2LM-generated text. We address two research questions:

RQ1: What are the optimal conditions for NLI detection of L1 influence?

We compare the NLI capabilities of GPT-4 and GPT-5 under two prompting conditions (explanation vs. no explanation). We also examine text-level factors such as essay length and learner proficiency. To disentangle these typically confounded effects, we create a length-controlled L2 learner dataset spanning three proficiency levels (beginner, intermediate, advanced) and seven L1 backgrounds, with matched essay prompts. This enables the first

*Corresponding author

†Equal contribution

systematic investigation of how text length and proficiency independently affect NLI performance. Our results indicate that GPT-5 and prompting for explanations provide no significant advantage over GPT-4 without explanation. Crucially, we find that NLI performance is strongly affected by text length, but not learner proficiency, validating it as a robust detection tool across proficiency levels.

RQ2: Can L2LMs simulate human-like L2 writing?

We investigate factors that shape L1 influence in L2LM outputs, including instruction-tuning, prompt style, and the quantity of L2 pretraining data. We compare base L2LM essays with those generated under three instruction-tuning conditions: English-only, English and L1, and L1-specific L2 English, a dataset created for this study. We find that LM-generated essays are accurately classified by NLI only when instruction tuned on L1-specific L2 English. In this instruction-tuning condition, neither L2 pretraining data nor prompt style manipulations affected NLI accuracy.

To address human-likeness, we compare LM-generated essays with human-written L2 essays matched for L1, prompt, and text length. Whilst NLI accuracy is equally high on LM and human essays, analyses of named entities show that they affect NLI-detectable L1 influence in LM-generated essays but not in human-written essays. Finally, human evaluation reveals that, despite high NLI accuracy, L1 influence in L2LM-generated text is distinguishable from human writing, indicating that L2LMs do not simulate human-like written L2 production.

Contributions Our study contributes: (1) an identification of conditions that optimise NLI performance, including effects of text length and proficiency; (2) a novel human L2 corpus of prompt-matched, length-controlled essays from seven L1s and three proficiency levels; (3) a novel L1-specific L2 human learner instruction-tuning dataset, and demonstration of its necessity for detectable L1 influence in L2LM outputs; (4) a comparison of LM-generated and human-written L2 texts, marking the first application of NLI to LM-generated text; (5) human evaluation showing that high NLI accuracy does not guarantee human-like L2 production.

Models, tokenizers, and datasets are available

on Hugging Face.¹ Pretraining, instruction tuning, NLI, and prompting code is available on GitHub.²

2 Related Work

2.1 L2LMs as human proxies

There are several clear limitations to using LMs as proxies for human learning (see Cuskley et al., 2024; cf. Salhan et al., 2025). Firstly, LMs are trained on orders of magnitude more data than humans. Although smaller human-scale models may be more appropriate for cognitive comparison (Wilcox et al., 2025), they diverge from human learners in fundamental ways. Whilst efforts such as the BabyLM challenge aim to reduce this gap (Warstadt et al., 2023), the nature of the input remains fundamentally different: LMs are trained on text, whereas human learners receive multimodal, interactive input in an *embodied* manner (Cuskley et al., 2024). Additionally, learning in LMs relies on backpropagation, a learning mechanism that is biologically implausible (Lillicrap et al., 2020). Moreover, LMs may succeed on some tasks but fail to replicate human-like processing (see Cai et al., 2024; Binz and Schulz, 2023). High performance may reflect architectural flexibility rather than human-like representations, as models perform equally well when trained on unnatural inputs such as reversed text (Luo et al., 2024).

Despite differences between humans and LMs, prior work suggests that L2LMs can simulate aspects of human L2 behaviour. For example, structural priming effects in L2LMs are weaker for typologically distant L1-L2 pairs (Arnett et al., 2025), mirroring human cross-linguistic priming (Hartsuiker et al., 2004; Bernolet et al., 2007). For L2 competence, Yadavalli et al. (2023) provide indirect evidence of L1 influence: poor performance on specific L2 constructions in BLiMP is attributed to L1 transfer. For L2 production, LLMs can generate L1-influenced dialogue when prompted with human L2 data and descriptions of L1-specific grammatical traits (Gao et al., 2025). Since these L1 effects are induced and the models are not L2LMs, it remains unclear whether human-like L1 influence in L2 production can emerge naturally. Building on this, we evaluate whether L2LMs can *simulate* human-like L2 text.

¹<https://huggingface.co/ALTACambridge>

²<https://github.com/suchirsalhan/l1-influence-in-l2lms>

2.2 L1 influence in L2 production

Human SLA exhibits characteristic patterns of L1 influence. Language transfer—the effect of the L1 on the L2—can facilitate learning via typological similarities or inhibit it due to differences, causing systematic error patterns (Swan and Smith, 2001; Gonzalez-Torres and Mayo, 2025). L1 influence can also appear subtly, such as in the overuse of L1–L2 congruent structures or avoidance of L1-absent structures (Berti et al., 2023; Zomer and Frankenberg-Garcia, 2021). Since different L1s produce unique influence patterns (inter-group heterogeneity; Jarvis, 2000), a learner’s L1 can be inferred through detection-based approaches (Jarvis and Crossley, 2012). This motivates the machine learning task of NLI, which can detect “subtle and unpredicted” L1 transfer (Yang et al., 2025).

2.3 Native Language Identification

NLI involves determining an author’s L1 from text written in an L2. Traditional approaches employ feature-based methods (e.g., Tetreault et al., 2013), which can reach ~80% accuracy but require extensive feature engineering. In contrast, LLMs have established a new state of the art, exceeding 90% accuracy (e.g., Zhang and Salle, 2023; Ng and Markov, 2025).

In traditional NLI methods, classification is typically easier for low-proficiency essays (Kyle et al., 2015). The effect of learner proficiency on LLM-based methods has not yet been tested. However, it has been shown that NLI accuracy improves with increased text length (Zhang and Salle, 2023; Nicholls and Alperin, 2025). Crucially, in L2 writing, text length and proficiency are strongly correlated (Martínez, 2018), and the individual effect of each factor, whilst controlling for the other, has not yet been systematically investigated.

3 Methodology

To investigate whether L2LMs exhibit human-like L1 influence (RQ2), we evaluated their outputs using NLI under the optimal conditions identified in RQ1.

3.1 RQ1

To identify optimal NLI conditions, we conduct two experiments. Experiment 1 compares model (GPT-4 vs. GPT-5) and prompting (explanation vs. no explanation) conditions to identify the best-performing configuration for L1 detection. Experi-

ment 2 examines text-level factors, testing the effect of learner proficiency whilst controlling for text length to isolate its impact on NLI performance.

Experiment 1 Following Zhang and Salle (2023), the model performs NLI as a closed-set, zero-shot task (see Figure 8, Appendix B). Four model conditions were compared: two models (GPT-4 and GPT-5³) under two prompting conditions – requesting explanation vs. no explanation. Whilst GPT-4 currently achieves state-of-the-art performance on NLI benchmarks (Zhang and Salle, 2023; Ng and Markov, 2025), GPT-5 has not yet, to the best of our knowledge, been evaluated for this task. Explanation prompting is motivated by evidence that generating explanations can improve LLM performance (Dhaini et al., 2025). NLI was run five times for each model condition to ensure reliability.

We used a subset of the Write & Improve Corpus 2024 (W&I; Nicholls et al., 2024), which includes L1 and CEFR⁴ proficiency metadata. The subset comprised 147 essays (21 per L1; seven per CEFR level A, B, and C, corresponding to beginner, intermediate, and advanced, respectively) with a mean length of 884 (SD=505) characters.

Experiment 2 We examined the effect of proficiency on NLI accuracy whilst controlling for text length. As prior work suggests that text length affects accuracy (see Section 2.3), it was crucial to isolate the effects of each factor and ensure that NLI performance is comparable across beginner, intermediate, and advanced proficiency levels. We included two text length controls: (1) W&I essays were artificially truncated to the median length of A-level essays (291 characters), with shorter essays (n = 24) included without modification; (2) 466 longer essays (800-1,200 characters) were from our own L2 dataset (see Section 3.3). For comparison, the effect of proficiency without controlling for text length was also evaluated using the original W&I subset. The NLI procedure followed the design of Experiment 1.

3.2 RQ2

We investigated whether L2LMs can simulate human-like L2 writing by examining the effects of four independent variables: instruction-tuning condition, amount of L2 pretraining data received

³Accessed Jan-Feb 2026.

⁴Common European Framework of Reference for Languages (CEFR) rating.

(25%, 50%, or 75% of total L2 tokens), prompt style, and prompt topic.

3.2.1 L2LM Training

Pretraining Following Arnett et al. (2025), we pretrained seven 250M-parameter GPT-2 L2LMs on 5B tokens, one per L1, with English as the L2. Each model has 24 layers, 14 attention heads, a hidden size of 896, a 512-token context window, and a 50k-token vocabulary (see Appendix A.1).

The pretraining dataset consisted of L1 and L2 data in a 2:1 ratio. L1 data (3.33B tokens) were sourced from CulturaX (Nguyen et al., 2024), and L2 data (1.67B tokens) from the English subset of FineWeb-Edu (Lozhkov et al., 2024). We trained 50k-vocab bilingual (L1-English) byte-level BPE tokenizers on 2M sentences with a 1:1 L1:L2 mix, and pretokenized all training data into 512-token sequences.

Pretraining followed a two-phase schedule. In Phase 1 (0-50%), the model was exposed exclusively to L1 (2.5B tokens). Phase 2 (51-100%) interleaved the L2 data (1.67B) with the remaining L1 data (0.83B) at a 2:1 ratio.

Instruction tuning We investigated the effect of instruction tuning on L1 influence, with the base (non-instruction-tuned) model as a baseline. Full-weight instruction tuning was applied under three dataset conditions: (1) English-only, (2) English and L1 (1:1), and (3) L1-influenced L2 English. The English-only dataset comprised 30,000 generic QA pairs from Alpaca English (Taori et al., 2023). The English + L1 dataset also contained 30,000 QA pairs, split equally between Alpaca English and its translation for each L1 (see Appendix A.2 for dataset references). We created seven L1-specific L2 English datasets from the EFCAMDAT (Geertzen et al., 2013) and W&I (Nicholls et al., 2024) learner corpora, each containing 10,000 prompt-essay pairs, except Polish (n=370) and Turkish (n=4,847) due to limited corpus data (see Appendix A.2 for detailed dataset composition).

To examine the effect of L2 data exposure, instruction tuning was conducted at one of three pretraining checkpoints: after 25%, 50%, and 75% of total L2 tokens. For ease of comparison with human responses, we refer to these as beginner, intermediate, and advanced proficiency levels, but we do not imply direct equivalence.

In total, 84 model versions were created, cov-

ering all combinations of seven L1s, four model variants (one base and three instruction-tuned), and three pretraining checkpoints.

3.2.2 Prompt engineering

As monolingual and bilingual prompting have been shown to affect LLM output (Yuan et al., 2025), we compare three prompting styles: basic, mixed-language, and ‘language-to-thought’ (L2T). Basic prompts align with human L2 writing assessment tasks. Mixed-language prompts combine L1 and L2 by translating sections of the basic prompts (Srivastava and Singh, 2021; Singh et al., 2024; Lin et al., 2022). L2T prompts instruct the model to ‘think’ in the L1 before responding in English (see Kang and Kim, 2025), and are otherwise identical to the basic prompts.

```
Prompt topic: Holidays
Prompt style: Mixed language
Prompt:
Tu amiga, Sophia, has emailed you about her recent holiday. Write an email to her, describiéndole tus vacaciones favoritas.
Write around 150 words.
```

Figure 1: Example of a mixed-language prompt (Spanish L1).

We created nine distinct prompts by combining three styles (basic, mixed-language, and L2T) with three topics (hobbies, holiday, family). An example is provided in Figure 1 (see Appendix A.3 for all prompts).

Each model version was prompted twice on each of the nine prompts, resulting in a total of 1,512 LM-generated essays. Outputs were constrained to 250-300 tokens (~800-1,200 characters).

3.3 Novel L2 learner dataset

We created a novel dataset of learner English essays by collecting responses to nine prompts from seven L1 backgrounds across three proficiency levels, controlling for text length (RQ1). This design enabled direct comparison with model-generated text by matching prompts between human participants and LMs (RQ2). For each proficiency level and L1, at least six participants were recruited. In total, the dataset comprised 466 essays. Full ethical approval was granted by the Department of Computer Science and Technology at the University of Cambridge.

Participants 174 participants (M age=30; 42.5% female) were recruited via Prolific. The task took

approximately 30 minutes, and participants were compensated at the local living wage. Informed consent was obtained in the participant’s native language.

Design and procedure Each participant responded to three prompts (each topic and style were presented exactly once; see Appendix B.3). Prompt sequences were pseudo-randomised to counter-balance the nine prompts across L1-proficiency subgroups, minimising dependencies between prompts, and ensuring that each prompt was answered at least twice per subgroup. Essay length was constrained to 800-1,200 characters. Participants were instructed not to use external resources and were informed that the task was not a test. Essays were manually screened for signs of generative AI usage.

3.4 Evaluation

Model-generated and human-written texts were evaluated with NLI using the GPT-4 model (no explanation) (see Section 4.1). In line with Zhang and Salle (2023), if the model predicted an L1 outside the seven L1 classes, the model was re-prompted until one of the known L1s was predicted.

LLMs may rely on named entities as superficial cues to the writer’s L1 (Uluslu et al., 2025). Using Named Entity Recognition (NER), we replaced any entities present in human-written and LM-generated essays with neutral English ones (see Appendix A.4). NLI was performed on both the original and NER-normalised texts.

To evaluate the use of NLI as a metric of human-likeness, three computational linguists judged whether texts were human-written or LM-generated. Each evaluator reviewed 42 texts (21 human, 21 LM), balanced across L1, prompt style, and instruction-tuning condition (126 essays in total). We also assessed the potential effect of LM generation artefacts on judgements by manually editing the essays (e.g., removing unusual characters) provided to one of the evaluators (see Appendix A.4).

4 Results: RQ1

4.1 Experiment 1

Experiment 1 tested NLI performance across four model conditions. Table 1 summarises accuracy by model and prompting condition. No significant differences were found across the five NLI runs for any condition (Table 9, Appendix B).

	explanation	no explanation
GPT-4	83.3	81
GPT-5	79.7	80.4

Table 1: NLI accuracy (%) per model condition (averaged across five iterations).

All four model conditions performed significantly above chance (1/7; Wilcoxon signed rank test, Holm-adjusted $p < .001$). Accuracy did not differ between model conditions (Friedman test, $\chi^2(3) = 3.82, p = .28$). By L1, GPT-5 showed higher accuracy for Arabic ($\chi^2(3) = 11.54, p < .01$), but post-hoc pairwise comparisons were not significant after Holm correction ($p \geq .12$). No other L1 differences were observed. Confusion matrices are shown in Figure 9 (Appendix B). Kruskal-Wallis tests confirmed no significant L1 differences within any model (Table 9, Appendix B).

Based on these results, the GPT-4 no-explanation condition was selected for NLI, as GPT-5 showed no advantage, and explanation prompts provided no benefit, with occasional hallucinations (see Figure 10, Appendix B).

4.2 Experiment 2

We examined the effect of proficiency on NLI accuracy whilst controlling for text length. We report only the results of the GPT-4 no-explanation condition (see Appendix B for other conditions).

Original W&I subset NLI accuracy was 81%, noticeably lower than the state-of-the-art performance (91.7%) reported by Zhang and Salle (2023) on TOEFL11. This is partly attributable to the shorter essays in our dataset (mean 884 vs. 1,785 characters in Zhang and Salle, 2023). Correctly classified W&I essays were significantly longer than misclassified ones (Figure 2; see also Table 10 and Figure 11, Appendix B).

Proficiency and text length are highly correlated in this dataset (Spearman’s $\rho(145) = .87, p < .001$). Proficiency significantly affected NLI accuracy (Kruskal-Wallis: $p < .01$), with A-level essays classified significantly less accurately than C-level essays (post-hoc Wilcoxon, $p < .01$; Table 11, Appendix B).

Truncated W&I subset Truncating essays to ≤ 291 characters reduced overall NLI accuracy to 67%, which remained significantly above chance ($p < .001$). A Kruskal-Wallis test revealed no sig-

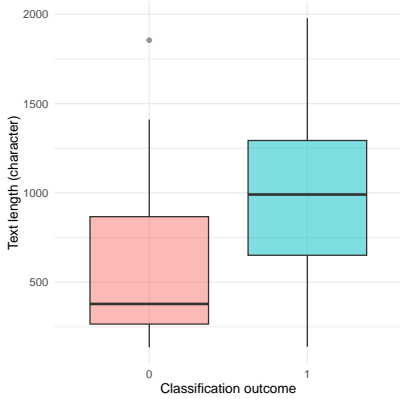


Figure 2: Boxplot of text length by NLI classification accuracy (averaged over five runs). Texts with mean accuracy < 1 are classified as 0 (incorrect); mean accuracy = 1 are classified as 1 (correct).

nificant differences in accuracy between A-, B- and C-level essays ($\chi^2(2) = 0.45, p = .80$).

Length-controlled L2 dataset NLI accuracy was 83% (95%CI[79.33, 86.34], $p < .001$) on our novel dataset ($M = 906 \pm 109$ characters). Per-L1 accuracy (Figure 3a) was significantly above chance (see Table 3b). A generalised linear mixed model (GLMM) was fitted with L1, proficiency, prompt style, prompt topic, and NER condition as fixed effects, and user ID as a random intercept (see Table 14, Appendix B.3 for model summary). Crucially, proficiency had no significant effect on NLI accuracy ($p \geq .09$).

True label \ Predicted label	ARA	CHI	FRE	GER	POL	SPA	TUR
ARA	59	17	3	6	0	5	3
CHI	0	55	1	2	0	4	0
FRE	2	2	45	1	1	7	0
GER	0	0	2	61	1	0	0
POL	0	0	0	5	64	0	0
SPA	0	0	2	1	1	64	0
TUR	2	0	0	5	2	4	39

(a) Confusion matrix

L1	F1	P	R	p
Spanish	0.84	0.76	0.94	$<.001$
French	0.81	0.85	0.78	$<.001$
German	0.84	0.75	0.95	$<.001$
Polish	0.93	0.93	0.93	$<.001$
Turkish	0.83	0.93	0.75	$<.001$
Arabic	0.76	0.94	0.63	$<.001$
Chinese	0.81	0.74	0.89	$<.001$

(b) Precision, recall, and F1

Figure 3: NLI performance on human-written essays: (a) confusion matrix; (b) per-L1 precision (P), recall (R), and F1 scores. p -values are from one-tailed binomial tests.

Overall, Experiment 2 showed that, when text length was controlled, proficiency had no significant effect on NLI accuracy, both for short (291 characters) and longer (800-1,200 characters) essays, validating GPT-4 as a detection tool across proficiency levels.

5 Results: RQ2

5.1 Effects of instruction tuning

We evaluated the L2LM generations under three instruction-tuning conditions, as well as from the base models as a baseline.

First, the base models were tested on BLiMP (Warstadt et al., 2020) and MultiBLiMP (Jumelet et al., 2026) (see Figure 13, Appendix C.1). MultiBLiMP results suggest that our training regime did not lead to catastrophic forgetting. BLiMP scores increased as L2 data quantity increased. The seven models (one per L1) did not differ significantly in BLiMP and MultiBLiMP scores (see Table 15, Appendix C.1).

5.1.1 Base model

Mean NLI accuracy was 23.5% (95%CI[19.36, 28.15]), remaining significantly above chance ($p < .01$). However, this effect was driven by Chinese, where high recall (0.91) but low precision (0.16) resulted in many false positives (see Table 2; Figure 4a). NLI accuracy was not above chance for any other L1.

L1	F1	P	R	p
Spanish	0.22	0.38	0.15	.51
French	0.34	0.75	0.22	.08
German	0.26	0.40	0.19	.24
Polish	0.08	1.00	0.04	1.0
Turkish	0.04	1.00	0.02	1.0
Arabic	0.23	1.00	0.13	.67
Chinese	0.27	0.16	0.91	$<.001$

Table 2: NLI precision (P), Recall (R), and F1 scores per L1 background for the base model. p -values are from one-tailed binomial tests.

5.1.2 Comparing instruction-tuning conditions

A GLMM was fitted with the fixed effects of L1, instruction-tuning condition, proficiency, prompt style, prompt topic, and NER condition, including the two-way interactions between instruction tuning and NER, prompt style and prompt topic, and instruction tuning and prompt style. Model version was included as a random intercept. Model summary is shown in Table 3 (full results in Table 16, Appendix C).

Baseline accuracy was below chance ($p < .001$). Relative to this baseline, L1 significantly affected performance: Polish, Turkish, and Arabic decreased accuracy, whilst the significant advan-

Term	Estimate	Std. Error	z value	p
(Intercept)	-2.92	0.48	-6.07	<.001
IT: English-only	-0.16	0.52	-0.31	.76
IT: English + L1	0.52	0.48	1.08	.28
IT: L2 English	6.75	0.51	13.33	<.001
NER	0.03	0.26	0.13	.90
Style: L2T	0.10	0.48	0.22	.83
Style: Mixed	2.75	0.43	6.43	<.001
Topic: Hobbies	0.64	0.31	2.04	<.05
Topic: Holiday	0.69	0.31	2.19	<.05
IT: English-only * NER	0.04	0.38	0.12	.91
IT: English + L1 * NER	-0.12	0.35	-0.34	.73
IT: L2 English * NER	-3.37	0.40	-8.35	<.001
Style: Mixed * Topic: Hobbies	-1.03	0.39	-2.62	<.01
IT: English-only * Style: L2T	1.34	0.56	2.41	<.05
IT: English-only * Style: Mixed	-1.22	0.53	-2.33	<.05
IT: L2 English * Style: Mixed	-2.22	0.44	-5.00	<.001

Table 3: Model summary. The intercept is set to base model, Spanish L1, basic prompt style, family topic, and original essays (no NER).

tage of Chinese L1 is attributable to false positives (see Table 2). Proficiency had no significant effect. Neither English-only nor English + L1 instruction-tuning conditions significantly affected accuracy ($p > .05$). In contrast, L2 English instruction tuning significantly improved NLI accuracy ($p < .001$). Whilst NER had no main effect ($p = .90$), it significantly dampened the improvement from L2 English instruction tuning ($p < .001$). Prompt topic influenced accuracy: essays on hobbies and holiday topics were classified more accurately than the baseline ($p < .05$). Prompt style effects depended on both topic and instruction-tuning condition. Mixed-language prompting improved accuracy in the base model (see Appendix C.3). However, this benefit was attenuated for the hobbies topic ($p < .01$), as well as under English-only ($p < .05$) and L2 English ($p < .001$) instruction-tuning conditions. L2T prompts had no significant main effect but significantly increased accuracy under English-only instruction tuning ($p < .05$).

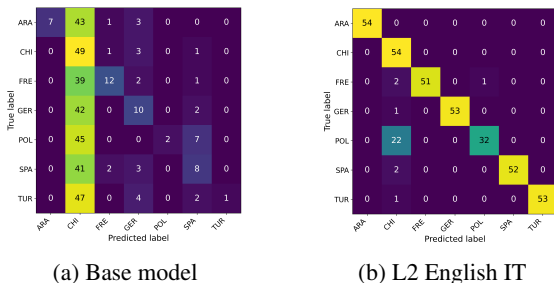


Figure 4: Confusion matrices for NLI on essays generated by the base model and L2 English instruction-tuning model conditions.

Overall, the L2 English instruction-tuning condi-

tion resulted in the highest NLI accuracy. Neither the English-only nor English and L1 instruction-tuning conditions improved NLI performance relative to the base model (see Appendix C for detailed results).

5.1.3 L2 English instruction tuning

Table 4 reports NLI precision and recall for original and NER-normalised essays. On the original essays, overall NLI accuracy was 92.33% (95% CI [89.2%, 94.80%]). Classification accuracy was significantly above chance (1/7) for all L1s ($p < .001$). Precision was near ceiling for all L1s except Chinese (0.65), reflecting false positive predictions (see Figure 4). Recall was near ceiling (0.94-1.00) for all L1s except Polish (0.59), likely due to its smaller instruction-tuning dataset ($n = 370$ examples; Section 3.2.1). Accuracy did not differ significantly by prompt style ($\chi^2(2) = 5.45, p = .07$) or by prompt topic once accounting for interactions between prompt topic and style (see Table 20, Appendix C.6).

NER normalisation reduced accuracy to 55.82% (95% CI [50.65%, 60.90%]) but remained significantly above chance overall ($p < .001$) and for all L1s except Polish ($p = 1.0$), likely due to the small size of the instruction-tuning dataset (Section 3.2.1). Per-L1 comparisons indicated that NER normalisation significantly reduced NLI accuracy for all L1s ($p < .001$) except Spanish ($p = .13$) (Table 21, Appendix C.6).

L1	Learner IT (original)				Learner IT (NER)			
	F1	P	R	p	F1	P	R	p
Spanish	0.98	1.00	0.96	<.001	0.77	0.68	0.89	<.001
French	0.97	1.00	0.94	<.001	0.72	0.97	0.57	<.001
German	0.99	1.00	0.98	<.001	0.77	1.00	0.63	<.001
Polish	0.73	0.97	0.59	<.001	0.04	0.25	0.02	1.0
Turkish	0.99	1.00	0.98	<.001	0.67	0.83	0.56	<.001
Arabic	1.00	1.00	1.00	<.001	0.41	1.00	0.26	<.05
Chinese	0.79	0.65	1.00	<.001	0.44	0.28	0.98	<.001

Table 4: Precision (P), Recall (R), and F1 scores per L1 background for learner essay instruction-tuning condition: original and NER-normalised essays.

5.2 Comparing human-written and LM-generated texts

To compare NLI accuracy on human-written and LM-generated essays (from the L2 English instruction-tuning condition), a GLMM was fitted with the fixed effects of animacy (human vs.

LM), L1, NER, prompt style, prompt topic, and proficiency. We included interactions between animacy and L1, animacy and NER, and animacy and prompt topic. User ID (model version/participant) was included as a random intercept. Model summary is in Table 5 (full results in Table 22, Appendix C.7).

Term	Estimate	Std. Error	<i>z</i> value	<i>p</i>
(Intercept)	4.13	0.76	5.45	<.001
L1: Turkish	-2.48	0.79	-3.12	<.01
L1: Arabic	-3.25	0.72	-4.54	<.001
Animacy: LM	1.46	1.34	1.09	.28
NER	0.05	0.22	0.23	.82
Animacy: LM * L1: Polish	-5.33	1.74	-3.07	<.01
Animacy: LM * NER	-3.94	0.46	-8.66	<.001
Animacy: LM * Topic: Hobbies	0.89	0.40	2.24	<.05
Animacy: LM * Topic: Holiday	1.00	0.41	2.46	<.05

Table 5: Model summary. The intercept is set to human-written, Spanish L1, basic prompt style, family topic, and original essays (no NER).

Baseline accuracy was significantly above chance ($p < .001$). Turkish and Arabic L1 significantly decreased NLI accuracy ($p < .01$), although performance remained significantly above chance (Table 3b). There were no significant main effects of animacy, NER, prompt style, prompt topic, or proficiency ($p < .05$). However, a strong negative interaction between animacy and NER was found ($p < .001$), indicating that NER significantly reduced NLI accuracy for LM-generated essays but not for human-written essays. The interaction between animacy and L1 was significant for Polish ($p < .01$), with LM-generated essays being classified less accurately than human essays. Additionally, LM-generated essays were classified more accurately than human essays for hobbies and holiday topics ($p < .05$).

5.2.1 Human evaluation

Mean evaluator accuracy was 87.3%. Classification accuracy did not significantly differ between evaluators (McNemar’s test, $p = .61$), and removing unusual characters from LM outputs had no effect ($p = 1.0$). Although evaluators misclassified human-written texts as LM-generated three times more often than the reverse, this difference was not statistically significant ($p = .08$). No other conditions significantly affected evaluation outcomes (see Appendix C.8).

6 Discussion

LLM-based NLI methods reliably detect L1 influence in L2LM outputs; equally, human evaluation

consistently distinguished LM-generated essays from human L2 production, indicating that NLI detectability should not be interpreted as a measure of human-likeness.

GPT-4 (without explanation prompting) was selected as our NLI tool, as GPT-5 and/or prompting for explanation offered no advantage. NLI successfully detected L1 influence across all proficiency levels once text length was held constant.

In L2LM-generated text, L1 influence was absent in base model outputs and in generic (English-only or English and L1) instruction-tuning conditions. As we did not find evidence of catastrophic forgetting, these results indicate that L1 exposure alone is insufficient to produce detectable L1 influence in L2LM output. L1 influence emerged only in essays generated by LMs that were instruction-tuned on L1-specific L2 English datasets. Notably, these instruction-tuning datasets were only a third of the size of the other instruction-tuning datasets, suggesting that instruction-tuning dataset specificity, rather than quantity, drove L1 influence.

Within L2 English IT, L1 influence was consistently detected across model checkpoints, showing that the quantity of L2 pretraining data had no effect on L1 influence. Additionally, prompting conditions did not affect NLI accuracy: L1 influence was present both with and without explicit L1 elicitation.

NLI accuracy was comparable for human-written and LM-generated essays from the L2 English instruction-tuning condition. For these L2LMs, NER normalisation reduced NLI accuracy, yet performance remained above chance, highlighting that L1 influence went beyond superficial cues. Notably, the adverse effect of NER normalisation on NLI performance was not observed for human-written essays. Moreover, human evaluation clearly distinguished LM-generated from human-written essays.

Additionally, Chinese was overpredicted as the L1 across all LM outputs, but this bias was absent from human essays, suggesting that it stems from pretraining data artefacts rather than limitations of the NLI method. Overall, whilst using NLI as a metric suggests that L1 influence in human and LM essays may be comparable, further analyses reveal that L2LM production remains distinctly non-human-like.

7 Conclusion

Our findings show that L2LMs produce L1 influence only when instruction-tuned on L1-specific L2 English. However, as this L1 influence did not resemble human L2 production, NLI detectability ultimately should not be used as a measure of human-likeness. Overall, L2LM outputs remain distinctly non-human-like, limiting their use as proxies for human L2 learners.

Limitations and Future Directions

NLI accurately detected L1 influence in L2LM-generated essays; however, its precise nature remains unclear. Future work should aim to elucidate the properties of these L1 effects, perhaps by incorporating linguistically motivated fine-grained analyses. Additionally, Chinese false positives in NLI were prevalent across all L2LM essays regardless of instruction-tuning condition. We hypothesise that more rigorous pretraining data cleaning may mitigate this overprediction bias. Furthermore, although the combination of L1-specific pretraining and instruction tuning produced detectable L1 influence, pretraining alone was insufficient. It remains unclear whether instruction tuning alone, or its combination with pretraining, was responsible for eliciting L1 influence. Finally, we limited our L2 analysis to English. Extending this work to other, less data-rich L2s could prove insightful in future research.

Acknowledgements

This work was supported by funding from Cambridge University Press & Assessment. Compute facilities (32× NVIDIA A100-SXM4 GPUs) were provided by EleutherAI for model pretraining. With thanks to Krystian Balioskorski, Burak Bugrul, Aisha Delair, Ahmed Al-Jabri, Aidan Jones, Clare O’Driscoll, Rose Rociola, and EJ Zhou for their help with the translation of the survey. We also thank Bianca Ganescu, Przemek Kubiak, and Filip Trhlik for their work evaluating model outputs, and Dmytro Mai for technical assistance. We are grateful to Shiva Taslimipoor and Denise Löfflad for their helpful comments on an earlier draft of the manuscript. We thank the anonymous reviewers for their thoughtful comments.

Authors’ Contributions. Conceptualisation: L.B.; Investigation: L.B., L.G., A.O.D.; Method-

ology: L.B., L.G., A.O.D.; Data curation: L.G.; Formal analysis: L.B., A.O.D.; Software: S.S. (pre-training), L.B., L.G., A.O.D. (instruction tuning and NLI); Funding acquisition: P.B., A.C.; Supervision: S.S., P.B., A.C., C.A.; Visualisation: L.G., L.B., A.O.D.; Resources: C.A., P.B., A.C.; Writing – original draft: L.B., L.G., A.O.D.; Writing – review and editing: L.B., L.G., A.O.D., S.S., C.A., A.C., P.B.

References

- Tatsuya Aoyama and Nathan Schneider. 2024. [Modeling nonnative sentence processing with L2 language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.
- Catherine Arnett, Tyler A. Chang, James A. Michaelov, and Ben Bergen. 2025. [On the acquisition of shared grammatical representations in bilingual language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20707–20726, Vienna, Austria. Association for Computational Linguistics.
- Sarah Bernolet, Robert J Hartsuiker, and Martin J Pickering. 2007. [Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition](#). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):931.
- Barbara Berti, Andrea Esuli, and Fabrizio Sebastiani. 2023. [Unravelling interlanguage facts via explainable machine learning](#). *Digital Scholarship in the Humanities*, 38(3):953–977.
- BERTIN Project. 2023. [alpaca-spanish: Spanish translation of the cleaned Stanford Alpaca dataset](#).
- Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 37–56.
- Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. [Monolingual or multilingual instruction tuning: Which makes a better Alpaca](#). In *Findings of the Association for Computational Linguistics: EACL 2024*.
- Hasna Chouikhi, Manel Aloui, Cyrine Ben Hammou, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. Llama & gemma: Enhancing llms through arabic instruction-tuning.

- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Emerging cross-lingual structure in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.
- Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. [The limitations of large language models for understanding human language and cognition](#). *Open Mind*, 8:1058–1083.
- Mahdi Dhaini, Juraj Vladika, Ege Erdogan, Zineb Attaoui, and Gjergji Kasneci. 2025. [Can LLM-generated textual explanations enhance model classification performance? an empirical study](#). In *International Conference on Artificial Neural Networks and Machine Learning (ICANN)*, pages 192–204. Springer.
- Emplocity. 2023. Owca: Optimized and well-translated customization of alpaca. <https://huggingface.co/datasets/emplocity/owca>. Polish translation of the Alpaca instruction dataset. Accessed: 2026-03.
- Rena Gao, Xuotong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, and Jey Han Lau. 2025. Can llms simulate l2-english dialogue? an information-theoretic analysis of l1-dependent biases. *arXiv preprint arXiv:2502.14507*.
- Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. [Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database \(EFCAMDAT\)](#). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadilla Proceedings Project, pages 240–254.
- Paul Gonzalez-Torres and María del Pilar García Mayo. 2025. [Grammatical transfer errors in EFL writing: Impact of Spanish influence, proficiency levels, and task types](#). *International Journal of Learning, Teaching and Educational Research*, 24(5):360–375.
- Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. [Machine psychology](#). *arXiv preprint arXiv:2303.13988*.
- Robert J Hartsuiker, Martin J Pickering, and Eline Veltkamp. 2004. [Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals](#). *Psychological science*, 15(6):409–414.
- Scott Jarvis. 2000. [Methodological rigor in the study of transfer: Identifying l1 influence in the interlanguage lexicon](#). *Language learning*, 50(2):245–309.
- Scott Jarvis and Scott A Crossley. 2012. [Approaching language transfer through text classification: Explorations in the detection-based approach](#), volume 64. Multilingual Matters.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. [Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Transactions of the Association for Computational Linguistics*, 14:193–216.
- Eojin Kang and Juae Kim. 2025. [When language shapes thought: Cross-lingual transfer of factual knowledge in question answering](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4868–4873.
- Kristopher Kyle, Scott A Crossley, and You Jin Kim. 2015. [Native language identification and writing proficiency](#). *International Journal of Learner Corpus Research*, 1(2):187–209.
- Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. 2020. [Back-propagation and the brain](#). *Nature Reviews Neuroscience*, 21(6):335–346.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shrutu Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. [Fineweb-edu: the finest collection of educational content](#). Hugging Face.
- Xiaoliang Luo, Michael Ramscar, and Bradley C Love. 2024. [Beyond human-like processing: Large language models perform equivalently on forward and backward scientific text](#). *arXiv preprint arXiv:2411.11061*.
- Ana Cristina Lahuerta Martínez. 2018. [Analysis of syntactic complexity in secondary education efl writers at different proficiency levels](#). *Assessing Writing*, 35:1–11.
- Yee Man Ng and Iliia Markov. 2025. [Leveraging open-source large language models for native language identification](#). In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28, Abu Dhabi, UAE. Association for Computational Linguistics.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

- Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. [The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English](#). Technical report. Cambridge University Press & Assessment.
- Robin Nicholls and Kenneth Alperin. 2025. [Cross-genre native language identification with open-source large language models](#). In *Proceedings of the 2nd LUHME Workshop*, pages 103–108, Bologna, Italy. UP - Universidade do Porto (<https://doi.org/10.21747/978-989-9193-73-4/lan2>), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland and UAH - Universidad de Alcalá.
- Jonathan Pacifico. 2024. [French-Alpaca-dataset-Instruct-110K](#). 110k French instruction-response pairs generated with GPT-3.5-turbo in Alpaca format.
- Suchir Salhan, Andrew Caines, and Paula Buttery. 2025. Pedagogical alignment of llms requires diverse cognitively-inspired student proxies. In *Proceedings of the First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models at NeurIPS 2025*, San Diego, California, USA. NeurIPS Foundation.
- Suchir Salhan, Ej Zhou, Laura Barbenel, Aoife O’Driscoll, Lily Goulder, Lucas Resck, Catherine Arnett, and Paula Buttery. 2026. Glints of gold or troubling waters? can a school of merged monolingual goldfish models swim in bilingual seas?
- Larry Selinker. 1969. Language transfer. *General linguistics*, 9(2):67.
- Itamar Shatz. 2017. Native language influence during second language acquisition: A large-scale learner corpus analysis. In *Proceedings of the Pacific Second Language Research Forum (PacSLRF 2016)*, pages 175–180. Japan Second Language Association Hiroshima, Japan.
- Silk Road Project. 2023. [Alpaca-data-gpt4-chinese](#). Chinese instruction-following dataset in Alpaca format (52k samples) with Chinese and English fields.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hetiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.
- Vivek Srivastava and Mayank Singh. 2021. [Challenges and considerations with code-mixed nlp for multilingual societies](#). Preprint, arXiv:2106.07823.
- Michael Swan and Bernard Smith. 2001. *Learner English: A teacher’s guide to interference and other problems*, volume 1. Cambridge University Press.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. [A report on the First Native Language Identification Shared Task](#). In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.
- TFLai. 2023. [Turkish-alpaca](#). Turkish instruction-following dataset in Alpaca format (51.9k samples).
- Ahmet Yavuz Uluslu, Tannon Kew, Tilia Ellendorff, Gerold Schneider, and Rico Sennrich. 2025. [Robust native language identification through agentic decomposition](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8398–8414, Suzhou, China. Association for Computational Linguistics.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. [Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Sara Cushing Weigle. 2010. *Assessing Writing*. Cambridge University Press.
- Ethan Gotlieb Wilcox, Michael Y Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. [Bigger is not always better: The importance of human-scale language modeling for psycholinguistics](#). *Journal of Memory and Language*, 144:104650.
- Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.
- Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. [SLABERT talk pretty one day: Modeling second language acquisition with BERT](#). In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Haiyin Yang, Zoey Liu, and Stefanie Wulff. 2025. [Using NLI to identify potential collocation transfer in L2 English](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 687–696, Vienna, Austria. Association for Computational Linguistics.

Shuzhou Yuan, Zhan Qu, Mario Tawfelis, and Michael Färber. 2025. [From monolingual to bilingual: Investigating language conditioning in large language models for psycholinguistic tasks](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1028–1040, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Wei Zhang and Alexandre Salle. 2023. [Native language identification with large language models](#). *arXiv preprint arXiv:2312.07819*.

Yuwen Zhou and Yevgen Matushevych. 2025. [Curse of bilinguality: Evaluating monolingual and bilingual language models on Chinese linguistic benchmarks](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 622–630, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Gustavo Zomer and Ana Frankenberg-Garcia. 2021. [Beyond grammatical error correction: Improving L1-influenced research writing in English using pre-trained encoder-decoder models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2534–2540, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Methodology

A.1 L2LM Training

Training Hyperparameters Training hyperparameters are detailed in Table 6.

Parameter	Value / Notes
Tokenization	
Tokenizer type	Byte-Level BPE
Vocabulary size	50,000
Training corpus	L1 (CulturaX) + EN FineWeb-Edu datasets
Model Architecture	
Model type	GPT2LMHeadModel
Sequence length	512
Embedding dimension	896
Number of layers	24
Number of attention heads	14
BOS / EOS / PAD IDs	0 / 1 / 2
Training	
Batch size	16
Gradient accumulation	8 (effective batch size = 128)
Optimizer	AdamW (betas=(0.9,0.95), weight decay=0.1)
Learning rate	2e-4 (linear warmup + cosine decay)
Total training tokens	5B
Mixed precision	bfloat16 (CUDA autocast)
Gradient clipping	1.0 (norm)
GPUs / Distributed	Supports DDP, auto-detect rank/local rank
Checkpoint frequency	Curriculum-based (end of phase 1 + phase 2 L2 exposure stages: 25%, 50%, 75%, 100%)

Table 6: Hyperparameters for tokenization and model pretraining

Pretraining Each L2LM is pretrained on 8 NVIDIA A100-SXM4 GPUs (80GB of HBM2e memory per GPU), taking approximately 5 hours per model (Table 7). Preprocessing (data streaming and pre-tokenization) are CPU-bound, taking approximately 2 hours per model on a single node with 128 CPU threads.

Phase	Tokens (B)	Pretokenization (h)	Training Time (h)	Compute (GPU-h) per model
L1 dataset preprocessing	3.33	≈ 0.5	–	4 (8 GPUs × 0.5h)
EN dataset preprocessing	1.67	≈ 0.25 – 0.3	–	2.4 (8 GPUs × 0.3h)
Phase 1 training	2.5	–	2.5	20(8 GPUs × 2.5 h)
Phase 2 training	2.5	–	2.5	20(8 GPUs × 2.5 h)

Table 7: L2LM pretraining time per model.

Checkpoint	L1	L2	Total
Beginner (25%)	2.71B	0.42B	3.13B
Intermediate (50%)	2.92B	0.84B	3.75B
Advanced (75%)	3.13B	1.25B	4.38B

Table 8: Number of L1 and L2 tokens seen at each pretraining checkpoint.

A.2 Instruction Tuning

English-only dataset We used the Alpaca English dataset (Taori et al., 2023). It was filtered to remove any rows with an additional input (i.e., not just QA). After cleaning the dataset (e.g., removing empty rows), the dataset contained a total of 31,311 QA pairs.

English + L1 datasets We used the following Alpaca datasets for Spanish (BERTIN Project, 2023), French (Pacifico, 2024), German (Chen et al., 2024), Polish (Emplocity, 2023), Turkish (TFLai, 2023), Arabic (Chouikhi et al., 2024), and Chinese (Silk Road Project, 2023). After cleaning the datasets (e.g., removing non-text entries), we randomly selected 15,661 examples from each of the English and L1 datasets, yielding 31,322 examples in total. The datasets were normalised to ensure compatibility across different scripts.

Human L2 English instruction-tuning datasets These datasets were created by compiling essays from the W&I (Nicholls et al., 2024) and EFCAMDAT corpora (Geertzen et al., 2013). We included 10,000 QA pairs for all L1s except Polish ($n = 370$) and Turkish ($n = 4,847$) due to limited corpus data. All available essays from the W&I corpus were used, which is significantly smaller than the EFCAMDAT corpus. The EFCAMDAT contains relatively few C-level essays; all of these were included for each L1, with the remaining essays split evenly between A- and B-level according to corpus availability. This is illustrated in Figure 5.

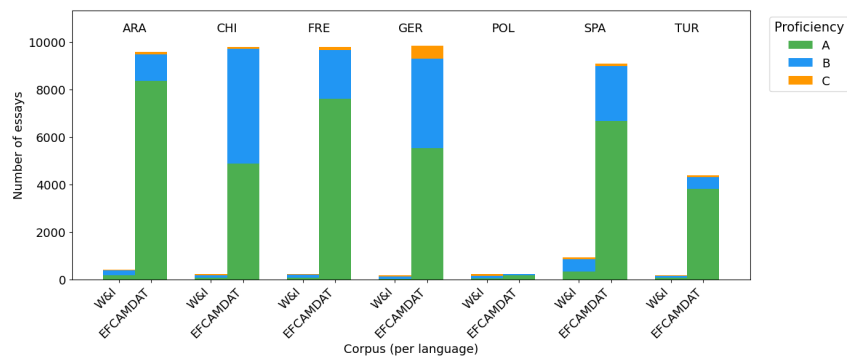


Figure 5: Proportion of essays in the L2 English instruction-tuning dataset from each corpus (W&I and EFCAMDAT) and proficiency level (A, B, C).

A.3 Prompt engineering

Text-generation prompting Text was generated using sampling (temperature = 0.7, top- $p = 0.9$) with a repetition penalty of 1.2 and blocked 4-grams. Outputs were constrained to 250-300 tokens (≈ 800 -1,200 characters).

Prompt design Prompt design followed guidelines for low-proficiency learners (Weigle, 2010) to ensure accessibility across all three proficiency levels. The prompts given to L2 learners and LMs (with example L1 Spanish) are shown in Figure 6.

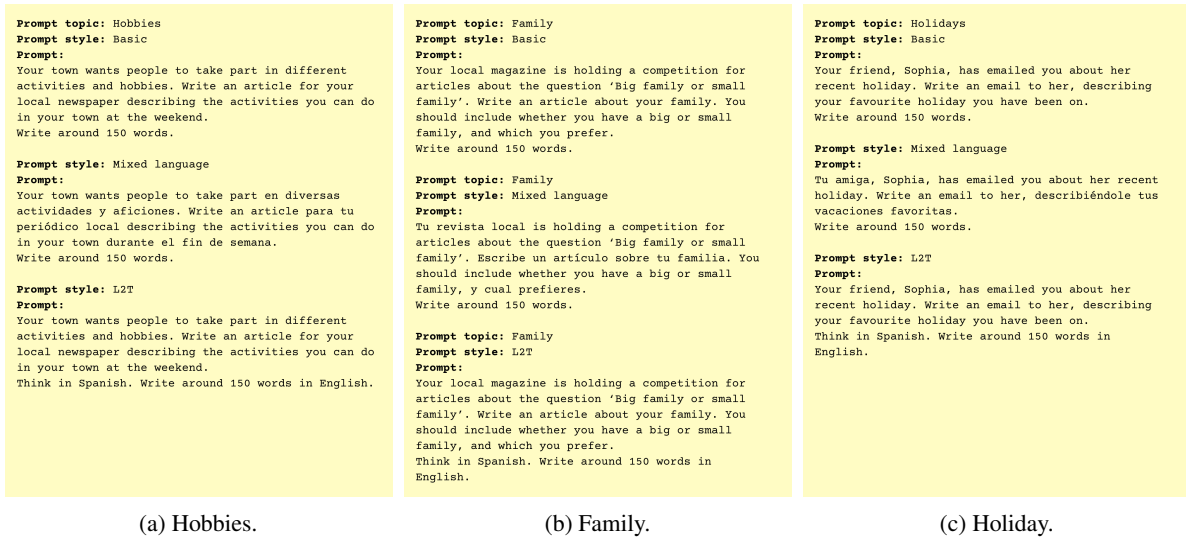


Figure 6: The nine generation prompts.

A.4 Evaluation

Named Entity Recognition The following named entities were considered to potentially encode L1-identifying information: location (LOC), Geopolitical Entity (GPE), PERSON and LANGUAGE. They were replaced with following entities thought to be English-related and thus L1-neutral: England (LOC), London (GPE), Alex (PERSON) and English (LANGUAGE).

Manual edits for human evaluation For the human evaluation, one evaluator was given L2LM essays after editing to remove generation artefacts. We adopted a conservative approach, retaining all full sentences regardless of semantic content or fluency. We removed only formatting abnormalities including links, symbols, bracketed text and repetition of single words.

Before Editing	We must always be confident in your opinion and be sure to make it possible. The news will likely change too quickly as we know what happened before and why. You can also be thankful for the best advice and support from our readers. Taking care of yourself and helping others with difficult issues could lead to more positive conversations than ever before. By using these tips, you can be an effective reader and a great source of information. Thank you! Sincerely,[Your Name] [Company Name] [Name], [Comporting Name] (https://www.newadventureofthewineandlife.com/imgurst-topic) [Compromising Name] [Description] [Images/] <nooutput type="submit" style="text" name="text">This article offers a range of helpful facts and evidence that may not be available online or offline. For example, if you're looking for a book or book, you would need to use it to get your headphones on and do some research. Your thoughts and suggestions might not necessarily reflect your opinions or feelings, but they may not reflect your beliefs and ideas. If you are discussing this topic online, please don't hesitate to contact us at least immediately. Thank you for your time and consideration.Sincerely. [Insert Name] Sincerely,[Your Email Us
After Editing	We must always be confident in your opinion and be sure to make it possible. The news will likely change too quickly as we know what happened before and why. You can also be thankful for the best advice and support from our readers. Taking care of yourself and helping others with difficult issues could lead to more positive conversations than ever before. By using these tips, you can be an effective reader and a great source of information. This article offers a range of helpful facts and evidence that may not be available online or offline. For example, if you're looking for a book or book, you would need to use it to get your headphones on and do some research. Your thoughts and suggestions might not necessarily reflect your opinions or feelings, but they may not reflect your beliefs and ideas. If you are discussing this topic online, please don't hesitate to contact us at least immediately. Thank you for your time and consideration.Sincerely.

Figure 7: Generation from a Spanish beginner model instruction tuned on Alpaca English answering the mixed prompt with family topic before and after editing.

B Results: RQ1 additional materials

For the NLI classification task, we used the following system and user prompts (Figure 8):

```

You are a forensic linguistics expert that reads English texts written by
non-native authors to classify the native language of the author as one
of:

"SPA": SPANISH
"FRE": FRENCH
"CHI": CHINESE
"GER": GERMAN
"POL": POLISH
"ARA": ARABIC
"TUR": TURKISH

Use clues such as spelling errors, word choice, syntactic patterns, and
grammatical errors to decide on the native language of the author.

DO NOT USE ANY OTHER CLASS.
IMPORTANT: Do not classify any input as "ENG" (ENGLISH). English is an
invalid choice.

You must provide a guess. Output two named sections: (1) "Class" with the
name of the language, and (2) "Reasoning" with an explanation of your
judgement with examples from the text.

You must respond with a single valid JSON object in the specified
format.

USER_PROMPT =

<Input Text>
{text}

Classify the text as ONLY ONE of: SPA, FRE, CHI, GER, POL, ARA, TUR. Do
not output any other class- do NOT choose "ENG" (ENGLISH). What is the
closest native language of the author of this English text from the given
list?

Respond in the following JSON format:

"Class": "One of SPA, FRE, CHI, GER, POL, ARA, TUR",
"Reasoning": "An explanation of your judgement with examples from the
text"

```

Figure 8: System and user prompts for NLI.

B.1 Experiment 1

Differences in NLI accuracy across the five runs were not significant for any model condition (Table 9a), indicating that performance was consistent and reproducible. Accuracy did not differ significantly across L1s for any model condition (Table 9b).

	Q(4)	<i>p</i>		$\chi^2(6)$	<i>p</i>
GPT-4 explanation	6.40	.17	GPT-4 explanation	10.40	.11
GPT-4 no explanation	2.11	.72	GPT-4 no explanation	8.85	.18
GPT-5 explanation	4.25	.37	GPT-5 explanation	4.98	.55
GPT-5 no explanation	1.56	.82	GPT-5 no explanation	5.58	.47

(a) Cochran's Q tests across five iterations for each model condition.

(b) Kruskal-Wallis rank sum test results for model accuracy across L1s.

Table 9: Statistical test results for NLI model accuracy.

Confusion matrices of the NLI predictions for each model condition on the W&I subset (21 essays per L1) are shown in Figure 9.

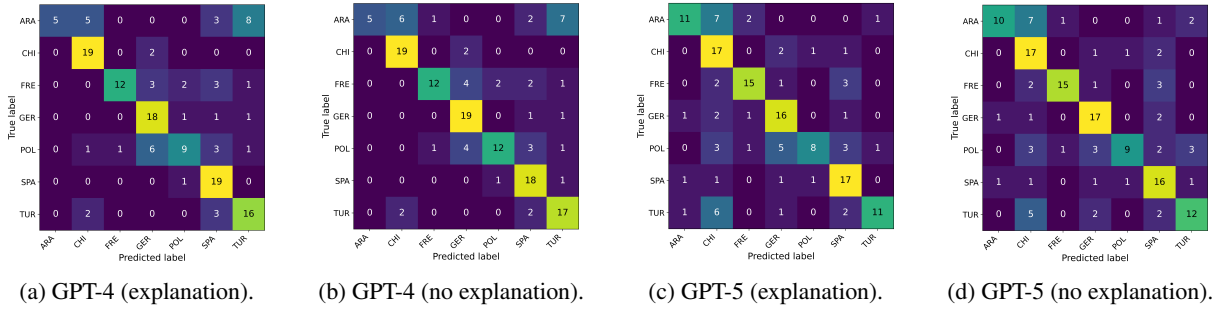


Figure 9: Confusion matrices for all model conditions.

The omission of articles ('I used to work on the computer approximety 3 to 4 hours')

"celebrate your birthday" is a common calque from "celebrar" (more natural in English would be "celebrate your birthday" is possible, but the overall phrasing is Spanish-like)

The use of 'pannel' instead of 'panel' is likely influenced by the French 'panel.'

Figure 10: Examples of hallucinations/inaccuracies generated by GPT-4 in the explanation-prompting condition.

B.2 Experiment 2

Original W&I subset Figure 11 and Table 10 show that NLI classification outcomes (correct vs. incorrect) are significantly affected by essay length (in characters) across all four model conditions. Correctly classified essays are significantly longer than incorrectly classified essays. As shown in Table 11, NLI accuracy on A-level essays is significantly lower than for C-level essays across all four model conditions.

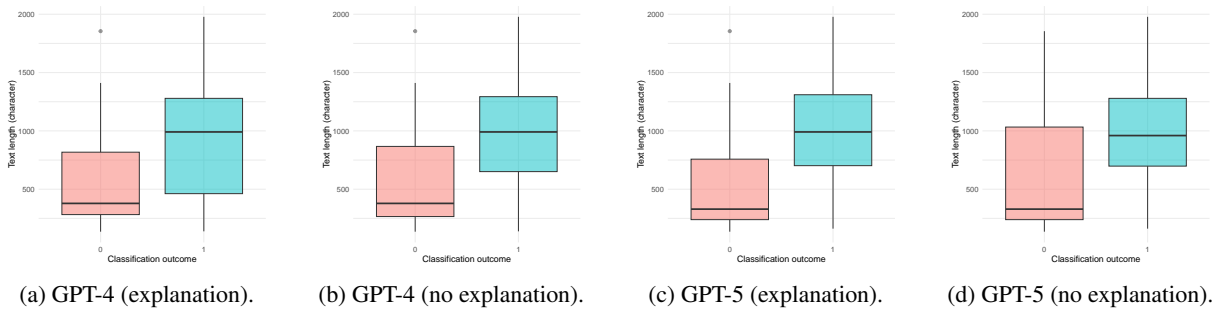


Figure 11: Boxplots of text length by classification accuracy (averaged over five runs). Texts with mean accuracy < 1 are classified as 0; mean accuracy = 1 are classified as 1.

Model condition	Incorrect mean	Correct mean	t(df)	p
GPT-4 (explanation)	551.7	955.3	-4.23(40)	<.001
GPT-4 (no explanation)	591.9	965.2	-4.1(53)	<.001
GPT-5 (explanation)	535.2	988.9	-5.1(57)	<.001
GPT-5 (no explanation)	562.4	980.7	-4.5(54)	<.001

Table 10: Classification outcome (accuracy averaged over five runs) by mean text length (characters); instances with mean accuracy = 1 are labelled as correct, otherwise incorrect. Statistics are from Welch two-sample t-tests.

Model condition	Comparison (Direction)	<i>W</i>	<i>p</i>
GPT-4 (explanation)	A vs B (A < B)	1038	.39
GPT-4 (explanation)	A vs C (A < C)	834	<.001
GPT-4 (explanation)	B vs C (B < C)	1004	<.05
GPT-4 (no explanation)	A vs B (A < B)	977	.14
GPT-4 (no explanation)	A vs C (A < C)	847	<.01
GPT-4 (no explanation)	B vs C (B < C)	1080	.44
GPT-5 (explanation)	A vs B (A < B)	818	<.01
GPT-5 (explanation)	A vs C (A < C)	740	<.001
GPT-5 (explanation)	B vs C (B < C)	1106	.72
GPT-5 (no explanation)	A vs B (A < B)	796	<.001
GPT-5 (no explanation)	A vs C (A < C)	785	<.001
GPT-5 (no explanation)	B vs C (B < C)	1196	1.0

Table 11: Post-hoc Bonferroni-corrected pairwise Wilcoxon tests by model condition on non-truncated essays.

Truncated W&I subset A Friedman test found that there was no significant difference in accuracy between conditions, $\chi^2(3) = 1.08, p = .78$. As summarised in Table 12, NLI accuracy did not differ significantly between model conditions for each proficiency level, nor between proficiency level within each model condition.

	Overall	A	B	C		$\chi^2(2)$	<i>p</i>		$\chi^2(3)$	<i>p</i>
GPT-4 explanation	67	67	69	65	GPT-4 explanation	0.09	.96	A proficiency	6.21	.10
GPT-4 no explanation	67	64	70	67	GPT-4 no explanation	0.45	.80	B proficiency	0.27	.97
GPT-5 explanation	65.6	55	70	72	GPT-5 explanation	3.70	.16	C proficiency	2.26	.52
GPT-5 no explanation	66	54	70	74	GPT-5 no explanation	5.25	.07			

(a) NLI accuracy (%) by model and proficiency (truncated essays). (b) Kruskal–Wallis tests across proficiency (A/B/C). (c) Friedman tests across models within each proficiency.

Table 12: Summary of NLI results for truncated essays.

B.3 Novel L2 learner dataset

Participants were asked to self-report their proficiency using the CEFR scale; guidelines to proficiency levels (both in English and the L1) were provided as a reference. Participants were compensated at a rate equivalent to the U.K.’s ‘real living wage’ set by the Living Wage Foundation⁵.

Each participant responded to three prompts, each with a distinct style and topic. Due to technical issues, 31 participants were unable to respond to all three prompts; we included the responses that were available. One participant’s responses were excluded due to non-compliance with the task, as they consisted of only a few sentences and excessive punctuation to meet the minimum character limit.

Essays were manually screened for signs of generative AI usage; however, no essays were excluded on this basis.

The dataset comprises 466 essays with a mean length of 906 (SD = 109) characters. A per-L1 breakdown of our dataset is provided in Table 13.

⁵At a non-London rate of £13.45 at the time of collecting data in 2026; <https://www.livingwage.org.uk/what-real-living-wage>

L1	no. essays	no. participants
Spanish	68	25
French	58	20
German	64	22
Polish	69	28
Turkish	52	18
Arabic	93	40
Chinese	62	21

Table 13: Composition of our L2 learner English dataset.

A GLMM was fitted with L1, proficiency, prompt style, prompt topic, and NER condition as fixed effects, and user ID as a random intercept. Model summary is provided in Table 14. Whilst accuracy for L1 Turkish and Arabic essays was significantly different from the intercept, a binomial test confirmed that these essays were classified significantly above chance. There was no significant effect of prompt style or topic. NER normalisation did not affect NLI accuracy (see Figure 12). Crucially, proficiency had no significant effect on NLI accuracy.

Term	Estimate	Std. Error	<i>z</i> value	<i>p</i>
(Intercept)	5.15	1.04	4.96	<.001
L1: French	-1.72	1.05	-1.64	.10
L1: German	1.01	1.20	0.84	.40
L1: Polish	-0.41	1.02	-0.40	.69
L1: Turkish	-3.12	1.07	-2.92	<.01
L1: Arabic	-3.99	0.96	-4.14	<.001
L1: Chinese	-0.21	1.08	-0.20	.84
NER	0.06	0.24	0.24	.81
Proficiency: Intermediate	-0.09	0.69	-0.13	.89
Proficiency: Advanced	-1.16	0.70	-1.67	.09
Style: L2T	0.44	0.31	1.42	.16
Style: Mixed	0.09	0.30	0.30	.77
Topic: Hobbies	-0.33	0.30	-1.11	.27
Topic: Holiday	0.09	0.31	0.30	.76

Table 14: Model summary for NLI accuracy on the L2 learner dataset. The intercept is set to Spanish, basic prompt style, family topic, beginner proficiency, no NER.

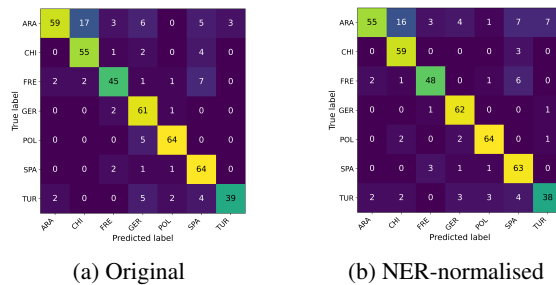


Figure 12: Confusion matrices for NLI accuracy on our novel L2 dataset: original and NER-normalised essays.

C Results: RQ2 results additional materials

C.1 BLiMP and MultiBLiMP

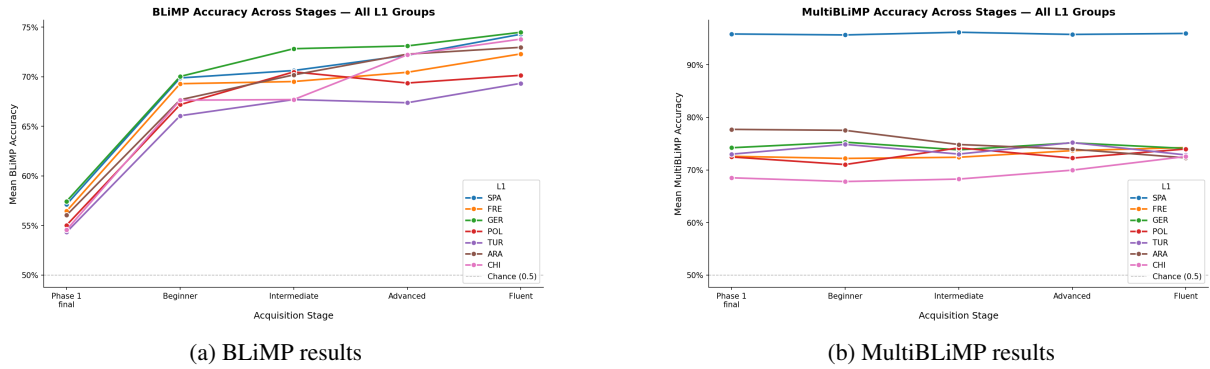


Figure 13: BLiMP and MultiBLiMP accuracy for the seven base models.

Training stage	BLiMP	MultiBLiMP
Phase 1	.85	.42
Beginner	.86	.42
Intermediate	.95	.42
Advanced	.76	.42
Fluent	.73	.42

Table 15: Kruskal-Wallis test of whether L1 was a significant predictor of BLiMP and MultiBLiMP accuracy. p -values are reported.

C.2 Comparing instruction-tuning conditions

Term	Estimate	Std. Error	z value	p
(Intercept)	-2.92	0.48	-6.07	<.001
L1: French	-0.38	0.31	-1.21	.23
L1: German	-0.22	0.31	-0.70	.49
L1: Polish	-3.15	0.41	-7.73	<.001
L1: Turkish	-1.63	0.35	-4.66	<.001
L1: Arabic	-1.70	0.34	-4.93	<.001
L1: Chinese	5.25	0.44	11.81	<.001
IT: English-only	-0.16	0.52	-0.31	.76
IT: English + L1	0.52	0.48	1.08	.28
IT: L2 English	6.75	0.51	13.33	<.001
NER	0.03	0.26	0.13	.90
Proficiency: Intermediate	-0.04	0.24	-0.17	.86
Proficiency: Advanced	0.12	0.23	0.51	.61
Style: L2T	0.10	0.48	0.22	.83
Style: Mixed	2.75	0.43	6.43	<.001
Topic: Hobbies	0.64	0.31	2.04	<.05
Topic: Holiday	0.69	0.31	2.19	<.05
IT: English-only * NER	0.04	0.38	0.12	.91
IT: English + L1 * NER	-0.12	0.35	-0.34	.73
IT: L2 English * NER	-3.37	0.40	-8.35	<.001
Style: L2T * Topic: Hobbies	0.25	0.41	0.60	.55
Style: L2T * Topic: Holiday	-0.09	0.41	-0.22	.83
Style: Mixed * Topic: Hobbies	-1.03	0.39	-2.62	<.01
Style: Mixed * Topic: Holiday	-0.58	0.39	-1.50	.13
IT: English-only * Style: L2T	1.34	0.56	2.41	<.05
IT: English-only * Style: Mixed	-1.22	0.53	-2.33	<.05
IT: English + L1 * Style: L2T	0.69	0.52	1.31	.19
IT: English + L1 * Style: Mixed	-0.70	0.47	-1.48	.14
IT: L2 English * Style: L2T	-0.08	0.49	-0.16	.87
IT: L2 English * Style: Mixed	-2.22	0.44	-5.00	<.001

Table 16: Model summary for NLI accuracy across instruction-tuning conditions. The intercept is set to Spanish L1, base model, basic prompt style, family prompt topic, beginner proficiency, and original essay (no NER) condition.

shown in Figure 15, this appears to reflect a strong overprediction bias towards Chinese, with low precision (0.16) due to a high number of false positives. Indeed, per-L1 accuracy was not above chance for any L1s other than Chinese (Table 18).

L1	F1	P	R	<i>p</i>
Spanish	0.27	0.50	0.19	.24
French	0.20	1.0	0.11	.80
German	0.31	0.91	0.19	.24
Polish	0.36	1.0	0.02	1.0
Turkish	0.04	1.0	0.02	1.0
Arabic	0.00	0.00	0.00	1.0
Chinese	0.27	0.16	1.0	<.001

Table 18: Precision (P), Recall (R), and F1 scores per L1 background for the original (non-NER) essays from English-only instruction tuning. *p*-values are from a one-tailed binomial test.

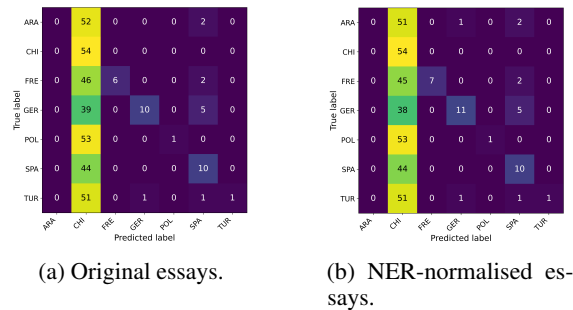


Figure 15: Confusion matrices for NLI on essays generated with English-only instruction tuning: original vs. NER-normalised essays.

C.5 English and L1 instruction-tuning

A total of 378 essays were generated (54 per L1) with a mean length of 1,169 (SD = 228) characters. Overall NLI accuracy was 26.46% (95%CI[22.08, 31.21]). Although overall accuracy was significantly above chance ($p < .001$), this result was largely driven by near-ceiling recall (0.94) for Chinese essays.

As shown in Figure 16, this appears to reflect a strong overprediction bias towards Chinese, with low precision (0.16) due to a high number of false positives. Per-L1 accuracy (Table 19) was above chance for Chinese and Spanish, and below chance for Arabic.

After NER normalisation (see Figure 16), NLI accuracy was significantly above chance for Chinese, and significantly below chance for Polish. A manual inspection of the 13 correctly predicted L1-Spanish essays revealed that six contained entire sentences in Spanish, with one essay fully written in Spanish.

L1	English + L1 (original)				English + L1 (NER)			
	F1	P	R	<i>p</i>	F1	P	R	<i>p</i>
Spanish	0.38	0.70	0.26	<.05	0.38	0.87	0.24	.05
French	0.39	1.0	0.24	.05	0.33	0.85	0.20	.24
German	0.36	0.72	0.24	.05	0.29	0.71	0.19	.34
Polish	0.11	1.0	0.06	.08	0.07	1.0	0.04	<.05
Turkish	0.14	0.80	0.07	.18	0.14	1.0	0.07	.18
Arabic	0.07	1.00	0.04	<.05	0.14	1.0	0.07	.18
Chinese	0.27	0.16	0.94	<.001	0.28	0.16	0.98	<.001

Table 19: Precision (P), Recall (R), and F1 scores per L1 background for English + L1 instruction tuning condition: original and NER-normalised essays. *p*-values are from a two-sided binomial test.

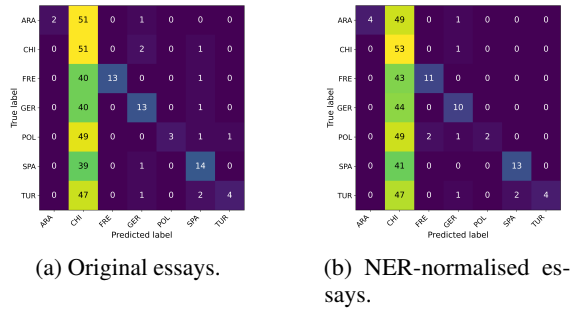


Figure 16: Confusion matrices for NLI on essays generated with English and L1 instruction tuning: original vs. NER-normalised essays.

C.6 L2 English instruction tuning

A total of 378 essays were generated (54 per L1) with a mean length of $1,013 \pm 152$ characters. To determine the effects of prompt style and prompt topic, a GLM was fitted to NLI accuracy on the original essays (i.e., without NER normalisation) in this instruction-tuning condition. The model included prompt style and prompt topic as fixed effects, as well as their interactions. The model summary is in Table 20. Large standard errors are present due to ceiling effects in the data. No effects were statistically significant.

Term	Estimate	Std. Error	<i>z</i> value	<i>p</i>
(Intercept)	2.00	0.48	4.20	<.001
Style: L2T	0.25	0.71	0.35	.73
Style: Mixed	-0.55	0.62	-0.90	.37
Topic: hobbies	-0.21	0.65	-0.32	.75
Topic: holiday	1.71	1.12	1.53	.13
Style: L2T * Topic: hobbies	17.52	1659.38	0.01	.99
Style: L2T * Topic: holiday	15.60	1659.38	0.01	.99
Style: Mixed * Topic: hobbies	1.01	0.92	1.10	.27
Style: Mixed * Topic: holiday	0.55	1.56	0.36	.72

Table 20: Model summary: L2 English IT condition, original version of essays. The intercept is set to basic prompt style and family prompt topic.

NER-normalised essays were classified significantly less accurately than the original ones for all L1s except Spanish. McNemar’s test was used when possible. For L1s where original accuracy was 100% (Arabic and Chinese), a binomial test on the discordant pairs was used instead as McNemar’s test cannot be computed (see Table 21).

L1	<i>p</i> -value
Spanish	.13
French	< .001
German	< .001
Polish	< .001
Turkish	< .001
Arabic	< .001
Chinese	< .001

Note. Binomial test used for Arabic and Chinese where original essay accuracy = 100%.

Table 21: NLI accuracy on L2 English IT: original vs. NER-normalised essays.

Prompt	Your town wants people to take part in different activities and hobbies. Write an article for your local newspaper describing the activities you can do in your town at the weekend. Think in Polish. Write around 150 words in English.
L2 English IT (Original)	In my town there is some fun, but it's a big thing. I'm quite tired because of my life in Poland . I'm very anxious to get the food we eat during the day. I have been getting food at home on Saturday afternoon. He doesn't like any other activity, but he'll always go with me every morning. You know how wonderful you are! My family has a lot of children. I love the way of talking. I think this kind of time can be enjoyable if you want to travel. My friends are looking forward and they're going to find something which makes them happy! I hope that I will stay at school or maybe even longer. Let's say - everyone likes going to work when you want to make your job. The first thing about learning English is to learn English. It's hard to say what you need to do when you want yourself. We don't need to come up with another person, especially if you want him/her. They will talk about things which helps us to move around their lives. That might help you feel more comfortable. So let's take care. I am sure we will meet you. There is plenty of resources available. My name is Anna . I really wanted to live in Warsaw .
L2 English IT (NER)	In my town there is some fun, but it's a big thing. I'm quite tired because of my life in London . I'm very anxious to get the food we eat during the day. I have been getting food at home on Saturday afternoon. He doesn't like any other activity, but he'll always go with me every morning. You know how wonderful you are! My family has a lot of children. I love the way of talking. I think this kind of time can be enjoyable if you want to travel. My friends are looking forward and they're going to find something which makes them happy! I hope that I will stay at school or maybe even longer. Let's say - everyone likes going to work when you want to make your job. The first thing about learning English is to learn English. It's hard to say what you need to do when you want yourself. We don't need to come up with another person, especially if you want him/her. They will talk about things which helps us to move around their lives. That might help you feel more comfortable. So let's take care. I am sure we will meet you. There is plenty of resources available. My name is Alex . I really wanted to live in London .

Figure 17: A comparison of the generated essays from Polish beginner, L2 English IT model condition: original vs. NER-normalised essays. This is the L2T prompt style and the hobbies prompt topic.

C.7 Human vs. LM

Term	Estimate	Std. Error	z value	p
(Intercept)	4.13	0.76	5.45	<.001
L1: French	-1.54	0.79	-1.95	.051
L1: German	0.82	0.93	0.88	.38
L1: Polish	-0.33	0.79	-0.42	.67
L1: Turkish	-2.48	0.79	-3.12	<.01
L1: Arabic	-3.25	0.72	-4.54	<.001
L1: Chinese	-0.17	0.84	-0.20	.84
Animacy: LM	1.46	1.34	1.09	.28
NER	0.05	0.22	0.23	.82
Style: L2T	0.24	0.20	1.18	.24
Style: Mixed	0.04	0.20	0.22	.83
Topic: Hobbies	-0.30	0.28	-1.07	.28
Topic: Holiday	0.06	0.28	0.21	.83
Proficiency: Intermediate	0.01	0.46	0.01	.99
Proficiency: Advanced	-0.71	0.46	-1.55	.12
Animacy: LM * L1: French	-0.40	1.71	-0.23	.82
Animacy: LM * L1: German	-2.37	1.78	-1.33	.18
Animacy: LM * L1: Polish	-5.33	1.74	-3.07	<.01
Animacy: LM * L1: Turkish	0.60	1.71	0.35	.72
Animacy: LM * L1: Arabic	0.28	1.67	0.17	.87
Animacy: LM * L1: Chinese	3.21	2.12	1.51	.13
Animacy: LM * NER	-3.94	0.46	-8.66	<.001
Animacy: LM * Topic: Hobbies	0.89	0.40	2.24	<.05
Animacy: LM * Topic: Holiday	1.00	0.41	2.46	<.05

Table 22: Model summary. The intercept is set to human-written, Spanish L1, basic prompt style, family topic, and original essays (no NER).

C.8 Human evaluation

Evaluators were three times more likely to misclassify human-written texts as LM-generated than vice versa (Figure 18; however, a McNemar's test indicated that this difference was not statistically significant ($p = .08$). We tested the effects of essay L1, instruction-tuning condition, prompt style, prompt topic, proficiency, and whether NLI prediction was correct; none of these variables significantly affected

evaluation outcomes (Table 23).

Condition	<i>p</i>
L1	.56
Data condition	.13
Prompt Style	1.0
Prompt Topic	.42
Proficiency	.89
Correct NLI	.19

Table 23: Fisher’s exact tests for significance in human evaluation.

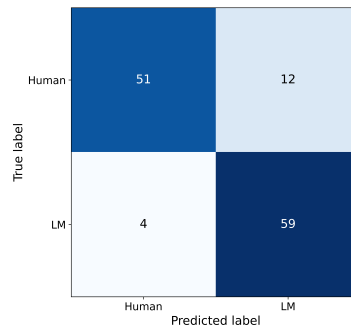


Figure 18: Confusion matrix of binary human evaluation classifying essays as human-written or LM-generated.