Report on Comparing human and LLM proofreading in L2 writing

By Hakyung Sung, Karla Csuros, and Min-Chang Sung

Intro Summary

“This study examines the lexical and syntactic interventions of human and LLM proofreading aimed at improving overall intelligibility in identical second language writings, and evaluates the consistency of outcomes across three LLMs”

Background

How does the rise of LLMs affect L2 learning?

With LLMs on the rise we can see their usefulness in L2 learning

- Classrooms
- Independent study

Is it really better than the rest of the tools?

Proofreading = correcting surface errors or improving clarity in written texts.

- Human proofreading shows large variation:
  - Harwood (2018): 14 proofreaders made 113–472 edits to the same essay.
  - Edits range from helpful to introducing new errors.
  - Influenced by fatigue, focus, and subjective judgment.
- LLM proofreading = a new frontier in L2 writing:
  - Builds on earlier automated written corrective feedback (WCF) tools
  - Integrated in prewriting, post writing, and revision stages
  - Studies show LLMs handle grammar well but struggle with context, idioms, and cultural awareness.

Previous proofreading tools

- Grammatical Error Correction systems (GECs)
  - Automated spelling correction
  - Grammar assisted
- Human tutors
  - Not automated
  - Hard to access

LLMs proofreading

- **Accessible**
- **Instant**
- **Spell checker**
- **Can overwrite words losing context**
- **Creates new text**

Human vs. LLM Proofreading

Human Proofreading
Restructures ideas for clarity; may be lenient.
Inconsistent (large variation across editors).
Understands tone and purpose.

LLM Proofreading
Rewrites whole sentences; aims for fluency.
Consistent but may misread context.
Lacks cultural & situational awareness.

Previous research = focused mainly on grammar correction and minimal edits.

Little focus on:

- Broader lexical sophistication and syntactic complexity.
- Comparing multiple LLMs

Unclear whether observed improvements:

- Are unique to a specific model
- Generalizable across other LLMs.

Research Questions

- How do humans and LLMs differ in lexical features of L2 writing?
- How do they differ in syntactic features?
- Are LLM outputs consistent across models?

Methodology

- The participants were college students learning English in ten regions
  - Japan (JPN)
  - Korea (KOR)
  - China (CHN)
  - Taiwan (TWN)
  - Indonesia (IDN)
  - Thailand (THA)
  - Hong Kong (HKG)
  - the Philippines (PHL)
  - Pakistan (PAK)
  - Singapore (SIN)

Region	A2_0	B1_1	B1_2	B2_0	Total
JPN	10	10	10	10	40
KOR	10	10	10	10	40
CHN	10	10	10	10	40
TWN	10	10	10	10	40
IDN	10	10	10	3	33
THA	10	10	10	2	32
HKG	–	10	10	10	30
PHL	–	10	10	10	30
PAK	–	10	10	3	23
SIN	–	–	10	10	20
Total	60	90	100	78	328

Writing Prompts

1) “It is important for college students to have a part time job”
2) “Smoking should be completely banned at all restaurants”

All five proofreaders revised the same eight essays. Edited tokens ranged from 40.00 – 59.63 (≈ 41% difference)

ID	Age	Sex	Degree	Experience (years)	L1 English
A	28	Female	BA	3	Canadian
B	32	Female	MS	5	Australian
C	27	Female	BS	3	American
D	38	Female	BS	10	British
E	31	Female	PhD	2	Australian

Why this set?

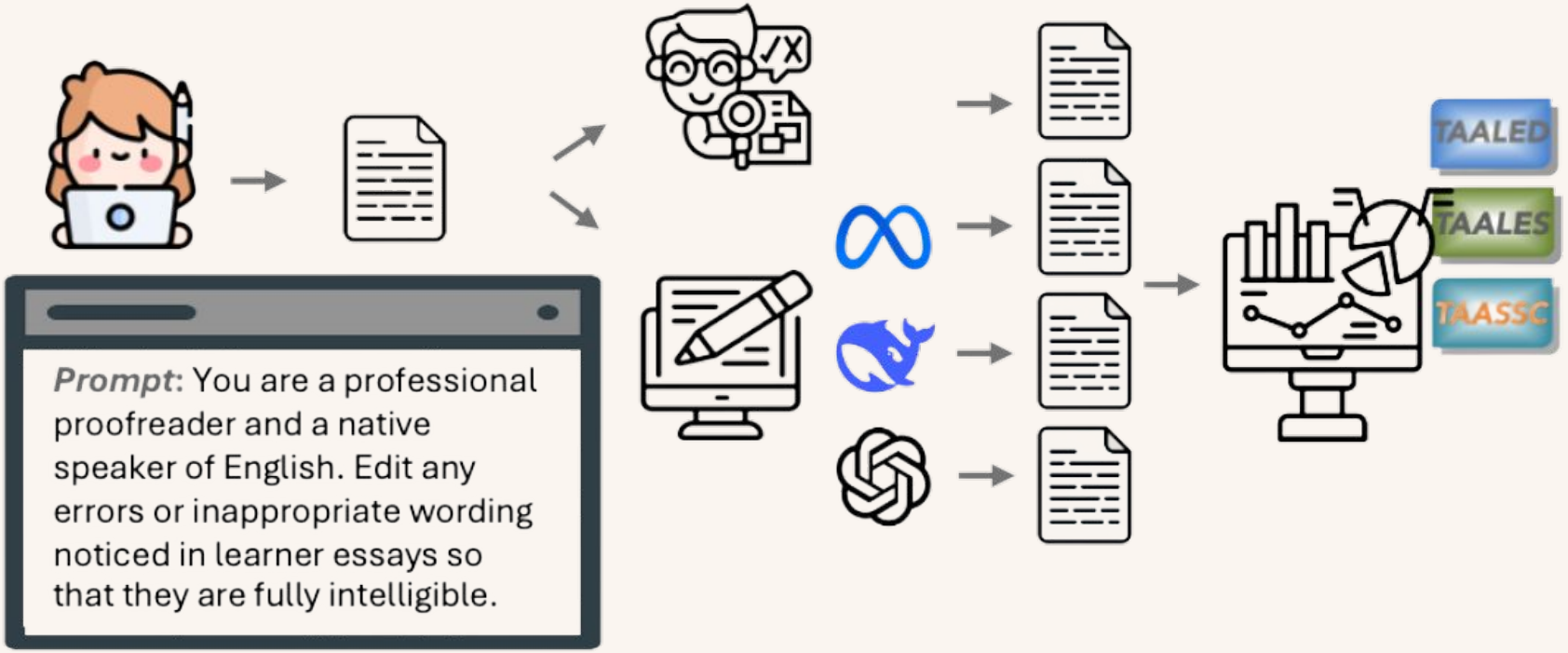
- Provides paired original and professionally proofread versions.
- Proficiency labels.
- Balanced regional coverage across ten regions.

Models Used

Chatgpt-4o - Common but hard to specialize.

Llama3.1-8b - Open model that is lighter to run.

Deepseek-r1-8b - Open model that is lighter to run.


Prompt: You are a professional proofreader and a native speaker of English. Edit any errors or inappropriate wording noticed in learner essays so that they are fully intelligible.
TAALED
TAALES
TAASSC

Syntactic Feature Levels

- Traditional methods use average T-unit length to infer complexity (Lu 2010, 2011).
- Fine-grained indices (Kyle & Crossley 2018) separate:
  - Clausal level: Nominal subjects per clause → connected ideas.
  - Phrasal level: Dependents per nominal → detail inside clauses.
  - Morphosyntactic level: Verb forms → tense/aspect variety.

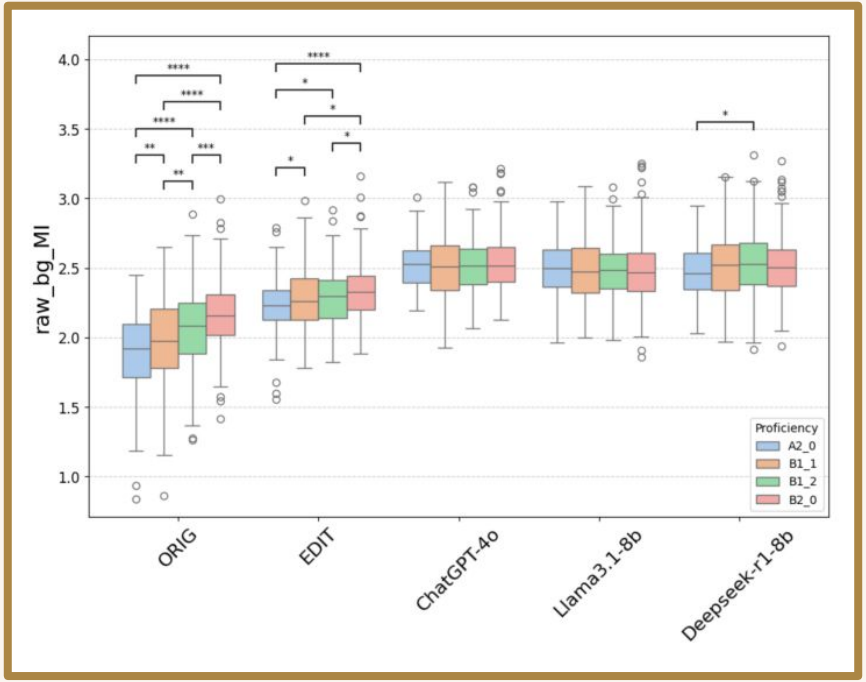
A T-unit = main + subordinate clauses (e.g., "I went to the store because I needed milk").

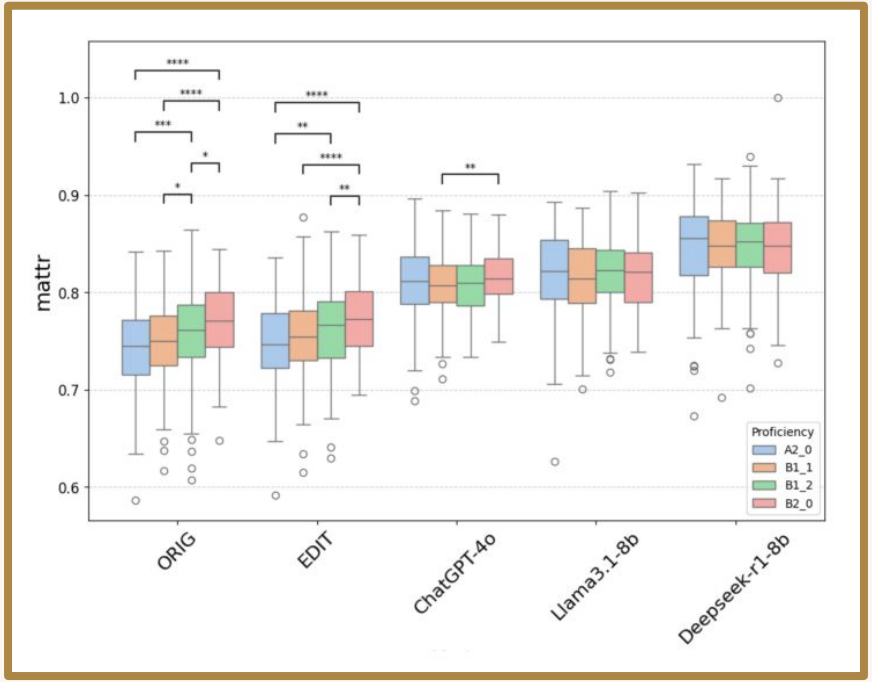
Results

Lexical Indices Summary

Index	EDIT	ChatGPT-4o	Llama3.1-8b	Deepseek-r1-8b
raw_bg_MI	+0.35 / 1.80***	+0.65 / 3.30***	+0.62 / 3.17***	+0.60 / 3.03***
usf	-1.37 / 0.15	-9.21 / 0.99***	-8.48 / 0.91***	-12.09 / 1.30***
b_concreteness	+0.00 / 0.02	-0.15 / 0.83***	-0.12 / 0.67***	-0.21 / 1.11***
cw_lemma_freq_log	-0.02 / 0.03	-0.30 / 0.54***	-0.26 / 0.47***	-0.37 / 0.67***
mattr	+0.01 / 0.18	+0.07 / 2.20***	+0.08 / 2.63***	+0.10 / 3.41***
ntypes	+0.63 / 0.05	+19.98 / 1.68***	+16.68 / 1.40***	+16.80 / 1.41***

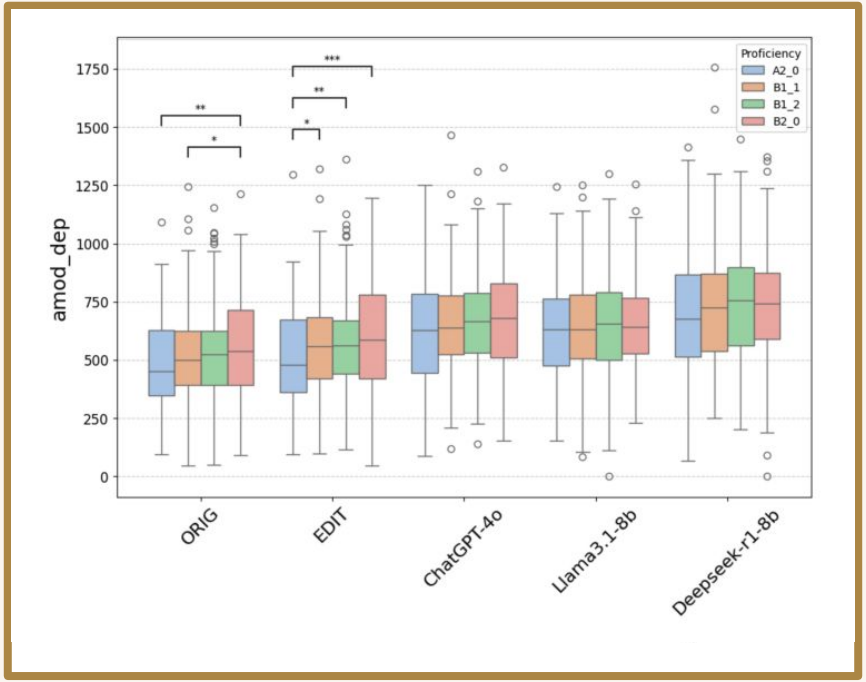
The LLM edits shift towards less common words
The LLM edits may seem less formal
Human edits did not shift towards more diverse vocabulary

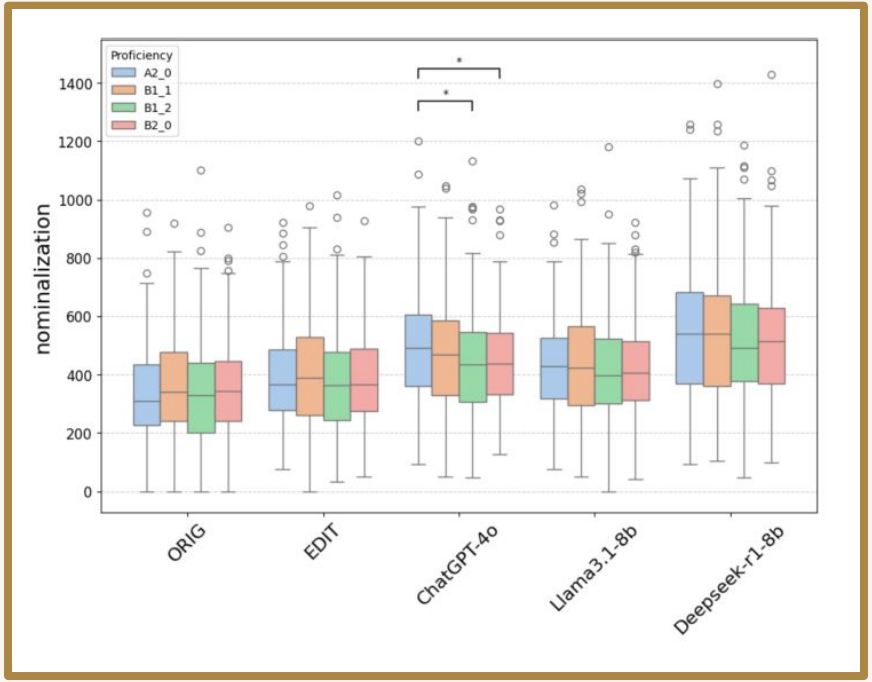

raw_bg_MI
4.0
3.5
3.0
2.5
2.0
1.5
1.0
ORIG
EDIT
ChatGPT-4o
Llama3.1-8b
Deepseek-r1-8b
Proficiency
A2_0
B1_1
B1_2
B2_0


mattr
1.0
0.9
0.8
0.7
0.6
ORIG
EDIT
ChatGPT-4o
Llama3.1-8b
Deepseek-r1-8b
Proficiency
A2_0
B1_1
B1_2
B2_0

Syntactic Indices Summary

Index	EDIT	ChatGPT-4o	Llama3.1-8b	Deepseek-r1-8b
mltu	-115.49 / 0.31	-105.73 / 0.28	+44.26 / 0.12	+118.42 / 0.31
all_clauses	+15.55 / 0.10	+133.76 / 0.84***	+99.12 / 0.62***	+179.00 / 1.12***
nonfinite_prop	-1.33 / 0.29	+2.01 / 0.44***	+2.63 / 0.57***	+5.52 / 1.20***
np	-21.30 / 0.08	+91.96 / 0.36**	+41.27 / 0.16	+194.91 / 0.76***
np_deps	-35.03 / 0.08	+79.21 / 0.17	+91.91 / 0.20	+217.81 / 0.47**
amod_dep	+17.54 / 0.01	+137.65 / 0.75***	+127.44 / 0.70***	+204.54 / 1.12***
nominalization	+58.12 / 0.40**	+152.04 / 1.05***	+102.85 / 0.71***	+213.63 / 1.47***
be_mv	+10.37 / 0.12	-56.53 / 0.63***	-41.60 / 0.47**	-84.02 / 0.94***
past_tense	-15.80 / 0.29	-17.38 / 0.32	-17.77 / 0.32	-19.31 / 0.35**


amod_dep
Proficiency
A2_0
B1_1
B1_2
B2_0
ORIG
EDIT
ChatGPT-4o
Llama3.1-8b
Deepseek-r1-8b


nominalization
Proficiency
A2_0
B1_1
B1_2
B2_0
ORIG
EDIT
ChatGPT-4o
Llama3.1-8b
Deepseek-r1-8b

Cross-Model Consistency

Pair	Lexical	Syntax
ChatGPT-4o – Llama3.1-8b	0.70	0.62
ChatGPT-4o – Deepseek-r1-8b	0.60	0.53
Llama3.1-8b – Deepseek-r1-8b	0.56	0.65

Cronbach's α = 0.83 (lexical) and 0.81 (syntactic) → strong internal consistency
Indicates all three models produced very similar results
Confirms proofreading behavior is not model-specific

Analysis

Both human and LLM proofreading improved vocabulary sequencing (lexical cohesion).

LLMs also boosted diversity and sophistication—sometimes erasing proficiency differences.

Example:

"I often can smell" → "I often catch a whiff."

Conclusion

Key Implications

- More attention should go to how to use LLM proofreading effectively, not which LLM to choose.
- LLMs are consistent but miss cultural/contextual nuance.
- Humans bring subjectivity and flexibility; together they complement each other.

Limitations:

- **Focused on Asian college learners**
- **Only English argumentative writing**
- **Doesn’t capture user perception of edits**

This paper is notably objective and well-designed:

- **Multiple models ensure transferability**
- **Balanced dataset and quantitative rigor**
- **Highlights human variability without bias**

Final takeaway:

- **AI acts as a writing partner, not a replacement.**
- **It improves structure and variety but risks erasing individuality.**

QUIZ

This was the in Class link so I added slides with the questions

Multiple Choice

In the tests which is NOT a problem that the LLMs had in proofreading L2 writing?

A. spelling errors

B. keeping context relevant and organization of content

C. cultural idioms / slangs

D. overcorrection (of already valid writing)

TAALED is "Tool for Automatic Analysis of ____"

A. lexical diversity (how varied vocabulary is)

B. lexical sophistication (how advanced vocabulary is)

C. syntax and complexity (sentence structure complexity)

D. lexical range (amount of lexical categories)

Why did the authors decide to use multiple LLMs?

A. To see if proofreading behavior is consistent across models

B. To improve model performance

C. To compare training data quality

D. To reduce processing time

Short answer question

What is the difference between LLM proofreading and Human proofreading in L2 writing?

Thank You

Class Name

Student Name