Evaluating Prompting Strategies for Grammatical Error Correction Based on Language Proficiency

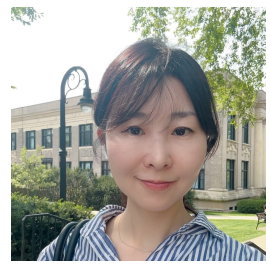
Min Zeng[†] Jiexin Kuang[†] Mengyang Qiu[‡]
Jayoung Song[¶] Jungyeul Park[†]

[†]Department of Linguistics, The University of British Columbia, Canada
[‡]Department of Psychology, Trent University, Canada
[¶]Department of Asian Studies, Pennsylvania State University, USA
{minzengz,jxkuang}@student.ubc.ca mengyangqiu@trentu.ca
jayoung.song@psu.edu jungyeul@mail.ubc.ca

LREC-COLING 2024
20-25 May, 2024

Authors




Min Zeng* Jiexin Kuang* Mengyang Qiu Jayoung Song Jungyeul Park

*Equally contributed authors.

- This paper proposes an analysis of prompting strategies for grammatical error correction (GEC) with selected large language models (LLM) based on language proficiency. GEC using generative LLMs has been known for overcorrection where results obtain higher recall measures than precision measures.
- Our method focuses on zero-shot and few-shot prompting and fine-tuning models for GEC for learners of English as a foreign language based on the different proficiency.
- We investigate GEC results and find that overcorrection happens primarily in advanced language learners' writing (proficiency C) rather than proficiency A (a beginner level) and proficiency B (an intermediate level).
- Fine-tuned LLMs, and even few-shot prompting with writing examples of English learners, actually tend to exhibit decreased recall measures.
- To make our claim concrete, we conduct a comprehensive examination of GEC outcomes and their evaluation results based on language proficiency.

Language Proficiency

For prompting GEC using GPTs, we use the Cambridge English Write & Improve (W&I) corpus, which is manually annotated with CEFR proficiency levels, consisting of beginner level A, intermediate level B, and advanced level C.

(1) *in addition more and more scientists agree with alien really exist

(2) In addition, more and more scientists agree that aliens really exist.

Proficiency A		Proficiency B		Proficiency C	
M:PUNCT	0.0933	M:PUNCT	0.1134	M:PUNCT	0.1183
R:ORTH	0.0602	R:PREP	0.0589	R:PREP	0.0517
R:PREP	0.0506	M:DET	0.0442	M:DET	0.0345
R:VERB:TENSE	0.0455	R:VERB	0.0414	R:VERB	0.0323
R:VERB	0.0419	R:VERB:TENSE	0.0393	R:VERB:TENSE	0.0273

Table 1: Most frequent errors and their ratio in W&I

Experimental Results

- For experiments, we use the development data set of W&I from BEA2019, which distinguishes language proficiency levels into A, B and C.
- We follow the experimental setting described in Suzgun et al. (2022) for GPT-2 (`gpt2-xl`) inferences, and we also adapt it to GPT-3.5 (`text-davinci-003`).

model	`gpt2-xl`
tokenizer	`gpt2-xl`
num_examplars	0-4 shots
max_model_token_length	256 if num_examplars is 0 else 512
delimiter left and right	{ }

- To evaluate the performance of language proficiency levels A, B, and C, we report ERRANT results (Bryant et al., 2017) as metrics that include true positive, false positive, false negative, precision, recall, and more importantly, F0.5 scores which emphasize precision than recall.

We used the prompts described in Table 2:

1-shot	ungrammatical	This is important thing.
	grammatical	This is an important thing.
2-shot	ungrammatical	Water is needed for alive.
	grammatical	Water is necessary to live.
3-shot	ungrammatical	And young people spend time more ther lifestile.
	grammatical	And young people spend more time on their lifestyles.
4-shot	ungrammatical	Both of these men have dealed with situations in an unconventional manner and the results are with everyone to see.
	grammatical	Both of these men have dealt with situations in an unconventional manner and the results are plain to see.

Table 2: Prompt examples

We used the following setting for fine-tuning parameters:

epochs	5
using masked language modeling	False
block size (train)	128
per_device_train_batch_size	4
save_steps	10000
save_total_limit	2

		A						B						C						all					
		TP	FP	FN	Prec	Rec	F0.5	TP	FP	FN	Prec	Rec	F0.5	TP	FP	FN	Prec	Rec	F0.5	TP	FP	FN	Prec	Rec	F0.5
GPT-2	zero-shot	70	3944	2878	0.0174	0.0237	0.0184	45	5204	2453	0.0086	0.018	0.0096	28	4860	1058	0.0057	0.0258	0.0068	143	14008	6389	0.0101	0.0219	0.0113
	1-shot	86	3447	2862	0.0243	0.0292	0.0252	58	4240	2440	0.0135	0.0232	0.0147	28	3730	1058	0.0075	0.0258	0.0087	172	11417	6360	0.0148	0.0263	0.0163
	2-shot	103	4175	2845	0.0241	0.0349	0.0257	69	5442	2429	0.0125	0.0276	0.0141	30	4905	1056	0.0061	0.0276	0.0072	202	14522	6330	0.0137	0.0309	0.0154
	3-shot	140	4445	2808	0.0305	0.0475	**0.0329**	95	5710	2403	0.0164	0.038	**0.0185**	38	4979	1048	0.0076	0.035	**0.009**	273	15134	6259	0.0177	0.0418	**0.02**
	4-shot	133	4347	2815	0.0297	0.0451	0.0319	84	5422	2414	0.0153	0.0336	0.0171	31	4790	1055	0.0064	0.0285	0.0076	248	14559	6284	0.0167	0.038	0.0189
GPT-3.5	zero-shot	1203	3770	1740	0.2419	0.4088	0.2634	940	4693	1556	0.1669	0.3766	0.1878	407	4183	677	0.0887	0.3755	0.1047	2550	12646	3973	0.1678	0.3909	0.1894
	1-shot	1300	3086	1643	0.2964	0.4417	0.3173	1068	3562	1428	0.2307	0.4279	0.2541	472	3086	612	0.1327	0.4354	0.1541	2840	9734	3683	0.2259	0.4354	0.2499
	2-shot	1443	2983	1500	0.326	0.4903	0.3494	1116	3157	1380	0.2612	0.4471	0.2849	486	2592	598	0.1579	0.4483	0.1814	3045	8732	3478	0.2586	0.4668	0.2839
	3-shot	1477	2646	1466	0.3582	0.5019	**0.38**	1114	3164	1382	0.2604	0.4463	0.2841	479	2416	605	0.1655	0.4419	0.1891	3070	8226	3453	0.2718	0.4706	0.2969
	4-shot	1330	2328	1613	0.3636	0.4519	0.3784	1089	2424	1407	0.31	0.4363	**0.329**	457	1870	627	0.1964	0.4216	**0.2199**	2876	6622	3647	0.3028	0.4409	**0.323**
FT GPT-2	zero-shot	1118	1479	1830	0.4305	0.3792	**0.4192**	928	1203	1570	0.4355	0.3715	**0.421**	383	792	703	0.326	0.3527	**0.331**	2429	3474	4103	0.4115	0.3719	**0.4029**
	1-shot	1127	1668	1821	0.4032	0.3823	0.3989	925	1325	1573	0.4111	0.3703	0.4022	382	913	704	0.295	0.3517	0.3048	2434	3906	4098	0.3839	0.3726	0.3816
	2-shot	1107	1700	1841	0.3944	0.3755	0.3904	937	1359	1561	0.4081	0.3751	0.401	383	919	703	0.2942	0.3527	0.3043	2427	3978	4105	0.3789	0.3716	0.3774
	3-shot	1073	1860	1875	0.3658	0.364	0.3655	874	1596	1624	0.3538	0.3499	0.353	381	1168	705	0.246	0.3508	0.2616	2328	4624	4204	0.3349	0.3564	0.339
	4-shot	1032	1911	1916	0.3507	0.3501	0.3505	818	1815	1680	0.3107	0.3275	0.3139	359	1310	727	0.2151	0.3306	0.2313	2209	5036	4323	0.3049	0.3382	0.311
SOTA	GECTOR	1046	632	2054	0.6234	0.3374	0.533	785	458	1836	0.6315	0.2995	0.5169	315	208	845	0.6023	0.2716	0.4843	2146	1298	4735	0.6231	0.3119	0.5194
	T5	1338	741	1762	0.6436	0.4316	0.586	1018	620	1603	0.6215	0.3884	0.5549	377	351	783	0.5179	0.325	0.4629	2733	1712	4148	0.6148	0.3972	0.5541

Table 3: Prompting results using GPT-2 (`gpt2-xl` and FT = fine-tuned), GPT-3.5 (`text-davinci-003`) and SOTA results by models of GECTOR (Omelianchuk et al., 2020) and T5 (Rothe et al., 2021).

Analysis and Discussion

1. Label-by-label evaluation approach
2. Is recall higher than precision in prompting GPT for the GEC task?
3. Results using various F-scores
4. Comparison between prompting GPT and SOTA
5. Discussion

Label-by-label evaluation approach

		TP	FP	FN	Prec	Rec	F0.5
M:PUNCT	A	189	171	134	0.525	0.5851	0.536
	B	203	132	133	0.606	0.6042	0.6056
	C	95	96	80	0.4974	0.5429	0.5059
R:VERB	A	21	60	113	0.2593	0.1567	0.2293
	B	17	55	113	0.2361	0.1308	0.2033
	C	6	43	51	0.1224	0.1053	0.1186
M	A	318	436	372	0.3703	0.3571	0.1691
	B	336	347	344	0.4919	0.4941	0.2458
	C	157	222	168	0.4142	0.4830	0.2180

Table 4: Detailed breakdown evaluation results for the most frequent errors, and Missing operation errors (FT GPT2, zero-shot).

Is recall higher than precision in prompting GPT for the GEC task?

- Consistent higher recall compared to precision showcases a tendency of over-correction in prompting GPT for the GEC task.
- We have observed that proficiency levels A and B, however, do not exhibit such a propensity.
- It holds true even for GPT-3.5, where recall consistently surpasses precision.
- Nevertheless, the difference between precision and recall measurements in levels A and B is considerably smaller compared to level C.

Results using various F-scores

	FT GPT-2			GPT-3.5		
	F0.5	F1	F2	F0.5	F1	F2
A	0.4192	0.4032	0.3885	0.3784	0.4030	0.4310
B	0.4210	0.4010	0.3827	0.3291	0.3625	0.4034
C	0.3310	0.3388	0.3470	0.2199	0.2680	0.3430
all	0.3907	0.4029	0.3792	0.3590	0.3230	0.4040

Table 5: Different F-scores with F0.5, F1 and F2. FT GPT-2 results are based on 0-shot, while GPT-3.5 (`text-davinci-003`) results are based on 4-shot.

Comparison between prompting GPT and SOTA

- State-of-the-art (SOTA) results continue to demonstrate superior performance compared to prompting GPT in the GEC task in all aspects of results including precision and recall measures regardless of proficiency levels.
- Our assumption is primarily based on the fact that SOTA models are usually subjected to extensive fine-tuning processes.

Discussion – Limitation

We examine a correlation between proficiency level C and native in prompting GPT in GEC as shown in Table 6, we are unable to identify any comparable behavior in prompting GPT in GEC for native-like proficiency C and native proficiency.

	TP	FP	FN	Prec	Rec	F0.5
C	383	792	703	0.326	0.3527	0.331
N	2429	3474	4103	0.4115	0.3719	0.4029

Table 6: Results between proficiency level C and native

Table 7 shows a behavior of prompting GPT in the GEC task proficiency specific errors, in which finding their correlation could be excessively challenging because of the performance of GEC for proficiency level C.

		TP	FP	FN	Prec	Rec	F0.5
M:PREP	B	24	29	31	0.4528	0.4364	0.4494
	C	9	23	17	0.2812	0.3462	0.2922
R:DET	B	15	30	41	0.3333	0.2679	0.3178
	C	7	12	23	0.3684	0.2333	0.3302

Table 7: Detailed breakdown evaluation results for `M:PREP` and `R:DET`

Conclusion

1. We investigated the strengths and limitations of prompting GPT for the GEC task based on different language proficiency levels.
2. We used our own implementations to calculate relevant metrics for label-by-label analysis.
3. We observed a tendency of over-correction in prompting GPT, and it is more obvious in the recent version of GPTs, where recall consistently surpasses precision.

References

Bryant, C., Felice, M., and Briscoe, T. (2017). Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Omelianchuk, K., Atrasevych, V., Chernodub, A., and Skurzhanskyi, O. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Rothe, S., Mallinson, J., Malmi, E., Krause, S., and Severyn, A. (2021). A Simple Recipe for Multilingual Grammatical Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Suzgun, M., Melas-Kyriazi, L., and Jurafsky, D. (2022). Prompt-and-Rerank: A Method for Zero-Shot and Few-Shot Arbitrary Textual Style Transfer with Small Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.