ACKSemEval-20252

Daniel Lee[*]
Adobe Inc.
dlee1@adobe.com

Harsh Sharma[*]
CU Boulder
harsh.sharma@colorado.edu

Jieun Han
KAIST
jieun_han@kaist.ac.kr

Sunny Jeong
New York University
sunny.jeong@nyu.edu

Alice Oh
KAIST
alice.oh@kaist.edu

Vered Shwartz
UBC
vshwartz@cs.ubc.ca

Abstract

13

MTtransformerLLMs

- 13LLMMT-
-
- BLEUCOMETM-ETA

-

1

MTRNN (Sutskever et al., 2014) transformer (**?**)[] koishekenov-etal-2023-memory, tang2020multilingualtranslationextensiblemultilingual, zhu-etal-2024-multilingual-contrastive, alves-etal-2023-steering, wang-etal-2023-document-level, zaranis-etal-2024-analyzing (Hershcovich et al., 2022) John WickRotten Tomatoes LLMs (Ponti et al., 2020) RAGKGsKG-MTMT (Conia et al., 2024)

(Pedersen, 2014; Díaz-Millón and Olvera-Lobo, 2023) - (Kim and Choi, 2015; Kim et al., 2022)

LLMMT13OpenAILLMGPT-4, GPT-4o, o1, o1-miniAnthropicClaude 3.5 Sonnet, 3.5 HaikuGemini 1.5 Flash, 1.5 ProMetaLlama3-8BGrokgrok-2DeepSeekR1-7BMTNLLB-200mBART-50 1.1 **??** Conia et al. (2025) XC-TranslateXC-Translate

1.1

5,082-

- BLEUn-gram (Papineni et al., 2002)
- COMET: (Rei et al., 2020)
- M-ETA (Conia et al., 2024)

Company	Models	Metrics *BLEU*	*COMET*	*M-ETA*
OpenAI	*o1*	[redacted]	0.9196	0.3752
	o1 Mini	0.3830	[redacted]	0.3306
	GPT-4o	0.3692	0.9087	0.3951
	GPT-4o Mini	0.3545	0.9046	0.2914
Anthropic	*Claude 3.5 Sonnet*	0.1961	0.8384	0.3969
	Claude 3.5 Haiku	0.1584	0.8056	0.2849
Google	*Gemini 1.5 Pro*	0.3810	0.9094	[redacted]
	Gemini 1.5 Flash	0.2965	0.9081	0.3316
xAI	*Grok 2*	0.3808	0.9143	0.3514
DeepSeek	*DeepSeek R1*	0.0066	0.4895	0.0026
Meta	*Llama 3*	0.0327	0.5529	0.0563
	Mbart-50	0.1451	0.8702	0.0791
	NLLB-200	0.2195	0.8899	0.1663

Table 1: -BLEUCOMETM-ETA
Color key: [redacted] = Highest BLEU, [redacted] = Highest COMET, [redacted] = Highest M-ETA.

BLEUCOMETn-gramM-ETA
1350-1235150

1 13 BLEUCOMET M-ETA BLEU o1 o1-mini COMET Gemini 1.5 Pro M-ETA Grok-2 GPT-4o MBART-50 NLLB-200 DeepSeek R1Llama 3 Claude 3.5 Haiku Sonnet
650459Grok 2266
(Popović, 2018) **??** 308266 **??** 5
?? **??**

Wikipedia2024Wikipedia 6 BLEU-COMET[0.26, 0.26, 0.25, 0.24, 0.27][0.84, 0.84, 0.83, 0.83, 0.83]M-ETA 0.00224 2 BLEU-COMET

WikidataNER (Tedeschi et al., 2021) 13 3

BLEUCOMET M-ETA BLEU COMET 650 0.41p 3.54e-28M-ETA 88.7 %

2

13LLMs

3

[*] These authors contributed equally.

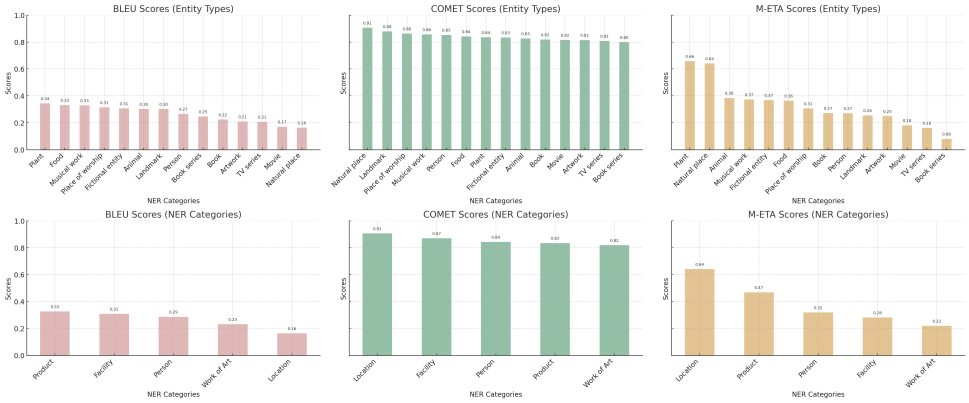

Figure 1: BLEUCOMETM-ETA

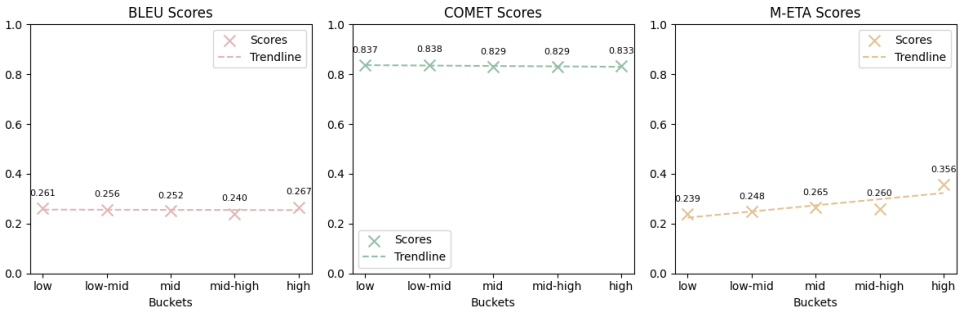

Figure 2: BLEUCOMETM-ETA

4

SemEval 2025 Task 2EA-MTSimone ConiaRevanth Gangi Reddy

References

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

Mar Díaz-Millón and María Dolores Olvera-Lobo. 2023. Towards a definition of transcreation: a systematic literature review. *Perspectives*, 31(2):347–364.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. KoCHET: A Korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505. International Committee on Computational Linguistics.

Youngsik Kim and Key-Sun Choi. 2015. Entity linking Korean text: An unsupervised learning approach using semantic relations. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 132–141. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311318, USA. Association for Computational Linguistics.

Daniel Pedersen. 2014. Exploring the concept of transcreation–transcreation as more than translation. *Cultus: The Journal of intercultural mediation and communication*, 7(1):57–71.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376. Association for Computational Linguistics.

Maja Popović. 2018. *Error Classification and Analysis for Machine Translation Quality Assessment*, pages 129–158. Springer International Publishing, Cham.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A

	BLEU	COMET	M-ETA	Annotator Score
o1		0.91	0.30	0.52
o1 Mini		0.91	0.38	0.50
GPT-4o	0.38		0.26	0.40
GPT-4o Mini	0.34	0.92	0.34	0.30
Claude 3.5 Sonnet	0.22	0.83	0.40	0.22
Claude 3.5 Haiku	0.14	0.80	0.24	0.12
Gemini 1.5 Pro		0.90		0.38
Gemini 1.5 Flash	0.29	0.92	0.28	0.34
Grok 2	0.32	0.92	0.36	
DeepkSeek R1	0.00	0.49	0.00	0.00
Llama 3	0.04	0.58	0.04	0.06
MBart05	0.17	0.87	0.10	0.22
NLLB-200	0.22	0.88	0.22	0.14

Table 2: BLEUCOMET M-ETA

Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA. = Highest Annotator Score.

Popularity Rank	Entity Type	BLEU	COMET	M-ETA
1	Plant		0.8354	
2	Book	0.2245	0.8188	0.2721
3	Person	0.2666	0.8526	0.2704
4	Artwork	0.2096	0.8148	0.2497
5	Food	0.3317	0.8421	0.3646
6	Movie	0.1707	0.8151	0.1812
7	Fictional entity	0.3070	0.8337	0.3685
8	Animal	0.3026	0.8257	0.3846
9	Landmark	0.3026	0.8784	0.2552
10	TV series	0.2078	0.8077	0.1628
11	Place of worship	0.3141	0.8625	0.3070
12	Natural place	0.1638		0.6410
13	Musical work	0.3297	0.8558	0.3735
14	Book series	0.2471	0.7992	0.0804

Table 3:

Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA.

Models	Metric	Entity Types													
		Animal	Artwork	Book	Book Series	Fictional Entity	Food	Landmark	Movie	Musical Work	Natural place	Person	Place of Worship	Plant	TV Series
o1	BLEU	0.3712	0.3236	0.3334		0.4659	0.4988	0.4284	0.2953	0.4477	0.2108				
	Comet	0.9393	0.9018	0.9029	0.8779			0.9537	0.8965		0.9720				0.8862
	M-ETA	0.5000	0.3602	0.3634	0.1940		0.5092	0.3413	0.2683	0.3912	1.0000	0.4118	0.4448	0.8182	0.2396
o1 Mini	BLEU	0.3698	0.3204	0.3504	0.3827		0.5301		0.2143	0.4869	0.2927	0.4034	0.4563	0.5268	0.3186
	Comet	0.9223			0.8770	0.9359	0.9430			0.9385		0.9410	0.9435	0.9324	0.8838
	M-ETA	0.0000	0.2762	0.3225	0.0448	0.4934	0.4479	0.2857	0.1847	0.5705	1.0000	0.3069	0.3785	0.7273	0.1506
GPT-4o	BLEU		0.2926	0.3204	0.3116	0.4544		0.4332	0.2423	0.4644	0.2717	0.4211	0.4491	0.6030	0.2835
	Comet	0.9046	0.8938	0.8975	0.8629	0.9122	0.9418	0.9495	0.8858	0.9217	0.9684	0.9276	0.9363	0.9420	0.8801
	M-ETA	0.6667	0.3782	0.4091	0.1493	0.5093	0.4847	0.3651	0.2840	0.5034	0.6667	0.3890	0.4038	0.9091	0.2927
GPT-4o Mini	BLEU	0.4254	0.2755	0.3049	0.3569	0.4230	0.4827	0.4144	0.2517	0.4520	0.3137	0.3789	0.4261	0.5220	0.2968
	Comet	0.8876	0.8782	0.8786	0.8737	0.9090	0.9374	0.9487	0.8943	0.9202	0.9720	0.9213	0.9346	0.9496	0.8845
	M-ETA	0.1667	0.2736	0.2960	0.0597	0.4164	0.4387	0.3016	0.1829	0.3735	0.6667	0.2692	0.3596	0.9091	0.1664
Claude-3.5 Sonnet	BLEU	0.1324	0.1507	0.1685	0.1644	0.2000	0.2349	0.2578	0.1257	0.3372	0.0000	0.1681	0.2243	0.2190	0.1674
	Comet	0.7358	0.8193	0.8329	0.8183	0.7991	0.8123	0.8787	0.8466	0.8939	0.8295	0.8340	0.8463	0.7539	0.8269
	M-ETA	0.8333	0.3448	0.3682	0.0597	0.4934		0.3333	0.2526	0.6142	0.6667	0.3688			0.2439
Claude-3.5 Haiku	BLEU	0.0407	0.1320	0.1377	0.1921	0.1568	0.2241	0.2089	0.1129	0.2528	0.0000	0.1255	0.1492	0.1348	0.1331
	Comet	0.6727	0.7875	0.7851	0.7962	0.7663	0.7913	0.8724	0.7834	0.8557	0.8263	0.8035	0.8346	0.7320	0.7981
	M-ETA	0.8333	0.2787	0.2960	0.1045	0.3979	0.4755	0.2937	0.2073	0.2599		0.2948	0.3880	0.8182	0.1535
Gemini-1.5-Pro	BLEU	0.4971			0.2932	0.4189	0.4678	0.4132		0.4914	0.2381	0.3786	0.4714	0.4760	0.2703
	Comet		0.8927	0.8920	0.8578	0.9166	0.9371	0.9501	0.8812	0.9371	0.9667	0.9291	0.9412	0.9522	0.8739
	M-ETA	0.5000				0.5570	0.5245				0.0000		0.4227	0.8182	
Gemini-1.5-Flash	BLEU	0.4989	0.2424	0.2656	0.2561	0.3868	0.4001	0.3482	0.2291	0.3242	0.0000	0.3308	0.3846	0.5315	0.2332
	Comet	0.9499	0.8876	0.8932	0.8548	0.9207	0.9292	0.9459	0.8907	0.9248	0.9634	0.9342	0.9393	0.9532	0.8721
	M-ETA		0.3045	0.3586	0.0896	0.4695	0.4417	0.3175	0.2439	0.3748	0.3333	0.3513	0.4006	0.9091	0.2023
Grok 2	BLEU	0.3493	0.2808	0.3042	0.3665	0.4617	0.4973	0.4121	0.2571	0.5490	0.3137	0.3912	0.4749	0.4891	0.3171
	Comet	0.9053	0.8923	0.8980		0.9250	0.9387	0.9499	0.8902	0.9356	0.9747	0.9358	0.9443	0.9516	
	M-ETA	0.1667	0.3113	0.3430	0.1343	0.5013	0.4417	0.2460	0.2213		1.0000	0.3163	0.3849	0.9091	0.2023
DeepSeek-R1-7B	BLEU	0.0315	0.0076	0.0076	0.0085	0.0034	0.0060	0.0137	0.0049	0.0071	0.0000	0.0044	0.0045	0.0000	0.0090
	Comet	0.5619	0.4891	0.4971	0.4952	0.4752	0.4664	0.4914	0.5036	0.4911	0.5595	0.4858	0.4714	0.4533	0.5038
	M-ETA	0.0000	0.0026	0.0036	0.0000	0.0053	0.0031	0.0079	0.0000	0.0000	0.0000	0.0013	0.0000	0.0000	0.0072
Llama 3	BLEU	0.0128	0.0154	0.0193	0.0407	0.0566	0.0601	0.0642	0.0130	0.0310	0.0000	0.0450	0.0570	0.0000	0.0238
	Comet	0.5322	0.4972	0.4997	0.5121	0.5633	0.5500	0.6507	0.5124	0.5784	0.8461	0.6225	0.6564	0.4895	0.5111
	M-ETA	0.0000	0.0223	0.0409	0.0000	0.1114	0.1258	0.0952	0.0261	0.0328	0.6667	0.0983	0.0946	0.0000	0.0316
mBART-Large-50	BLEU	0.2799	0.1540	0.1573	0.1808	0.1977	0.1655	0.1982	0.0716	0.1282	0.1291	0.1684	0.1949	0.0529	0.1125
	Comet	0.8659	0.8647	0.8692	0.8330	0.8814	0.8848	0.9301	0.8330	0.8897	0.9638	0.8878	0.9007	0.8821	0.8341
	M-ETA	0.0000	0.0823	0.0927	0.0000	0.1088	0.1227	0.1746	0.0488	0.0643	1.0000	0.0888	0.0726	0.2727	0.0488
NLLB-200	BLEU	0.3999	0.1781	0.1805	0.2620	0.2638	0.2141	0.3061	0.1044	0.3141		0.2258	0.2895	0.2180	0.2159
	Comet	0.9060	0.8813	0.8870	0.8429	0.8948	0.8719	0.9447	0.8800	0.8998	0.9566	0.9183	0.9146	0.9107	0.8576
	M-ETA	0.5000	0.1244	0.1252	0.0149	0.1565	0.1810	0.1667	0.1010	0.3776	0.3333	0.1575	0.1924	0.4545	0.0488

Table 4: BLEUCOMETM-ETA

Color key: = Highest BLEU, = Highest COMET, 0.5pt = Highest M-ETA.

Type of Error	Definition
Literal Translation	Translation follows the meaning of the source language.
Phonetic Translation	Translation follows how it sounds in the source language.
Word-Level Translation	Translation is done word-for-word from the source language.
Incorrect Entity Name	Used a different or less appropriate entity name.
Incorrect Grammar	Grammar mistakes in the target language.
Incorrect Language	Translated into the wrong language.
Incorrect Formatting	Formatting is wrong, but the translation itself is correct.
Added/Deleted Content	Extra parts added or parts missing compared to the source.
Incorrect Response	Output that doesn't match the source text meaning.
Partial Translation	Only part of the source text is translated.
Romanized Korean	Latin alphabet used instead of proper Korean script.
Gibberish	Output makes no sense at all.

Table 5:

Models	Metric	Popularity Level				
		Low *142 - 12943*	Low-mid *12944 - 29440*	Mid *29441 - 62685*	Mid-high *62686 - 157350*	High *157351 - 6974823*
o1	BLEU	0.3789	0.3764	0.3760		
	COMET	0.9167	0.9173		0.9184	0.9297
	M-ETA	0.3171	0.3259	0.3415	0.3579	0.5321
o1 Mini	BLEU	0.3883		0.3723	0.3621	0.4180
	COMET			0.9145		
	M-ETA	0.3028	0.3005	0.3129	0.3013	0.4455
GPT-4o	BLEU	0.3713	0.3689	0.3676	0.3496	0.3910
	COMET	0.9082	0.9122	0.9036	0.9094	0.9109
	M-ETA	0.3618	0.3492	0.4006	0.3589	0.5066
GPT-4o Mini	BLEU	0.3720	0.3615	0.3373	0.3449	0.3607
	COMET	0.9078	0.9082	0.8960	0.9058	0.9059
	M-ETA	0.2530	0.2680	0.2681	0.2841	0.3894
Claude 3.5 Sonnet	BLEU	0.2088	0.2114	0.1996	0.1672	0.1977
	COMET	0.8490	0.8530	0.8374	0.8264	0.8285
	M-ETA	0.3567	0.3777	0.3812	0.3761	0.4995
Claude 3.5 Haiku	BLEU	0.1881	0.1794	0.1557	0.1332	0.1393
	COMET	0.8269	0.8219	0.8032	0.7918	0.7829
	M-ETA	0.2266	0.2538	0.2803	0.2781	0.3914
Gemini 1.5 Pro	BLEU	0.3707	0.3776	0.3743	0.3691	0.4239
	COMET	0.9108	0.9127	0.8995	0.9066	0.9183
	M-ETA					
Gemini 1.5 Flash	BLEU	0.2864	0.2827	0.2948	0.2923	0.3305
	COMET	0.9079	0.9117	0.9035	0.9058	0.9120
	M-ETA	0.2571	0.2964	0.3425	0.3195	0.4444
Grok 2	BLEU		0.3779		0.3498	0.3869
	COMET	0.9130	0.9157	0.9118	0.9135	0.9173
	M-ETA	0.3018	0.3147	0.3405	0.3488	0.4648
DeepSeek R1	BLEU	0.0091	0.0087	0.0051	0.0052	0.0041
	COMET	0.4955	0.4924	0.4852	0.4837	0.4892
	M-ETA	0.0041	0.0020	0.0020	0.0020	0.0020
Llama 3	BLEU	0.0369	0.0281	0.0376	0.0315	0.0286
	COMET	0.5652	0.5634	0.5513	0.5373	0.5466
	M-ETA	0.0539	0.0315	0.0550	0.0586	0.0765
MBart-50	BLEU	0.1446	0.1411	0.1464	0.1453	0.1442
	COMET	0.8700	0.8742	0.8675	0.8660	0.8735
	M-ETA	0.0640	0.0558	0.0744	0.0809	0.1172
NLLB-200	BLEU	0.2317	0.2399	0.2236	0.2006	0.2042
	COMET	0.8915	0.8933	0.8871	0.8912	0.8884
	M-ETA	0.1667	0.1848	0.1784	0.1547	0.1600

Table 6: BLEUCOMETM-ETA

Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA.

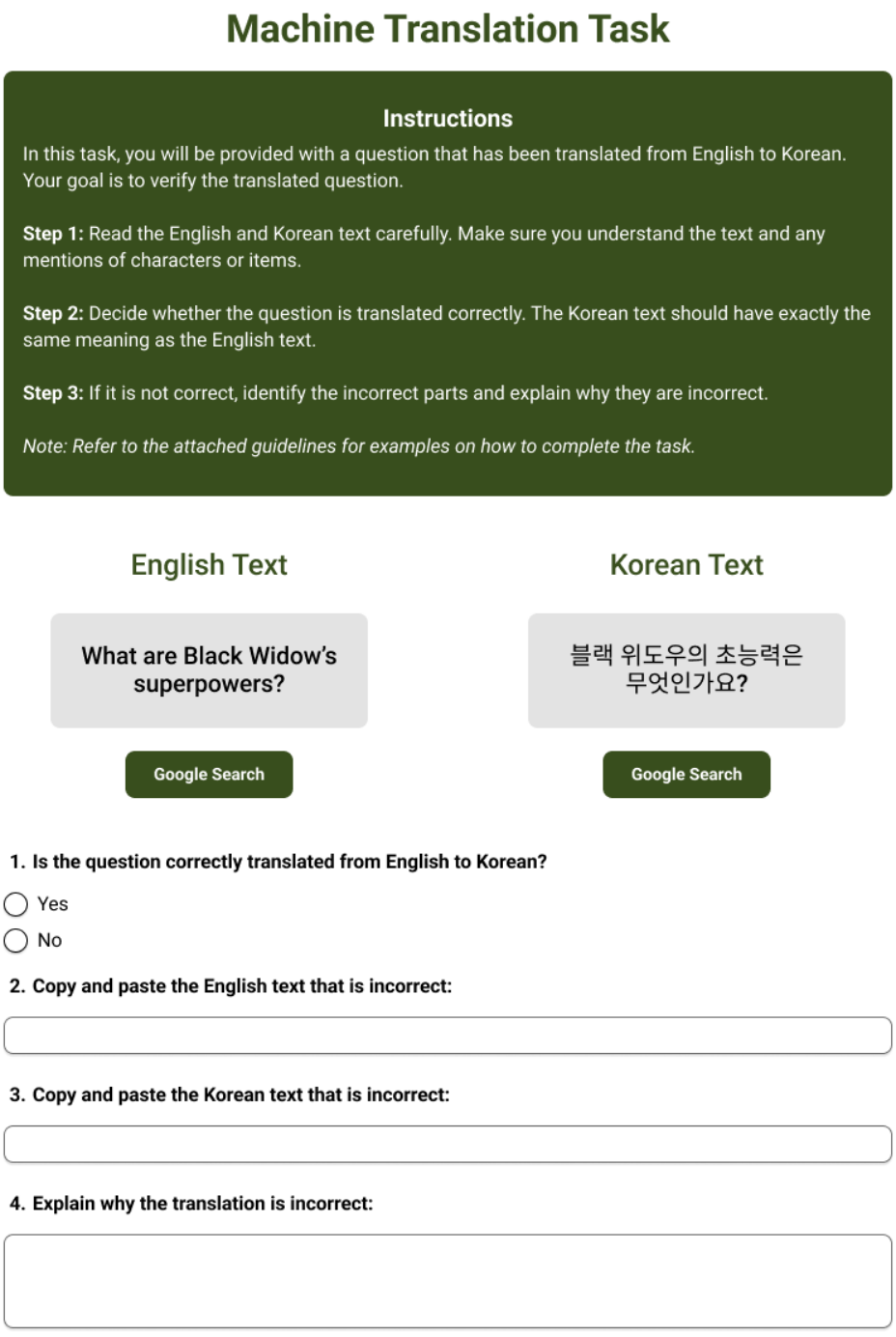

Figure 3: