MAGNET - MAchines GeNErating Translations: A CALAMITA Challenge

Mauro Cettolo[1,*,†], Andrea Piergentili[1,2,†], Sara Papi[1], Marco Gaido[1], Matteo Negri[1] and Luisa Bentivogli[1]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Italy

Abstract

We propose MAGNET - MAchines GeNErating Translations, a CALAMITA Challenge which aims at testing the ability of large language models (LLMs) in the hot topic of automatic translation, focusing on Italian and English (in both directions) to overcome the marginality with which Italian is considered by the machine translation community. We propose a benchmark composed of two portions with different distribution policies (one free to use, the other not discloseable), allowing to handle data contamination issues. The publicly available section of the benchmark is distributed on Hugging Face, whereas in this report we describe the details of our challenge, including the prompt formats to be used. Additionally, we report the performance of five models, including a LLM and different sized translation models, in terms of four evaluation metrics, whose scores allow an overall evaluation of the quality of the automatically generated translations.

Keywords

Machine translation, English-Italian, FLORES+, Bleu, ChrF, Bleurt, Comet, Llama3-8B-Instruct, mBART50, NLLB

1. Introduction and Motivation

Machine Translation (MT) refers to the process, carried out by a computer program, of translating text from one language to another without human involvement. The idea of using digital computers to translate natural languages dates back to the 1940s, making MT one of the oldest fields of artificial intelligence. Since then, the improvement in translation quality has been constant and achieved through increasingly effective approaches (rule-, example- and statistical-based); however, the most significant advances have likely been observed over the last few years, thanks to the introduction of neural networks. Neural models specifically trained for accomplishing the translation task, like DeepL Translator,[1] reach outstanding quality, even if the so-called human parity has not been achieved yet, especially in unrestricted domains and for language pairs not involving English. Recently, an alternative neural-based method is gathering a lot of interest due to its undoubted potential; it consists in prompting generative large language models (LLMs), like GPT models [1, 2] and the LLama model family [3, 4, 5], to translate a text. Whatever the approach, the MT research community is much focused on the development and validation of models covering English and few other languages, paying little attention or completely neglecting the vast majority of the more than 7,000 languages spoken in the world, including Italian. On the other hand, the global MT market size was valued at USD 847.24 million in 2021 and is expected to expand at a compound annual growth rate of 16.4% in 2024-2031, reaching USD 2107.56 million by 2027.[2] Being Europe, and then Italy, one of the leading regions for the MT market, CALAMITA [6] cannot miss MT. Therefore we propose the challenge of testing the LLMs ability in the hot topic of automatic translation, focusing on Italian and English (in both directions) to overcome the marginality with which Italian is considered by the MT community.

2. Challenge: Description

The MAGNET challenge provides a framework for assessing the ability of LLMs in translating Italian text into English and vice-versa. It is organized following the blueprint of other long-standing MT shared tasks, such as those proposed in the WMT[3] and IWSLT[4] conferences, where Organizers prepare and distribute *development* and *test* sets, define the training conditions, possibly providing specific training data, establish the evaluation modalities, typically via automatic metrics and occasionally enriched by human evaluations, collect and evaluate participants' submissions, and finally disclose the results.

The MAGNET challenge supplies a benchmark divided in two portions: one based on a publicly available MT benchmark and a private one (see Section 3). This allows participants not only to evaluate their models but possibly to also fine-tune them, by exploiting the open portion of the MAGNET benchmark for development purposes.

Multiple evaluation metrics are employed so as to have a comprehensive overview of the quality of the translations generated by a specific model. Indeed, shared tasks on automatic metrics are still being organized,[5] as evidence of the fact that none of the metrics designed up to now by the scientific community has proven capable of covering every single aspect that defines a "good" translation by itself .

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
†These authors contributed equally.
cettolo@fbk.eu (M. Cettolo); apiergentili@fbk.eu (A. Piergentili); spapi@fbk.eu (S. Papi); mgaido@fbk.eu (M. Gaido); negri@fbk.eu (M. Negri); bentivo@fbk.eu (L. Bentivogli)
https://mt.fbk.eu/author/cettolo/ (M. Cettolo); https://mt.fbk.eu/author/apiergentili/ (A. Piergentili); https://mt.fbk.eu/author/spapi/ (S. Papi); https://mt.fbk.eu/author/mgaido/ (M. Gaido); https://mt.fbk.eu/author/negri/ (M. Negri); https://mt.fbk.eu/author/bentivogli/ (L. Bentivogli)
0000-0001-8388-497X (M. Cettolo); 0000-0002-4494-8886 (A. Piergentili); 0000-0002-4494-8886 (S. Papi); 0000-0003-4217-1396 (M. Gaido); 0000-0002-8811-4330 (M. Negri); 0000-0001-7480-2231 (L. Bentivogli)

[1]https://en.wikipedia.org/wiki/DeepL_Translator

[2]https://www.linkedin.com/pulse/machine-translation-mt-market- size-2024-suhoe/

[3]https://www2.statmt.org/wmt24/translation-task.html

[4]https://iwslt.org/2024/#shared-tasks

[5]https://www2.statmt.org/wmt24/metrics-task.html

In addition, in order to allow for comparisons, scores measured on the translation generated by Llama3-8B-Instruct and a number of other models are made available (see Section 4).

3. Data description

We test LLMs' ability to translate between Italian and English using a parallel corpus composed of two parts: an OPEN portion and a CLOSED one.

OPEN For the OPEN portion of the MAGNET benchmark we propose FLORES+, the latest version of FLORES-200[6] [7], a multilingual MT evaluation benchmark released under CC BY-SA 4.0 by FAIR researchers at Meta. It consists of English sentences sampled in equal amounts from Wikinews (an international news source), Wikijunior (a collection of age-appropriate non-fiction books), and Wikivoyage (a travel guide), translated into more than 200 languages, including Italian. *Dev* and *devtest* sets consisting of about 1,000 segments each are provided. See Section 3.3 for statistics on this portion of the MAGNET benchmark.

CLOSED The CLOSED subset is a MT test set developed by FBK by collecting texts of English and Italian news, and then commissioning their professional translation to a specialized company. This resource is private and not publicly accessible. See Section 3.3 for statistics on this portion of the MAGNET benchmark.

Both subsets allow for the evaluation of MT quality in both translation directions, i.e. English→Italian and Italian→English. The decision to split our benchmark in two subsets is primarily motivated by their current distribution policy, which is inherently linked to growing concerns about *data contamination* [8]. Data contamination refers to the possibility that the input-output pairs used in LLM tests occur in the huge data sets typically used for pre-training and fine-tuning; such overlap can lead to inflated benchmark scores, creating an overly favorable impression of an LLM's abilities. Although it is challenging to determine with certainty whether the models being evaluated were trained on popular datasets scraped from the web, this possibility should be taken seriously. To promote sound evaluation and mitigate the effects of biased or potentially misleading results due to data contamination, one approach is to rely exclusively on – or at least include among the benchmarks – "safe" datasets that are either private or have very controlled/limited distribution. Therefore, pairing a larger, widely used public dataset (FLORES+) with a smaller, in-house dataset – the CLOSED subset – aims to strike a balance between the thoroughness and the reliability of the evaluation.

3.1. Data format

The datasets are organized in a parallel text format, i.e. every entry is composed of a sentence in one language and the corresponding translation. The OPEN portion of the benchmark is publicly available on Hugging Face,[7] whereas access to the CLOSED portion is only provided to the Organizers of the task.

3.2. Prompts

Table 1 reports the simple prompt formats we propose. Both contain a simple translation instruction first, followed by the source sentence, and then the target language translation in a new line. We include four iterations of this format in the actual prompts before appending the input, so as to activate LLMs' in-context learning ability [1].

Both the source and the translation are surrounded by the characters < and >. This instructs the model to reproduce this format in its output as well. We do so to address LLMs' tendency to include unwanted extra comments in their outputs. Such comments would compromise all automatic evaluations (see Section 4) due to the presence of extra content in the candidate outputs, which is penalized by the string-based metrics and alters the vector representations used by the model-based metrics to compute similarity scores.

3.3. Detailed data statistics

In Table 2 detailed statistics are provided on the various sections of the benchmark in terms of number of segments (#seg), and of English (|en|) and Italian (|it|) words.

4. Metrics

We evaluate LLMs' performance in translation using a set of four automatic metrics selected in light of the ongoing challenges in MT evaluation, which still pose an open problem. New metrics are indeed continually proposed, and evaluation campaigns aimed at assessing these metrics are organised periodically (for example, the annual WMT Metrics Shared Task [9]). Broadly, automatic metrics can be divided into string-based metrics and metrics using pretrained models, with either group having both strengths and weaknesses [10]. Therefore, for a more comprehensive translation quality evaluation accounting for their complementarity, we propose to adopt a couple of metrics from each group, selected among the most commonly used ones:

- string-based: BLEU[8] [11] and CHRF[9] [12] via sacreBLEU [13]
- pretrained models-based: BLEURT [14] (checkpoint: `BLEURT-20`) and COMET [15] (model: `wmt22-comet-da`).

All of them are quality metrics, that is the higher the score the better the translation. The overview of the scores from all these metrics allows for a robust assessment of the quality of individual models, and a fair comparison between different models as well.

We provide reference performance on our challenge of one of the most popular open LLMs, and four state-of-the-art MT models:

[6] https://github.com/openlanguagedata/flores
[7] https://huggingface.co/datasets/FBK-MT/MAGNETbenchmark4CALAMITA24

[8] sacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0`
[9] sacreBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.0.0`

prompt	content
en-it	Translate the following sentence into Italian: <On Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool that can sort cells by type: a tiny printable chip that can be manufactured using standard inkjet printers for possibly about one U.S. cent each.> <Nella giornata di lunedí, alcuni scienziati della Scuola di Medicina dell'Università di Stanford hanno annunciato l'invenzione di un nuovo strumento diagnostico capace di ordinare le cellule in base al tipo: un chip minuscolo che può essere stampato utilizzando stampanti a getto di inchiostro al costo di circa 1 centesimo di dollaro l'uno.>
it-en	Translate the following sentence into English: <Nella giornata di lunedí, alcuni scienziati della Scuola di Medicina dell'Università di Stanford hanno annunciato l'invenzione di un nuovo strumento diagnostico capace di ordinare le cellule in base al tipo: un chip minuscolo che può essere stampato utilizzando stampanti a getto di inchiostro al costo di circa 1 centesimo di dollaro l'uno.> <On Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool that can sort cells by type: a tiny printable chip that can be manufactured using standard inkjet printers for possibly about one U.S. cent each.>

Table 1
Examples of the format of prompts proposed for MT Challenge. Prompt en-it is designed for the translation from English into Italian, prompt it-en for the opposite direction. In both cases, for instructing Llama3-8B-Instruct only one single shot taken from the OPEN dev set is shown, while in experiments of Section 4 four shots are provided to the model.

Data	Set	#seg	|en|	|it|
OPEN	dev	997	21.0k	23.0k
	devtst	1012	21.9k	24.3k
CLOSED	UK	589	10.6k	11.2k
	US	599	10.0k	9.7k
	IT	547	10.8k	10.3k

Table 2
Statistics of the benchmark in terms of number of segments and of (detokenized) words on English and Italian sides.

Llama-3-8B-Instruct:[10] a LLM from the Llama 3 model family [5]. It is an instruction-tuned model, i.e. it is fine-tuned to align its outputs with the desired response characteristics [16], in this case for assistant-like chat. Therefore, we provide the 4-shot prompts described in Section 3.2 as input for the model in a chat format, with *user* role messages with the instruction and the input and *assistant* role messages with the corresponding output.[11]

HelsinkiMT:[12] the Language Technology Research Group at the University of Helsinki made available under the CC-BY-4.0 license a set of neural MT models trained with MarianNMT[13] on OPUS data,[14] including English-Italian[15] and Italian-English[16] models.

mBART50:[17] a multilingual neural translation model that covers any pair from a set of 50 languages, English and Italian included [17]. Built by Meta/Facebook on the fairseq toolkit,[18] it is released under the MIT license. Its network has approximately 600M parameters.

NLLB:[19] No Language Left Behind (NLLB) is also a multilingual neural translation model that covers any pair from more than 200 languages, including the two we are interested in. The code was developed by Meta/Facebook as a branch of fairseq and is released under the MIT license. Five different NLLB models are available under the CC-BY-NC 4.0 license, which mainly differ in size, ranging from the smallest with 600M parameters to the largest with 54.5B parameters. On the basis of their manageability and official performance claimed by the authors, we decided to include two NLLB models in this investigation, the distilled variant with 1.3B parameters (**NLLB_1.3B**) and the one with 3.3B parameters (**NLLB_3.3B**).

Table 3 provides the scores measured for each model on all evaluation sets of the benchmark, except for the OPEN *dev* set, since we reserved that subset as the source of the exemplars used for few-shot prompting with Llama-3-8B-Instruct. First of all, we note that the performance of the three multilingual translation models mBART50, NLLB_1.3B and NLLB_3.3B are strictly in increasing order according to their number of parameters, with respect to all metrics (with only one microscopic exception). In general, Llama-3-8B-Instruct performs better than mBART50 and worse than NLLB_1.3B.

The behavior of HelsinkiMT is more difficult to frame: there are cases in which it is definitely the best performing model (CLOSED-IT, it→en) or at least competitive with NLLB_3.3B (CLOSED-UK, en→it; CLOSED-IT, en→it); others in which it is only slightly better than mBART50 (OPEN devtst, it→en; CLOSED-US, it→en). This can probably be explained by the fact that HelsinkiMT is not a single model, rather a collection of models specifically trained for covering the translation between specific languages. That is, HelsinkiMT en→it and it→en models were trained independently, on different training data. Therefore, it is possible that their performance when compared to that of other models may not be consistent across the various sections of our benchmark.

In summary, we can state that Llama-3-8B-Instruct, a general purpose, generative model only conditioned towards performing translation by four task exemplars, compares well to translation models; likely, fine-tuning Llama-3-8B-Instruct on the translation task could allow it to achieve even better performance. However, it should be considered that this version of Llama-3-8B-Instruct – which is also the smallest of that model family – has 8B parameters, more than twice the parameters of NLLB_3.3B and an order of magnitude more than mBART50.

[10] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[11] https://huggingface.co/docs/transformers/main/en/chat_templating
[12] https://github.com/Helsinki-NLP/Opus-MT
[13] https://marian-nmt.github.io/
[14] https://opus.nlpl.eu/
[15] https://huggingface.co/Helsinki-NLP/opus-mt-en-it
[16] https://huggingface.co/Helsinki-NLP/opus-mt-it-en
[17] https://huggingface.co/facebook/mbart-large-50
[18] https://github.com/facebookresearch/fairseq
[19] https://github.com/facebookresearch/fairseq/tree/nllb

system	it→en				en→it			
	BLEU	ChrF	BLEURT	COMET	BLEU	ChrF	BLEURT	COMET
OPEN – devtst								
HelsinkiMT	29.39	60.00	0.7568	0.8656	27.53	57.61	0.7422	0.8521
mBART50	27.34	57.64	0.7371	0.8494	23.88	54.34	0.7322	0.8502
NLLB_1.3B	**35.08**	62.42	0.7732	0.8774	29.31	58.04	0.7773	0.8749
NLLB_3.3B	35.03	**63.04**	**0.7781**	**0.8805**	**29.95**	**58.74**	**0.7871**	**0.8811**
Llama-3-8B-Instruct	32.04	62.03	0.7778	0.8795	26.36	56.60	0.7710	0.8758
CLOSED – UK								
HelsinkiMT	48.06	71.78	0.8038	0.8949	**57.35**	**76.99**	0.7998	0.8836
mBART50	43.77	68.79	0.7789	0.8776	47.46	70.68	0.7910	0.8837
NLLB_1.3B	52.48	73.83	0.8072	0.8954	55.12	74.62	0.8160	0.8933
NLLB_3.3B	**54.61**	**75.09**	**0.8096**	0.8968	56.00	75.28	**0.8210**	**0.8937**
Llama-3-8B-Instruct	46.61	71.02	0.8088	**0.8985**	39.29	66.50	0.7948	0.8840
CLOSED – US								
HelsinkiMT	39.26	62.25	0.7459	0.8571	39.02	64.41	0.7395	0.8394
mBART50	37.54	60.78	0.7314	0.8437	34.19	60.79	0.7309	0.8420
NLLB_1.3B	42.72	64.76	0.7449	0.8544	39.91	64.40	0.7580	0.8566
NLLB_3.3B	**43.36**	**65.23**	0.7483	0.8585	**40.35**	**64.63**	**0.7681**	**0.8583**
Llama-3-8B-Instruct	39.08	62.53	**0.7502**	**0.8613**	28.73	58.24	0.7355	0.8469
CLOSED – IT								
HelsinkiMT	**59.14**	**77.83**	**0.7814**	**0.8515**	**48.90**	**74.47**	0.8278	0.8898
mBART50	39.00	63.98	0.7101	0.8029	37.24	66.65	0.7858	0.8679
NLLB_1.3B	49.17	69.88	0.7361	0.8251	46.48	72.32	0.8212	0.8896
NLLB_3.3B	50.33	70.67	0.7373	0.8271	47.67	73.56	**0.8285**	**0.8928**
Llama-3-8B-Instruct	43.89	68.96	0.7660	0.8496	37.19	67.64	0.7996	0.8797

Table 3
Translation results on benchmark of MT models and LLMs. The best scores for each translation direction, subset, and metric are signalled in bold.

5. Limitations

Nowadays, LLMs are trained on huge amounts of data mostly crawled from the web. Therefore, as already pointed out in Section 3, it is hard to be sure that there is no data contamination, that is no overlap between training and evaluation data. Data contamination makes the evaluation of LLMs unreliable since their performance may be inflated.

Concerning our specific case, the risk that OPEN/FLORES+ data are contaminated is not negligible; however the results shown in Table 3, which are good but realistic, do not seem to indicate any contamination.

In theory, the contamination risk of the CLOSED section is lower than for the CLOSED one, since the translations of the original texts have never been released. On the other hand, original texts are available on the web (although only for private use), therefore it cannot be ruled out that the models "know" them, in some way. For example, the exceptionally high results of HelsinkiMT on the CLOSED-IT set seem to be an anomaly, likely due to data contamination.

6. Ethical issues

Our proposal does not focus on ethically charged topics. While the data we propose for the evaluation of automatic translation may mention sensitive topics or be afflicted by ethical issues such as social biases (e.g., gender bias), here we focus solely on MT quality evaluation and leave the investigation of ethical aspects to other resources and analyses.

7. Data license and copyright issues

The OPEN section of our benchmark is part of the FLORES+ dataset which is licensed under the *Creative Commons Attribution Share Alike 4.0 International*,[20] which requires derivatives to be distributed under the same or a similar, compatible license. We opted for the same license.

There is no license associated with the CLOSED part of our benchmark as it is not distributed and can only be used by CALAMITA Organizers for evaluation purposes.

Acknowledgments

The work presented in this paper is funded by the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People) and the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing

[20] https://github.com/openlanguagedata/flores/blob/main/LICENSE

Systems, volume 33, 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] OpenAI, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[4] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

[5] A. Dubey, et al., The Llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[6] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[7] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia-Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, 2022. arXiv:arXiv:1902.01382.

[8] C. Deng, Y. Zhao, X. Tang, M. Gerstein, A. Cohan, Investigating data contamination in modern benchmarks for large language models, in: Proc. of NAACL (Volume 1: Long Papers), Mexico City, Mexico, 2024, pp. 8706–8719. URL: https://aclanthology.org/2024.naacl-long.482.

[9] M. Freitag, N. Mathur, C.-k. Lo, E. Avramidis, R. Rei, B. Thompson, T. Kocmi, F. Blain, D. Deutsch, C. Stewart, C. Zerva, S. Castilho, A. Lavie, G. Foster, Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent, in: Proc. of WMT, Singapore, 2023, pp. 578–628. URL: https://aclanthology.org/2023.wmt-1.51.

[10] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, A. Menezes, To ship or not to ship: An extensive evaluation of automatic metrics for machine translation, in: Proc. of WMT, Online, 2021, pp. 478–494. URL: https://aclanthology.org/2021.wmt-1.57.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in: Proc. of ACL, Philadelphia, USA, 2002, pp. 311–318.

[12] M. Popovic, chrF: character n-gram F-score for automatic MT evaluation, in: Proc. of WMT, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049.

[13] M. Post, A Call for Clarity in Reporting BLEU Scores, in: Proc. of WMT, Belgium, Brussels, 2018, pp. 186–191. URL: https://www.aclweb.org/anthology/W18-6319.

[14] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: Proc. of ACL, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704.

[15] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, A. F. T. Martins, COMET-22: Unbabel-IST 2022 submission for the metrics shared task, in: Proc. of WMT, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 578–585. URL: https://aclanthology.org/2022.wmt-1.52.

[16] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, G. Wang, Instruction tuning for large language models: A survey, 2024. URL: https://arxiv.org/abs/2308.10792. arXiv:2308.10792.

[17] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, 2020. URL: https://arxiv.org/abs/2008.00401. arXiv:2008.00401.