The 2024 GEM Shared Task on Multilingual Data-to-Text Generation and Summarization: Overview and Preliminary Results

Simon Mille[1], **João Sedoc**[2], **Yixin Liu**[3], **Elizabeth Clark**[4], **Agnes Axelsson**[5], **Miruna Clinciu**[6], **Yufang Hou**[7], **Saad Mahamood**[8], **Ishmael Obonyo**[9], **Lining Zhang**[2]

[1]ADAPT, Dublin City University, [2]New York University, [3]Yale University,
[4]Google DeepMind, [5]Delft University of Technology, [6]Heriot Watt University,
[7]IBM Research, [8]Trivago, [9]Technical University of Kenya

Correspondence: simon.mille@adaptcentre.ie, jsedoc@stern.nyu.edu

Abstract

We present an overview of the GEM 2024 shared task, which comprised both data-to-text generation and text summarization. New datasets were compiled specifically for the task to reduce the data contamination issue in large language models (LLMs) that the participants were likely to use. The paper describes the tasks, datasets, participating systems, evaluation methods, and some preliminary results. The full results will be presented at INLG '24. In this paper, we provide (i) the metrics results for English texts on six different data-to-text test sets for which we collected new reference texts, and (ii) the metrics results for Swahili on the text summarization test set.

1 Introduction

Since its inception, the Generation, its Evaluation and Metrics initiative (GEM (Gehrmann et al., 2021)) has had the objective to contribute to measuring progress in the field of Natural Language Generation (NLG), via the creation of datasets and tools for automatic and human assessments of text generation systems on different NLG tasks (McMillan-Major et al., 2021; Mille et al., 2021; Dhole et al., 2023; Gehrmann et al., 2022, 2023; Zhang et al., 2023; Nawrath et al., 2024). In the past few years, large language models (LLMs) have been widely used in NLG; they have been trained on enormous amounts of data, to the point that it can be unclear what they have seen or not during training time (Balloccu et al., 2024). To challenge these models, the NLG community has recently been developing methods for creating ad-hoc input data that the models cannot have been exposed to. For instance, Axelsson and Skantze (2023) propose to build dynamically counterfactual and fictional inputs for data-to-text generation, and Kasner and Dušek (2024) released a tool for collecting new test sets using public APIs; the creation or compilation of reference texts for the collected inputs remains an open issue.

In parallel, the interest for multilingual Natural Language Processing has been growing, with the organisation of shared tasks that included under-resourced languages, such as Universal Dependency parsing (Zeman et al., 2018) for syntactic parsing, MSR (Mille et al., 2018) for surface realisation, LowResourceEval (Klyachko et al., 2020) for morphological analysis, LowresMT (Ojha et al., 2020, 2021) and WMT (Libovický and Fraser, 2021) for machine translation, as well as WebNLG (Cripwell et al., 2023) for data-to-text generation.

Inspired by the current state of affairs, this edition of the GEM shared task[1] has two main objectives: (i) to assess LLMs—and more broadly NLG systems—using new ad-hoc datasets that no model could have already been exposed to, and (ii) to encourage participants to come up with approaches suitable across languages (including low-resource languages). We created data for two tasks, namely *data-to-text generation* and *text summarization*. The data-to-text task comprises 6 types of inputs: in-domain factual data, in-domain counterfactual data, in-domain fictional data, out-of-domain factual data, out-of-domain counterfactual data, and out-of-domain fictional data. We accepted output texts in 9 languages: Arabic, English, Chinese, German, Hindi, Korean, Russian, Spanish and Swahili; small sets of new human-written references were compiled for all 6 test sets in English and Swahili. For the summarization task, we scraped recent news articles in Swahili, extracting a summary from the web page they appeared in. The other two summarization subtasks we planned (cross-lingual summarization and book chapter summarization) did not attract participants, so we do not elaborate on them here. For all tasks,

[1] https://gem-benchmark.com/shared_task

Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, pages 17–38
September 23 –27, 2024. ©2024 Association for Computational Linguistics

we apply both automatic and human evaluation methods.

In the remainder of this paper, we present the timeline of the task and comment on the incomplete results (Section 2). We then provide an overview of the tasks and datasets involved (Section 3), followed by descriptions of the participating systems (Section 4) and the evaluation methods employed (Section 5). Finally, we present the results available at the time of publication (Section 6).

2 Timeline and status at publication time

The task was advertised in 2023 across different channels, and was officially launched on February 20^{th} 2024, when a pre-registration page was made publicly available. Every team who pre-registered their system was sent the data for the task(s) they selected, with no obligation to submit outputs. All system outputs were collected on April 11^{th} 2024. The following months were dedicated to organising the human evaluation process, and suffered multiple delays, mainly due to the fact that we took a late decision to compile new reference texts for English and Swahili (see Section 5.2).

As a result, at the time of publication of this paper, several evaluations are still ongoing. We only sent the participants the following completed evaluation results: the data-to-text metrics results for English (6 test sets, 7 systems), and the summarization metrics results for Swahili (1 test set, 2 systems). The data-to-text metrics results for Swahili (6 test sets, 3 systems), the human evaluation results for English (6 test sets, 7 systems), Swahili (6 test sets, 3 systems) and Spanish (6 test sets, 3 systems), and the summarization human evaluation results for Swahili (1 test set, 2 systems) are not yet released and are planned to be presented during the INLG conference in September 2024.

3 Overview of tasks

The GEM 2024 shared task consists of two different types of tasks: data-to-text generation and text summarization. Table 1 shows the input/output pairs for each task. Notably, no training or development data was provided to participants for either task. Given the prevalence of large language models, our primary objective was to design test data that was previously unseen by these models. To achieve this, we carefully crafted separate test sets for both the data-to-text and summarization tasks, which are described in detail in this section.

Task	Input	Output
Data-to-text	Table	Text
Summarization	Full text	Short summary

Table 1: Input/output specifications for the tasks.

3.1 Data-to-text task

The data-to-text (D2T) task consists in generating texts from input triple sets in the WebNLG fashion, where each triple is made of *Subject* | *Property* | *Object*. Figure 1 shows a sample triple set that contains 2 triples (i.e., of size 2). Both triples are about Nie Haisheng (the *Subject*); the first one states his birth date (1964-10-13), while the second one states his occupation (fighter pilot). The expected output in English would be one or two sentences such as "*Nie Haisheng is a fighter pilot born on October 13th 1964*" or "*Nie Haisheng, who was born on October 13th 1964, was a fighter pilot*".

The GEM data-to-text task contains 2 subtasks:

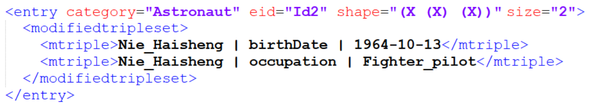
- WebNLG-based (D2T-1): We use the official WebNLG 2020 test set (Castro Ferreira et al., 2020); even though the WebNLG test set contains properties and entities not seen in the training/dev data, we consider the whole WebNLG dataset as in-domain since all splits (training/dev/test) had been available online for more than 3 years before the GEM task was launched. The dataset contains 220 different DBpedia properties and the original dataset specifications can be found on the WebNLG website.[2]

- Wikidata-based (D2T-2): We queried Wikidata to collect 1,800 triples sets containing between 2 and 7 properties for a random set of persons, following the method described in Axelsson and Skantze (2023). The dataset contains 74 different properties, none of which were in WebNLG; furthermore, almost none of the entities are in WebNLG either, so the Wikidata-based tests are considered out-of-domain.[3]

For each subtask, there are 3 parallel test sets, as proposed in Axelsson and Skantze (2023):

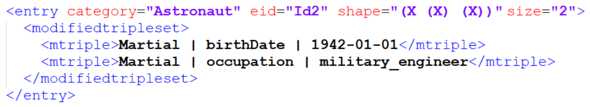
[2] https://synalp.gitlabpages.inria.fr/webnlg-challenge/challenge_2020/

[3] Note that the vocabulary of properties of DBpedia and Wikidata are different, but 17 of the 74 Wikidata properties have a direct equivalent with a DBpedia property, e.g., *Occupation*/*occupation* in Figures 1 and 3.



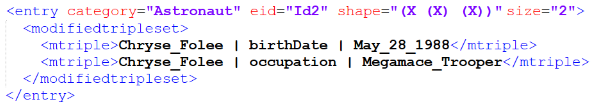
```
<entry category="Astronaut" eid="Id2" shape="(X (X) (X))" size="2">
  <modifiedtripleset>
    <mtriple>Nie_Haisheng | birthDate | 1964-10-13</mtriple>
    <mtriple>Nie_Haisheng | occupation | Fighter_pilot</mtriple>
  </modifiedtripleset>
</entry>
```

Figure 1: WebNLG Factual input (D2T-1-FA)



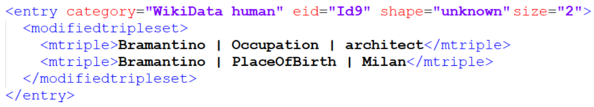
```
<entry category="Astronaut" eid="Id2" shape="(X (X) (X))" size="2">
  <modifiedtripleset>
    <mtriple>Martial | birthDate | 1942-01-01</mtriple>
    <mtriple>Martial | occupation | military_engineer</mtriple>
  </modifiedtripleset>
</entry>
```

Figure 2: WebNLG Counterfactual input (D2T-1-CFA)



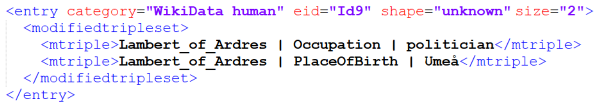
```
<entry category="Astronaut" eid="Id2" shape="(X (X) (X))" size="2">
  <modifiedtripleset>
    <mtriple>Chryse_Folee | birthDate | May_28_1988</mtriple>
    <mtriple>Chryse_Folee | occupation | Megamace_Trooper</mtriple>
  </modifiedtripleset>
</entry>
```

Figure 3: WebNLG Fictional input (D2T-1-FI)



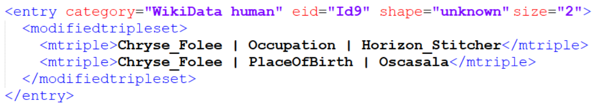
```
<entry category="WikiData human" eid="Id9" shape="unknown" size="2">
  <modifiedtripleset>
    <mtriple>Bramantino | Occupation | architect</mtriple>
    <mtriple>Bramantino | PlaceOfBirth | Milan</mtriple>
  </modifiedtripleset>
</entry>
```

Figure 4: Wikidata Factual input (D2T-2-FA)



```
<entry category="WikiData human" eid="Id9" shape="unknown" size="2">
  <modifiedtripleset>
    <mtriple>Lambert_of_Ardres | Occupation | politician</mtriple>
    <mtriple>Lambert_of_Ardres | PlaceOfBirth | Umeå</mtriple>
  </modifiedtripleset>
</entry>
```

Figure 5: Wikidata Counterfactual input (D2T-2-CFA)



```
<entry category="WikiData human" eid="Id9" shape="unknown" size="2">
  <modifiedtripleset>
    <mtriple>Chryse_Folee | Occupation | Horizon_Stitcher</mtriple>
    <mtriple>Chryse_Folee | PlaceOfBirth | Oscasala</mtriple>
  </modifiedtripleset>
</entry>
```

Figure 6: Wikidata Fictional input (D2T-2-FI)

- Factual (FA): The information in these inputs is factually correct. For the WebNLG-based task, this test set is the one used for the WebNLG 2020 shared task (Castro Ferreira et al., 2020). Figures 1 and 4 show sample inputs for the D2T-1-FA and D2T-2-FA subtasks respectively.

- Counterfactual (CFA): Entities in the factual dataset are switched based on their Wikidata class (e.g., a person entity is replaced by another person entity, a date by another date, etc.). Figures 2 and 5 show counterfactual inputs derived from Figures 1 and 4, respectively; the properties are the same as in the FA and FI datasets of the subtask (see FI below), but the Subject and Object values are replaced by other existing ones of the same category. In Figure 2, for instance, the information about Marcus Valerius Martialis, known in English as Martial, is factually wrong: Martial was a Roman poet born between 38 and 41 AD. The category feature may not match the new data, but the shape is correct as it is the same as in the original data.

- Fictional (FI): Entities in the factual datasets are replaced by made up entities (obtained via LLM prompting). Figures 3 and 6 show fictional inputs derived from Figures 1 and 4, respectively. In Figure 6 for instance, both the Subject (*Chryse_Folee*) and Object (*Oscasala* and *Horizon_Stitcher*) values are fictional; the properties are the same as in the other 2 subtask datasets (FA and CFA). There is no shape available. The same fictional name appears in the WebNLG example in Figure 3 and the Wikidata example in Figure 6—the same fictional entities may appear several times in different contexts and are not supposed to represent a coherent narrative about anything or anyone.

3.2 Summarization task

Text summarization is the task of producing a concise text sequence that captures the key information from a longer input text. The GEM summarization (Summ) task focuses on news article summarization. We follow the data collection pipeline of XL-Sum (Hasan et al., 2021) to create the task data. The articles are collected from the BBC website.[4] The summaries are extracted from the leading bold paragraph in the web pages containing the news articles, which summarizes the article's information in one or two sentences. To minimize the risk of potential data contamination, we only collect articles published between 2023 and 2024. We collect 2,978 articles in total in English, Spanish, and Swahili. Since all the submissions to the summarization task were in Swahili, we only conducted human evaluation with this subset, where 100 examples were sampled for the evaluation.

3.3 Languages

While the summarization task focused on Swahili, we encouraged submissions in multiple languages for data-to-text, namely Arabic (ar), English (en), Chinese (zh), German (de), Hindi (hi), Korean (ko), Russian (ru), Spanish (es) and Swahili (sw), and told the participants that a subset of these languages

[4] https://www.bbc.com/

Team	D2T-1	D2T-2	Summ	Languages
CUET_SSTM (Rahman et al., 2024)			x	sw
DCU-ADAPT-modPB (Osuji et al., 2024)	x			en, hi, ko, sw
DCU-NLG-PBN (Lorandi and Belz, 2024)	x	x		ar, de, en, es, hi, ko, ru, sw, zh
DCU-NLG-Small (Mille et al., 2024)	x	x		ar, de, en, es, hi, ko, ru, sw, zh
DipInfo-UniTo (Oliverio et al., 2024)	x	x		en
OSU CompLing (Allen et al., 2024)	x	x		en , es
RDFpyrealb (Lapalme, 2024)	x	x		en
SaarLST (Jobanputra and Demberg, 2024)	x	x		en

Table 2: Overview of participating systems

only would be used in the human evaluation, depending on the number of submissions for each (see Section 5.1). The inputs were exactly the same for all the output languages, that is, we did not provide DBpedia triples in Swahili to serve as input the the generation in Swahili; instead, inputs with English labels as in Figures 1 to 6 were used.

4 Participating systems

About 40 teams pre-registered, and 9 submitted outputs; one team eventually withdrew their submission. Table 2 lists the final teams and the subtask(s) and language(s) they addressed. The three DCU teams submitted multiple systems but were asked to choose a primary system for the human evaluation; for the sake of clarity we only report metrics scores for the primary systems, and point the reader to the respective system description papers in this volume for more details about non-primary submissions.

Pre-registration After handing out a preliminary survey to collect interest in the tasks and languages for the shared task, we asked all registered teams to carry out a pre-registration of their planned experiment(s). The objective of the pre-registration is to log in the details of a specific experiment before it is carried out; it is an important initial step to guarantee that the experiment is conducted fairly, and to help avoid potential biases derived from the researchers' interest (van Miltenburg et al., 2021). We asked participants to pre-register selected information (i.e. intended systems, hardware, additional data, automatic metrics, etc.) through a Qualtrics form (see Appendix E for screenshots of the form).

In the following, the summarization baseline and the team submissions are briefly described; an overview of participation to the tasks is provided in Table 2.

The Summarization baseline uses GPT-3.5 following the prompt design from Goyal et al. (2022). The specific prompt is "*Summarize the above article briefly in 1 sentence*" translated into Swahili, "*Fanya muhtasari wa kifungu kilicho hapo juu kwa kifupi katika sentensi 1.*". The system prompt is the default. All output is checked for language id to ensure that the output is in Swahili.

CUET_SSTM (Rahman et al., 2024) uses an integrated extractive-abstractive summarizer. For the extractive summarizer, the authors used the BERT Extractive Summarizer, which shortens long texts of more than 512 tokens. For the abstractive summarizer, they used two pre-trained models (T5-Small, mBART-50) to generate the summaries. The integrated model is trained on the XLSUM Swahili dataset combined with 1,000 manually summarized texts from the given Swahili news classification dataset.

DCU-ADAPT-modPB (Osuji et al., 2024) adopts an NLG+MT approach based on a pipeline neural architecture. It leverages the fine-tuned Flan-T5-large model for the ordering and structuring of input triples. Additionally, a GPT-4 prompt-based model was integrated for surface realisation, generating the final text outputs and employing few-shot prompting with five examples for the final text generation tasks in English. For multilingual text generation in Korean, Arabic, and Swahili, a prompt-based model—the Cohere-command-r-plus neural machine translation model—was incorporated, also using five examples for the translation. For Hindi, the IndicTrans2 model was used.

DCU-NLG-PBN (Lorandi and Belz, 2024) fine-tuned the Mistral 7B Instruct model, using Low-Rank Adaptation (LoRA) to enhance performance while maintaining computational efficiency. The system generates text in English, which is then translated into multiple languages (Chinese, German, Russian, Spanish, Korean, Hindi, Swahili, and Arabic) using a machine translation system (Google Translate).

DCU-NLG-Small (Mille et al., 2024) combined the FORGe rule-based generator and a post-processing step with a T5-Base model fine-tuned on a parallel dataset of English rule-based-generated texts and human- or LLM-produced texts. For languages other than English, they used the off-the-shelf machine translation system NLLB, which is freely available on HuggingFace.

DipInfo-UniTo (Oliverio et al., 2024) focuses on English and employs a three-step pipeline called the SGA (split-generate-aggregate) pipeline to generate verbalizations. The process begins with a data unit splitting phase, where the initial triples are divided into subsets of three or fewer triples, with an effort to maintain the relationships between them. The next step involves generating verbalizations for each subset of triples using Mistral-7B, which has been fine-tuned on a training and development set from WebNLG 3.0 dataset for English. Finally, in the last step, a pre-trained Mistral-7B model is used for sentence aggregation with a zero-shot prompting technique, merging the generated sentences into a more fluent and coherent text.

OSU CompLing (Allen et al., 2024) experimented with a data filtering and knowledge distillation approach for English, Spanish, Chinese, and English. They leverage the expertise of ChatGPT (GPT 4.0) to generate training data for factual, counterfactual, and fictional triple sets. Data filtering was done with automatic model judgments for error detection. Spanish and English filtered synthetic data was used to fine-tune Llama2.

RDFpyrealb (Lapalme, 2024) employs a symbolic method to address the English D2T challenge. One objective is to contrast the outcomes of computationally demanding techniques that may not always be easy to control with a streamlined, swift, and reliable symbolic method. The design is straightforward: every RDF triple represents a statement, where the subjects and objects of the triple are nearly identical to those of the sentence. The predicate in the triple represents a verb phrase that defines the sentence's syntax. The narrative-building mechanism arranges predicates sequentially, giving rise to a coherent tale. It also combines sentence components when they share the same subject or predicate. The final realization is performed using pyrealb, a French-English realizer which is used in some data to text applications.

SaarLST (Jobanputra and Demberg, 2024) employs a retrieval-augmented generation (RAG) pipeline to generate verbalization. Most RAG pipelines use a dense retriever while this pipeline contains a symbolic retriever – `PropertyRetriever`. The `PropertyRetriever` leverages available WebNLG training and validation sets to retrieve instances with the most similar properties. The retrieved examples and prompting instructions combined form the final few-shot prompt. In the final verbalization step, the pipeline prompts an ensemble of `Mixtral` and `Command-R` models to generate coherent verbalization.

5 Evaluation methods

In Section 3, we detailed the procedure for creating the inputs used in both the D2T and Summ tasks. Initially, these inputs lacked corresponding reference texts. Due to the significant time and resource investments required to create input-output pairs, we strategically delayed collecting human references until we had identified the languages submitted by participants. This section first provides an overview of the language selection and the reference text creation procedures, and then describes the automatic and human evaluations we ran on each submission to the shared task.

5.1 Selection of evaluated languages

As shown in Table 2, for the D2T task, all team submitted English outputs, 3 teams submitted Spanish outputs (DCU-NLG-PBN, DCU-NLG-Small and OSU CompLing), and 3 teams submitted Hindi, Korean and Swahili outputs (DCU-ADAPT-modPB, DCU-NLG-PBN and DCU-NLG-Small); only the two DCU-NLG teams submitted outputs for all other languages. For the Summ task, the only participating team submitted Swahili outputs. The task budget allowed for carrying out human evaluations in 3 languages, and our original plan was to include English and at least one low-resource language. We selected English and Swahili because they had the most submissions, and Spanish to include an additional team in the human evaluation of a language other than English. For English and Swahili, we carry out both automatic and human evaluations, whereas for Spanish, we rely solely on human evaluation.

5.2 Creation of new reference texts in English and Swahili

As mentioned in Section 3, the inputs for both the data-to-text and the summarization tasks have been collected specifically for the present task. Since we recruited bilingual Swahili-English speakers in person for the evaluation of Swahili texts, we also asked them to write reference texts in these two languages for all the D2T test set inputs; there are in total 1,080 input (180 inputs sampled from each of the 6 test sets, see Section 5.4.2), and one text was collected for each input.

The annotators were provided (i) a one-page document with instructions, and (ii) a document with definitions of the 211 different properties found in the sampled test sets, which we drafted ourselves.[5] One meeting with the task organisers and the annotators took place where questions could be asked, and during which the annotators collectively wrote and discussed English and Swahili texts for about 10 input tables. For each English/Swahili text pair created, each annotator received $0.5.

To collect the texts, we used a variation of the evaluation interface (see Section 5.4.4) in which instead of ratings, annotators were shown 2 boxes, one the text in each language. Packages of 12 to 18 input tables were created, and annotators (i) downloaded a package, (ii) submitted the texts for all inputs of the package, and (iii) then had the possibility to download another package not yet used by anyone. For quality control, we collected 2 annotations from different persons for 60 texts.

Due to some delays, we were not able to complete the collection of the above-mentioned texts by the time of publication of this paper. We launched a last-minute set of tasks on Amazon Mechanical Turk (AMT) and Prolific to get the English texts, using some of the English evaluators recruited as described in Section 5.4.1. These are the reference texts we use in the evaluations of the present paper; Appendix C contains a brief assessment of the quality of the collected texts.

5.3 Automatic evaluation

For the **D2T** task, we use a classic set of reference-based metrics for English and Swahili outputs, taking as reference the texts collected as described in Section 5.2; the six D2T test sets contain 180 input/reference pairs each (180 inputs, one reference per input, see Section 5.4.2). The metrics include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), chrF++ (Popović, 2017) and BERTScore (Zhang et al., 2019). To easily run the evaluation on any pair of predicted and reference files, we released a Notebook[6] largely based on the original WebNLG 2020 code.[7]

For the **Summ** task, we used the BBC automatically generated summaries following the procedure used in the XLSum task. While there were only 200 human evaluations, we use the entire 2,993 test set for the evaluation of the Swahili summarization task. These are several sentences long and provide a baseline summary. Since there were quality issues in the automatically extracted reference summaries, we performed data filtering to resolve these issues, which resulted in 1,367 examples in total. The metrics include ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021).

5.4 Human evaluation

In addition to the automatic evaluations, we also asked human raters to evaluate a subset of the outputs from each submission to the shared task. In this section, we provide details on the evaluator recruitment and training processes, the data sampling, and the evaluation criteria and task design.

5.4.1 Recruitment and training of evaluators

To ensure alignment between the recruited evaluators and the D2T task, we designed a qualification task that consisted of five rating checks and one attention check. For each rating check a handcrafted tabular set of data predicates was presented alongside a text generated from the table. In four of the five rating checks, the text presented contained deliberate errors such as issues with fluency, grammatically, omissions, and additions. Evaluators were asked to assess each text on a 7-point Likert scale on four quality criterion: *fluency*, *grammatically*, *no-additions*, *no-omissions*. In the case of rating checks with deliberate errors for specific quality criterion, evaluators were expected to rate these criteria either neutral (4-rating) and/or lower than neutral. Unaffected aspects were to be rated as higher than neutral. The fifth rating check contained no issues, so evaluators were expected to rate all quality criteria neutral or above.

For the recruitment of English and Spanish evaluators in the D2T task, we used Zhang et al.'s (2023) *qualification task*, where the evaluator is expected to successfully complete a task after receiving a short training. We recruited 23 evaluators in English (15%) pass rate) and 13 in Spanish (22% pass rate) respectively on MTurk and Prolific.

[5] https://github.com/mille-s/GEM24_D2T_StratifiedSampling/tree/main/documents

[6] https://github.com/mille-s/WebNLG-2020_Metrics

[7] https://github.com/WebNLG/GenerationEval.git

Task	Criterion name	Quality type	Frame of reference	Aspect
Table to text	No-Omissions	Correctness	Relative to input	Content
	No-Additions	Correctness	Relative to input	Content
	Grammaticality	Correctness	Output in its own right	Form
	Fluency	Goodness	Output in its own right	Form and Content
Summarization	Understandability	Goodness	Output in its own right	Form and Content
	Faithfulness	Correctness	Relative to input	Content
	Saliency	Goodness	Relative to input	Content
	Grammaticality	Correctness	Output in its own right	Form
	Coherence	Goodness	Output in its own right	Content
	Compactness	Goodness	Output in its own right	Content

Table 3: Properties of our criteria according to the taxonomy by Belz et al. (2020).

On the other hand, recruiting evaluators from low-resource languages (Swahili in our case) on crowdsourced platforms is more challenging. Thus, for both tasks, we recruited 14 students who are Swahili native speakers from the Technical University of Kenya and Moi University. To help these students understand the task, we (i) set up meetings to explain the task in detail, (ii) carried out a few tasks together, and (iii) formed a Google group for questions and discussion.

5.4.2 Data sampling and packaging

For the **D2T** task, we selected 180 data points (∼10%) from each of the six test sets (D2T-1-FA, D2T-1-CFA, D2T-1-FI, D2T-2-FA, D2T-2-CFA, D2T-2-FI, see Section 3), stratifying only by input size and excluding inputs of size 1, which are usually trivial to generate from. Thus, each of the six test sets contains 30 inputs for each input size, ranging from 2 to 7. This allows us to analyze the metrics results broken down by input size. The code for sampling and creating the corresponding pairs of HTML tables and system outputs as used in the human evaluation is available on GitHub.[8]

Once sampled, the input/output pairs were packaged to be sent to the evaluators. For Swahili, we created 75 packages of 36 input/output pairs. For Spanish, we created 270 packages of 12 input/output pairs. For English, we created 1,080 packages of 7 - 8 input/output pairs. The packages for English and Spanish are substantially smaller that those for Swahili because the evaluators for these two languages were recruited on Amazon Mechanical Turk, where proposed tasks are usually short. The Swahili evaluators were recruited in person and could be trusted to complete larger packages.[9]

The **Summ** outputs were not sampled nor packaged at the time this paper was written.

5.4.3 Quality criteria

The criteria used for the evaluation should capture aspects of the quality of the meaning and form. Table 3 lists the criteria used in both tasks and lists their properties according to Belz et al.'s (2020) taxonomy.

D2T task Our selection of criteria (see Table 4) reflects closely the evaluations carried out in the context of some recent data-to-text shared tasks such as WebNLG (Cripwell et al., 2023) or E2E (Dušek et al., 2020). We evaluated four dimensions, namely whether or not the text represents faithfully the input table (*No-Omissions*, *No-Additions*), whether or not the text contains grammatical errors (*Grammaticality*), and whether or not the output text flows well on its own (*Fluency*).

Criterion name	Definition
No-Omissions	ALL the information in the table is present in the text.
No-Additions	ONLY information from the table is present in the text.
Grammaticality	The text is free of grammatical and spelling errors.
Fluency	The text flows well and is easy to read; its parts are connected in a natural way.

Table 4: Criteria used for data-to-text generation

Summ task The objective of the evaluation is to assess the quality of a summary given an input text. The summaries are evaluated along the dimensions defined in Zhang et al. (2023), shown in Table 5 with their respective definitions. The objective of

[8] https://github.com/mille-s/GEM24_D2T_StratifiedSampling. Thanks to Liam Cripwell and Michel Lorandi for making the WebNLG 2023 sampling code available, which we used as a starting point.

[9] The Swahili packages represent about one hour of work; we tried packages of the same size on Mechanical Turk and received complaints from Turkers that the tasks were too long.

Criterion name	Definition
Understandability	Can the worker understand the summary and is the summary worth being annotated.
Faithfulness	All of the information in the summary can be found in the article; the summary accurately reflects the contents of the article.
Saliency	The summary captures the most important information of the article and does not include parts of the article that are less important.
Grammaticality	The summary is free of grammatical and spelling errors.
Coherence	The summary is presented in a clear, well-structured, logical, and meaningful way.
Compactness	The summary does not contain duplicated information.

Table 5: Criteria used for summarization

the first criterion, *Understandability*, is to give the annotator a chance to not provide the ratings for the rest of the criteria in case the quality of the text does not allow for it. Two criteria (*Faithfulness* and *Saliency*) require the evaluators to compare the summary with the input, while the remaining three (*Grammaticality*, *Coherence*, *Compactness*) capture intrinsic qualities of the summary. Two criteria are highly specific to the summarization task, namely *Saliency* and *Compactness*, which aim at capturing respectively whether the main points of the original text were captured, and whether the resulting summary is indeed compact and does not contain unnecessary repetitions.

5.4.4 Survey Design

We designed evaluation surveys for data-to-text and summarization using HTML, CSS, and Jinja. We launched our survey on Amazon Mechanical Turk and Prolific. For all tasks, evaluators were shown the input and one output (see Table 1). For all criteria, direct assessment was used, and the answers were collected using a labeled 7-point scale (see Figure 7). The evaluation interfaces are shown in Figures 8 and 9 in Appendix A.

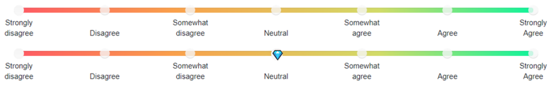


Figure 7: Rating Scale (7-point)

Designing an effective survey requires an understanding of the subject matter and awareness of potential biases that could compromise validity, and we drew our inspiration from HCI research practices (Müller and Sedley, 2015). We aimed to create a reliable and impactful survey by minimising biases and tailoring each aspect to elicit meaningful, accurate responses. See Appendix B for more discussion on the choices behind the survey design.

6 Results

In this section, we present the results of the metrics evaluation for the English data-to-text task and the Swahili summarization task.

6.1 Metrics results for the D2T task

Table 8 shows the **BLEU**, **METEOR**, **chrF++** and **BERT's F1** scores of all primary systems on the three D2T-1 and the three D2T-2 test sets respectively (FA = Factual, CFA = Counterfactual, FI = Fictional, see Section 3) for the English language. For all the results broken down by input size, see the plots in Appendix D. As mentioned above, for calculating the scores in Table 8, we use the references created by our AMT-recruited annotators (see Section 5.2). For comparison, we also report here the scores obtained with the entire WebNLG test set (1,779 texts) and all the WebNLG references (Table 6), and the scores obtained with the same set of 180 data points as in Table 8, but selecting only one random WebNLG reference when more than one is available.[10]

Comparison between the D2T-1 and the D2T-2 subtasks. For all six systems that participated in both subtasks, the scores are substantially higher for the D2T-1 task than for the D2T-2 for the factual (FA) and fictional (FI) datasets, but, surprisingly, not for the counterfactual (CFA) dataset, where scores are always higher in the D2T-2 subtask. For DCU-NLG-PBN, DipInfo-UniTo, OSU-CompLing and SaarLST (i.e. all submissions that are not primarily based on a rule-based system), BERTScore is even equal or higher for all 3 datasets of the D2T-2 task.

D2T-1 scores. All seven submissions obtained a (generally substantially) higher score for all metrics on the factual (FA) dataset, which was expected since this is the only dataset for which reference texts were available when the task was running. For all seven submissions, BERT systematically

[10] The number of references used can affect the scores of some metrics, for example, BLEU.

System ID	BLEU ↑	METEOR ↑	chrF++ ↑	BERT F1 ↑
DCU-ADAPT-modPB	49.8	0.400	0.655	0.955
DCU-NLG-PBN	**52.26**	**0.410**	0.679	**0.956**
DCU-NLG-Small	51.43	0.395	0.662	0.954
DCU-NLG-Small-noT5	40.55	0.372	0.620	0.943
DipInfo-UniTo	51.36	**0.410**	**0.681**	0.955
OSU CompLing	43.09	0.389	0.65	0.950
RDFpyrealb	42.38	0.390	0.642	0.946
SaarLST	39.86	0.400	0.655	0.947

Table 6: Metrics scores on the D2T-1-FA English test set using all WebNLG data points (1,779) and all reference texts (2.5 texts per data point on average).

System ID	BLEU ↑	METEOR ↑	chrF++ ↑	BERT F1 ↑
DCU-ADAPT-modPB	28.27	0.338	0.561	0.936
DCU-NLG-PBN	**32.5**	**0.356**	**0.6**	**0.937**
DCU-NLG-Small	29.17	0.337	0.571	0.933
DipInfo-UniTo	30.47	0.348	0.585	0.93
OSU CompLing	27.01	0.339	0.575	0.931
RDFpyrealb	26.26	0.339	0.567	0.927
SaarLST	25.61	0.354	0.59	0.931

Table 7: Metrics scores on the D2T-1-FA English test set using the 180 data points of the human evaluation and 1 randomly selected WebNLG reference text per data point.

		D2T-1			**D2T-2**		
	System	**FA**	**CFA**	**FI**	**FA**	**CFA**	**FI**
BLEU ↑	DCU-ADAPT-modPB	30.78	26.98	26.54	n/a	n/a	n/a
	DCU-NLG-PBN	29.08	25.2	26.02	23.96	30.34	20.46
	DCU-NLG-Small	27.0	22.98	20.85	19.48	24.9	16.88
	DipInfo-UniTo	**32.31**	**29.01**	**28.24**	27.22	**32.01**	**21.26**
	OSU CompLing	30.03	24.45	21.44	24.97	27.06	16.9
	RDFpyrealb	26.37	21.67	21.97	19.97	25.05	16.28
	SaarLST	29.7	23.48	20.76	**28.25**	26.47	20.16
METEOR ↑	DCU-ADAPT-modPB	0.332	0.299	0.318	n/a	n/a	n/a
	DCU-NLG-PBN	0.33	0.297	0.322	0.295	0.348	0.3
	DCU-NLG-Small	0.314	0.279	0.292	0.26	0.3	0.267
	DipInfo-UniTo	0.346	**0.315**	**0.342**	0.304	0.354	0.307
	OSU CompLing	0.335	0.293	0.306	0.295	0.334	0.282
	RDFpyrealb	0.331	0.291	0.31	0.287	0.335	0.286
	SaarLST	**0.347**	0.307	0.331	**0.32**	**0.359**	**0.315**
chrF++ ↑	DCU-ADAPT-modPB	0.555	0.515	0.539	n/a	n/a	n/a
	DCU-NLG-PBN	0.555	0.513	0.549	0.49	0.581	0.49
	DCU-NLG-Small	0.537	0.488	0.507	0.438	0.51	0.442
	DipInfo-UniTo	0.58	**0.543**	**0.587**	0.512	0.592	0.502
	OSU CompLing	0.566	0.514	0.537	0.496	0.567	0.475
	RDFpyrealb	0.551	0.495	0.527	0.479	0.561	0.472
	SaarLST	**0.581**	0.524	0.557	**0.538**	**0.597**	**0.518**
BERT F1 ↑	DCU-ADAPT-modPB	**0.935**	0.924	0.921	n/a	n/a	n/a
	DCU-NLG-PBN	0.933	0.923	0.92	0.936	**0.937**	**0.924**
	DCU-NLG-Small	0.93	0.918	0.914	0.925	0.923	0.914
	DipInfo-UniTo	0.933	**0.926**	**0.924**	**0.937**	0.936	0.923
	OSU CompLing	0.932	0.92	0.915	0.934	0.93	0.917
	RDFpyrealb	0.928	0.918	0.917	0.921	0.923	0.916
	SaarLST	0.931	0.921	0.917	0.934	0.929	0.919

Table 8: Metrics scores for the English D2T task (180 data points, 1 AMT reference text per data point).

sores the counterfactual (CFA) texts higher than the fictional (FI) texts, while METEOR and chrF++ exhibit the opposite behaviour. BLEU behaves very similarly to BERT.

D2T-2 scores. For all systems except SaarLST, the scores for all metrics on the counterfactual dataset (CFA) are higher than for the other two datasets (FA, FI); for these systems, only BERTScore sometimes gets slightly higher scores for Factual (FA) datasets. BLEU and BERT usually score FA texts clearly higher than fictional (FI) ones, while for METEOR and chrF++, FA and FI texts receive very similar scores.

From the perspective of system submissions, DipInfo-UniTo scores comparatively high for all metrics on all datasets. DCU-NLG-PBN and DipInfo-UniTo seem to degrade less than other systems when comparing the FA scores to the CFA and FI scores for D2T-1; for D2T-2, the submissions using rule-based components (DCU-NLG-Small and RDFpyrealb) have less drop than others from FA to FI (these two also have comparable scores overall). SaarLST seems to be the system that suffers the least when exposed to the out-of-domain data (D2T-2). When comparing the results on the AMT references (Table 8) and the ones with the WebNLG references (Tables 6 and 7), one can note that SaarLST for instance obtains higher scores on the D2T-1-FA dataset with AMT references than on the dataset with WebNLG references, while DCU-NLG-Small, which used a component fine-tuned using BLEU on the WebNLG dataset, obtains higher scores with WebNLG references than with AMT references.

At this point, and without the results of the human evaluation, it is unclear to what extent all the score differences mentioned above are due to the properties of the inputs and outputs, or to some features of the reference texts. A more in-depth analysis of the results will be provided at a later stage along with the human evaluation results.

6.2 Metrics results for the summarization task

We use ROUGE, BARTScore, and BERTScore for the automatic evaluation of the summarization system. For BARTScore, we use a multilingual BART checkpoint introduced in Tang et al. (2020).[11] Similarly, we use a multilingual BERT checkpoint[12] for BERTScore. Apart from the submitted system, we also evaluate a strong baseline that prompts GPT-3.5[13] to generate summaries with one sentence (see Section 4).

The evaluation results are reported in Table 9. CUET_SSTM achieves better performance in ROUGE scores, while GPT-3.5 achieves a higher BARTScore. Regarding BERTScore, CUET_SSTM achieved a higher recall score but a lower precision score, which is correlated with the fact that the average summary length of CUET_SSTM is much smaller. We note that since GPT-3.5's summaries are generated in a zero-shot manner, comparing its summaries using reference-based evaluation metrics may not always be accurate (Goyal et al., 2022; Liu et al., 2023). However, these results indicate that CUET_SSTM is able to achieve a relatively strong performance under the reference-based evaluation.

System	R1	R2	BARTS.	BERTS.	Len.
GPT-3.5	27.12	10.42	-6.305	69.33/73.18	31.10
CUET_SSTM	29.33	15.87	-6.791	71.05/71.37	19.59

Table 9: Automatic evaluation results of the submitted summarization system and the basline. R1 and R2 are ROUGE-1 and ROUGE-2 respectively. BARTS. and BERTS. are BARTScore and BERTScore. Len. is the average number of words in summaries. For BERTScore, we report both the precision/recall scores.

7 Conclusions

We presented an overview of the two tasks of the 2024 GEM shared task, multilingual data-to-text generation and news article summarization in Swahili. For both tasks, we collected new data with the objective provide challenging inputs to the large language models that we supposed most teams were going to use. For the data-to-text task, 7 teams submitted outputs in one or more languages, and we report on the metrics evaluation for English outputs only. The results of the evaluation show that despite the variety of system types (LLMs, rule-based, combination of the two), all systems seem to suffer when exposed to (i) out-of-domain data, and (ii) counterfactual or fictional data. The unexpectedly high scores obtained by all systems on the counterfactual out-of-domain dataset remain to be explained, possibly in the light of the human evaluation results. For the summarization task in Swahili, we received only one submission, which is competitive with a zero-shot GPT-3.5 baseline according to the metrics evaluation.

We were not able to complete all evaluations at the time the paper is published, and the data-to-text metrics results for Swahili, the human evaluation results for English, Swahili and Spanish, and the summarization human evaluation results for Swahili will be reported in a separate publication.

Acknowledgements

We thank Google for funding our crowdsourcing annotations. Mille's contribution was funded by the European Union under the Marie Skłodowska-Curie grant agreement No 101062572 (M-FleNS).

[11] https://huggingface.co/facebook/mbart-large-50

[12] https://huggingface.co/google-bert/bert-base-multilingual-cased

[13] https://platform.openai.com/docs/models/gpt-3-5-turbo

References

Mohd Azry Abdul Malik, Muhammad Firdaus Mustapha, Norafefah Mohamad Sobri, Nor Fatihah Abd Razak, Mohamad Nurifaizal Mohd Zaidi, Ahmad Aizat Shukri, and Muhammad Amir Luqman Zalimie Sham. 2021. Optimal reliability and validity of measurement model in confirmatory factor analysis: Different likert point scale experiment. *Journal of Contemporary Issues and Thought*, 11(1):105–112.

Alyssa Allen, Ash Lewis, Yi-Chien Lin, Tomiris Kaumenova, and Mike White. 2024. OSU Compling at the GEM'24 data-to-text task. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language generation from knowledge graphs. *arXiv preprint arXiv:2307.07312*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, William Soto-Martinez, et al. 2023. The 2023 webnlg shared task on low resource languages overview and evaluation results (webnlg 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2023. Nl-augmenter: A framework for task-sensitive natural language augmentation.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123–156.

Ron Garland. 1991. The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, 2.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv

Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R. Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Mayank Jobanputra and Vera Demberg. 2024. TeamsaarLST at the GEM'24 data-to-text task: Revisiting symbolic retrieval in the LLM-age. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Zdeněk Kasner and Ondřej Dušek. 2024. Beyond reference-based metrics: Analyzing behaviors of open llms on data-to-text generation. *arXiv preprint arXiv:2401.10186*.

Elena Klyachko, Alexey Sorokin, Natalia Krizhanovskaya, Andrew Krizhanovsky, and Galina Ryazanskaya. 2020. Lowresourceeval-2019: a shared task on morphological analysis for low-resource languages. *Preprint*, arXiv:2001.11285.

Guy Lapalme. 2024. RDFPYREALB at the GEM'24 data-to-text task: Symbolic english text generation from RDF triples. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Michela Lorandi and Anya Belz. 2024. DCU-NLG-PBN at the GEM'24 data-to-text task: Open-source LLM PEFT-Tuning for effective data-to-text generation. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (SR'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.

Simon Mille, Kaustubh D. Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. *ArXiv*, abs/2106.09069.

Simon Mille, Mohammed Sabry, and Anya Belz. 2024. DCU-NLG-Small at the GEM'24 data-to-text task: Rule-based generation and post-processing with T5-base. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Hendrik Müller and Aaron Sedley. 2015. Designing surveys for hci research. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 18.

Marcel Nawrath, Agnieszka Nowak, Tristan Ratz, Danilo Walenta, Juri Opitz, Leonardo Ribeiro, João Sedoc, Daniel Deutsch, Simon Mille, Yixin Liu, Sebastian Gehrmann, Lining Zhang, Saad Mahamood, Miruna Clinciu, Khyathi Chandu, and Yufang Hou. 2024. On the role of summary content units in text summarization evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 272–281, Mexico City, Mexico. Association for Computational Linguistics.

Atul Kr. Ojha, Chao-Hong Liu, Katharina Kann, John Ortega, Sheetal Shatam, and Theodorus Fransen. 2021. Findings of the loresmt 2021 shared task on covid and sign language for low-resource languages. *Preprint*, arXiv:2108.06598.

Atul Kr. Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Suzhou, China. Association for Computational Linguistics.

Michael Oliverio, Pier Felice Balestrucci, Alessandro Mazzei, and Valerio Basile. 2024. DipInfo-UniTo at the GEM'24 data-to-text task: Augmenting LLMs with the split-generate-aggregate pipeline. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Colm O'Muircheartaigh, Jon A. Krosnick, and Armen Helic. 2000. Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. Working Papers 0103, Harris School of Public Policy Studies, University of Chicago.

Chinonso Cynthia Osuji, Rudali Huidrom, Kolawole John Adebayo, Thiago Castro Ferreira, and Brian Davis. 2024. DCU-ADAPT-modPB at the GEM'24 data-to-text generation task: Model hybridisation for pipeline data-to-text natural language generation. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. 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 of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Samia Rahman, Momtazul Arefin Labib, Hasan Murad, and Udoy Das. 2024. CUET_SSTM at the GEM'24 summarization task: Integration of extractive and abstractive method for long text summarization in Swahili language. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Screenshots of evaluator interface

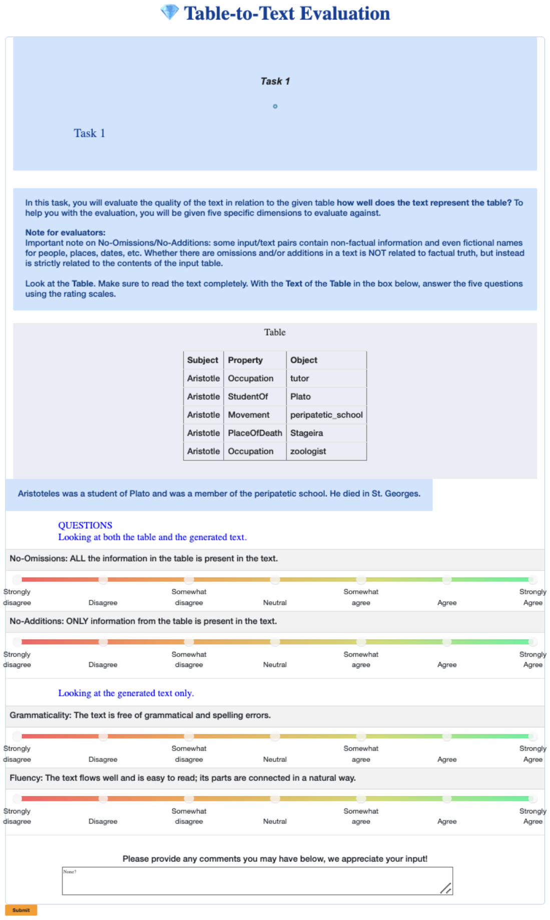


Table-to-Text Evaluation

Task 1

Task 1

In this task, you will evaluate the quality of the text in relation to the given table **how well does the text represent the table?** To help you with the evaluation, you will be given five specific dimensions to evaluate against.

Note for evaluators:
Important note on No-Omissions/No-Additions: some input/text pairs contain non-factual information and even fictional names for people, places, dates, etc. Whether there are omissions and/or additions in a text is NOT related to factual truth, but instead is strictly related to the contents of the input table.

Look at the **Table**. Make sure to read the text completely. With the **Text** of the **Table** in the box below, answer the five questions using the rating scales.

Table

Subject	Property	Object
Aristotle	Occupation	tutor
Aristotle	StudentOf	Plato
Aristotle	Movement	peripatetic_school
Aristotle	PlaceOfDeath	Stageira
Aristotle	Occupation	zoologist

Aristoteles was a student of Plato and was a member of the peripatetic school. He died in St. Georges.

QUESTIONS
Looking at both the table and the generated text.

No-Omissions: ALL the information in the table is present in the text.

No-Additions: ONLY information from the table is present in the text.

Looking at the generated text only.

Grammaticality: The text is free of grammatical and spelling errors.

Fluency: The text flows well and is easy to read; its parts are connected in a natural way.

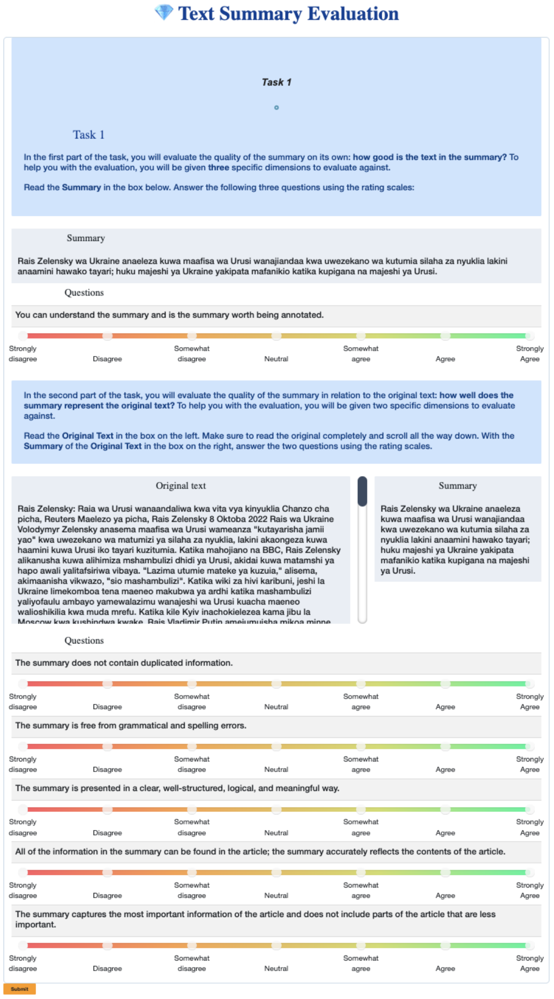
Please provide any comments you may have below, we appreciate your input!

None?

Figure 8: Data-to-text UI

B Justification of the survey design

A 7-point Likert scale offers respondents a broader range of options, enabling evaluators to express their opinions with greater nuance and precision. This expanded scale reduces the likelihood that respondents will default to a middle option out of uncertainty, thereby enhancing the accuracy of the data collected. By providing more choices, a 7-point scale allows for a more accurate reflection of respondents true feelings. Research has demonstrated that increasing the number of points on a Likert scale not only improves the reliability of the data but also reduces the potential for random error (Abdul Malik et al., 2021). On the other hand, closed-ended questions can introduce biases that may affect the data. For instance, the phrasing of questions, the order of response options, and the inclusion of a neutral midpoint can all influence how respondents' interpret and answer questions.



Text Summary Evaluation

Task 1

Task 1

In the first part of the task, you will evaluate the quality of the summary on its own: **how good is the text in the summary?** To help you with the evaluation, you will be given **three** specific dimensions to evaluate against.

Read the **Summary** in the box below. Answer the following three questions using the rating scales:

Summary

Rais Zelensky wa Ukraine anaeleza kuwa maafisa wa Urusi wanajiandaa kwa uwezekano wa kutumia silaha za nyuklia lakini anaamini hawako tayari; huku majeshi ya Ukraine yakipata mafanikio katika kupigana na majeshi ya Urusi.

Questions

You can understand the summary and is the summary worth being annotated.

In the second part of the task, you will evaluate the quality of the summary in relation to the original text: **how well does the summary represent the original text?** To help you with the evaluation, you will be given two specific dimensions to evaluate against.

Read the **Original Text** in the box on the left. Make sure to read the original completely and scroll all the way down. With the **Summary** of the **Original Text** in the box on the right, answer the two questions using the rating scales.

Original text

Rais Zelensky: Raia wa Urusi wanaandaliwa kwa vita vya kinyuklia Chanzo cha picha, Reuters Maelezo ya picha, Rais Zelensky 8 Oktoba 2022 Rais wa Ukraine Volodymyr Zelensky anasema maafisa wa Urusi wameanza "kutayarisha jamii yao" kwa uwezekano wa matumizi ya silaha za nyuklia, lakini akaongeza kuwa haamini kuwa Urusi iko tayari kuzitumia. Katika mahojiano na BBC, Rais Zelensky alikanusha kuwa alihimiza mshambulizi dhidi ya Urusi, akidai kuwa matamshi ya hapo awali yalitafsiriwa vibaya. "Lazima utumie mateke ya kuzuia," alisema, akimaanisha vikwazo, "sio mashambulizi". Katika wiki za hivi karibuni, jeshi la Ukraine limekomboa tena maeneo makubwa ya ardhi katika mashambulizi yaliyofaulu ambayo yamewalazimu wanajeshi wa Urusi kuacha maeneo walioshikilia kwa muda mrefu. Katika kile Kyiv inachokielezea kama jibu la Moscow kwa kushindwa kwake. Rais Vladimir Putin ameiumuisha mikoa minne

Summary

Rais Zelensky wa Ukraine anaeleza kuwa maafisa wa Urusi wanajiandaa kwa uwezekano wa kutumia silaha za nyuklia lakini anaamini hawako tayari; huku majeshi ya Ukraine yakipata mafanikio katika kupigana na majeshi ya Urusi.

Questions

The summary does not contain duplicated information.

The summary is free from grammatical and spelling errors.

The summary is presented in a clear, well-structured, logical, and meaningful way.

All of the information in the summary can be found in the article; the summary accurately reflects the contents of the article.

The summary captures the most important information of the article and does not include parts of the article that are less important.

Figure 9: Text summarization UI

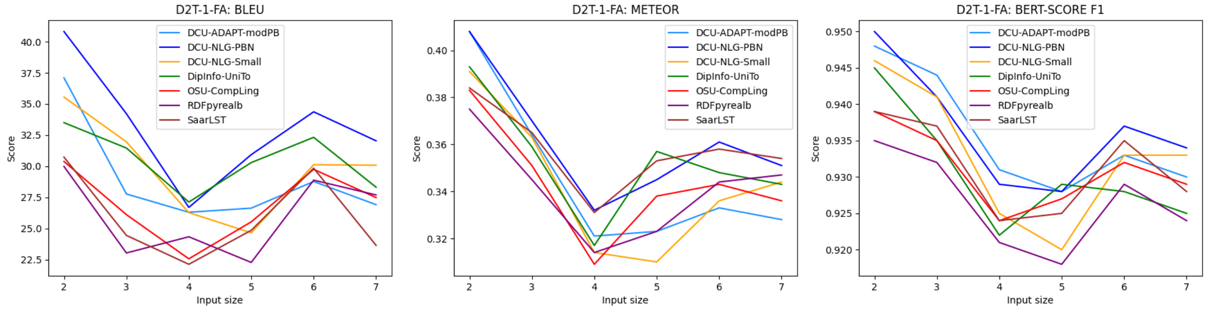
Garland (1991) examined the impact of including or excluding a neutral midpoint on a Likert scale in surveys and found that removing the midpoint can reduce social desirability bias but may push respondents toward more extreme ratings, potentially distorting results. This highlights the need to carefully consider the inclusion of a midpoint, as it can significantly influence survey outcomes. Later, O'Muircheartaigh et al. (2000) found that offering a middle alternative reduces random measurement error, increasing the reliability of responses without affecting validity. Contrary to concerns, their study suggests that including a midpoint improves data quality and does not increase acquiescence bias. Therefore, we decided to include the midpoint as “Neutral” 💎 , as presented in Figure 7.

C Informal assessment of the quality of English texts collected on AMT

While collecting texts on AMT, the authors applied manual and automatic filters. When the 1,080 (180*6) final texts were collected, one of the authors of the present paper selected randomly about 10 texts for each of the 6 datasets (60 texts in total), and checked whether or not the texts were adequately verbalising their respective input table. For 20 of these texts (1/3), some problems were detected, such as omissions, nonsensical contents, pasting irrelevant text, additions, or inaccurate verbalisation of some triples (e.g. inversions of Subject and Object or wrong semantics of the property). Additions are being noticed in particular (but not only) on the factual data, suggesting that some workers used language models to create the texts despite clear instructions not to do so. The rest of these texts (2/3) were judged of excellent quality.

D English D2T metrics evaluation broken down by input size

Figures 11 to 16 show the plots of the results in Table 8, broken down by input size (from size 2 to size 7). Figure 10 shows the same using the WebNLG references (i.e. for the D2T-1-FA dataset), for comparison.

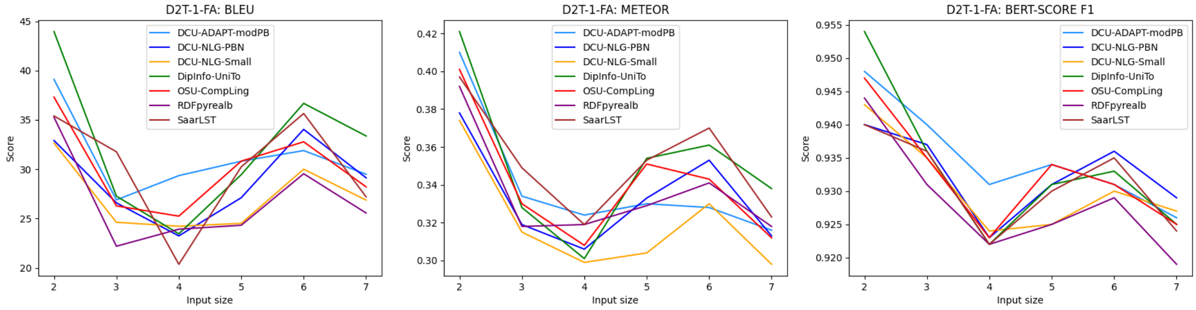


(a) BLEU

(b) METEOR

(c) BERTScore-F1

Figure 10: Metrics scores per input size (D2T-1-FA) using one randomly selected original WebNLG reference for the 180 sampled data points used in the human evaluation.

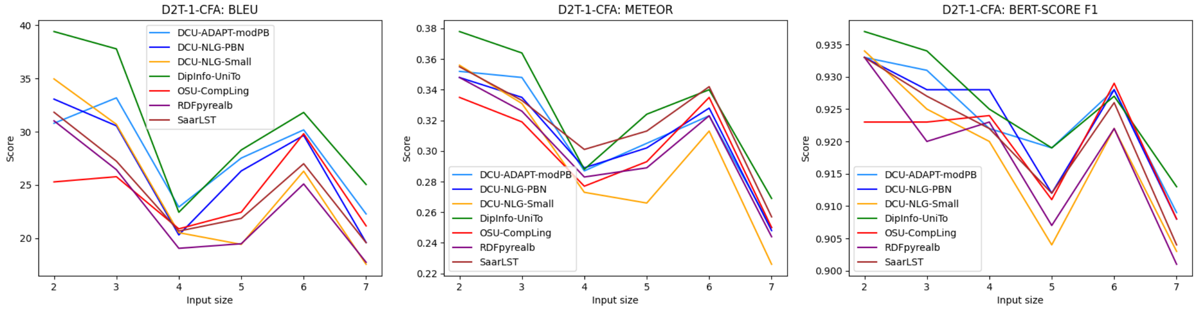


(a) BLEU

(b) METEOR

(c) BERTScore-F1

Figure 11: Metrics scores by input size on the D2T-1-FA English task (1 AMT reference text per data point)

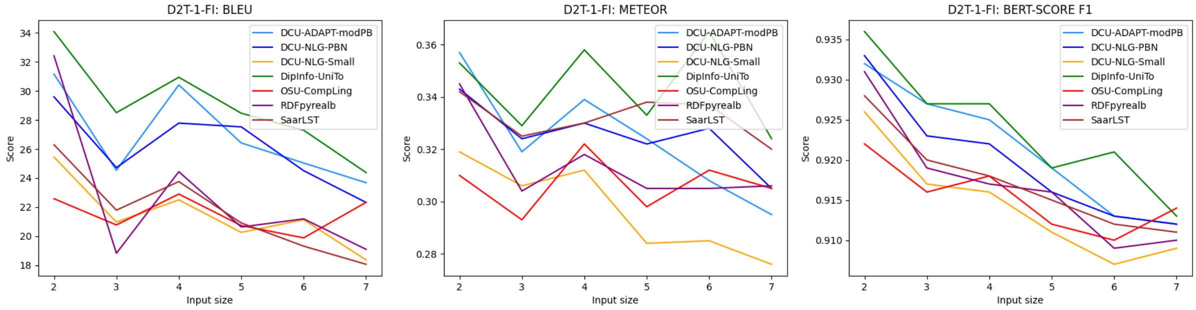


(a) BLEU

(b) METEOR

(c) BERTScore-F1

Figure 12: Metrics scores by input size on the D2T-1-CFA English task (1 AMT reference text per data point)

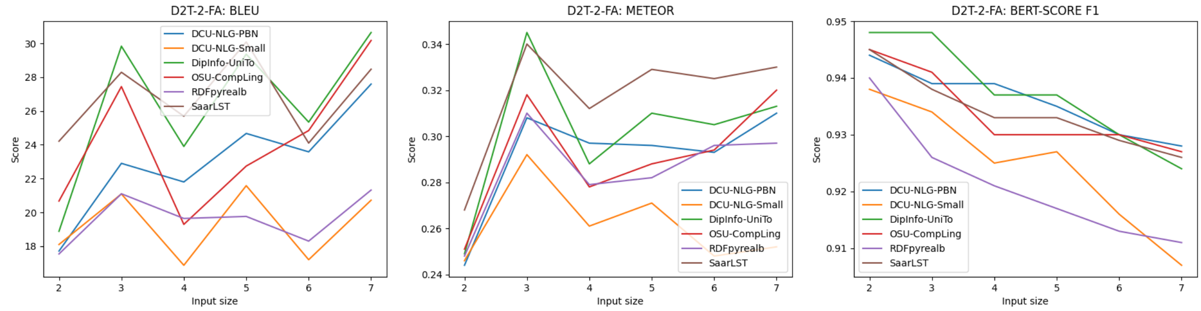


(a) BLEU

(b) METEOR

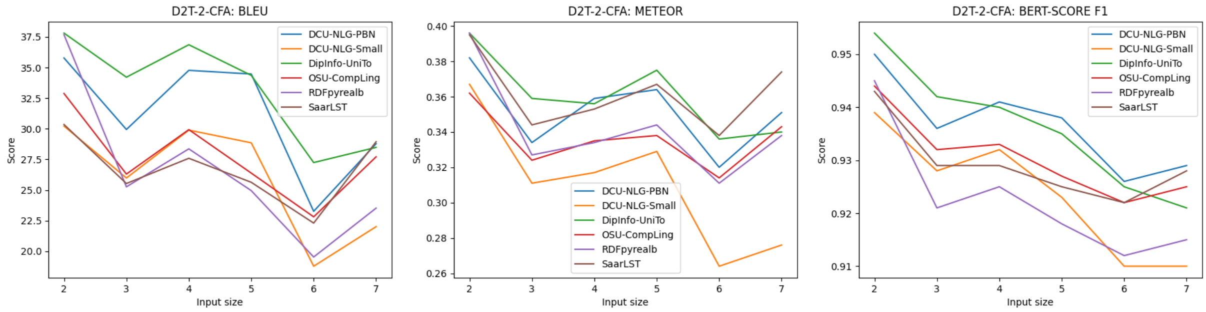
(c) BERTScore-F1

Figure 13: Metrics scores by input size on the D2T-1-FI English task (1 AMT reference text per data point)



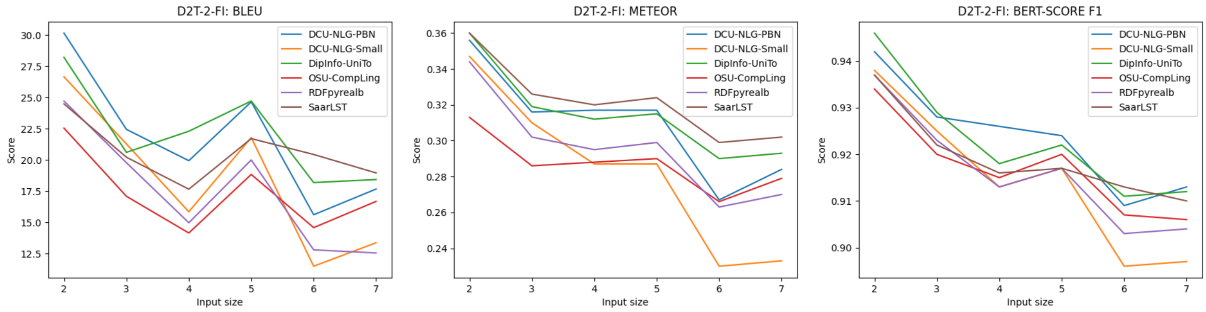
(a) BLEU (b) METEOR (c) BERTScore-F1

Figure 14: Metrics scores by input size on the D2T-2-FA English task (1 AMT reference text per data point)



(a) BLEU (b) METEOR (c) BERTScore-F1

Figure 15: Metrics scores by input size on the D2T-2-CFA English task (1 AMT reference text per data point)

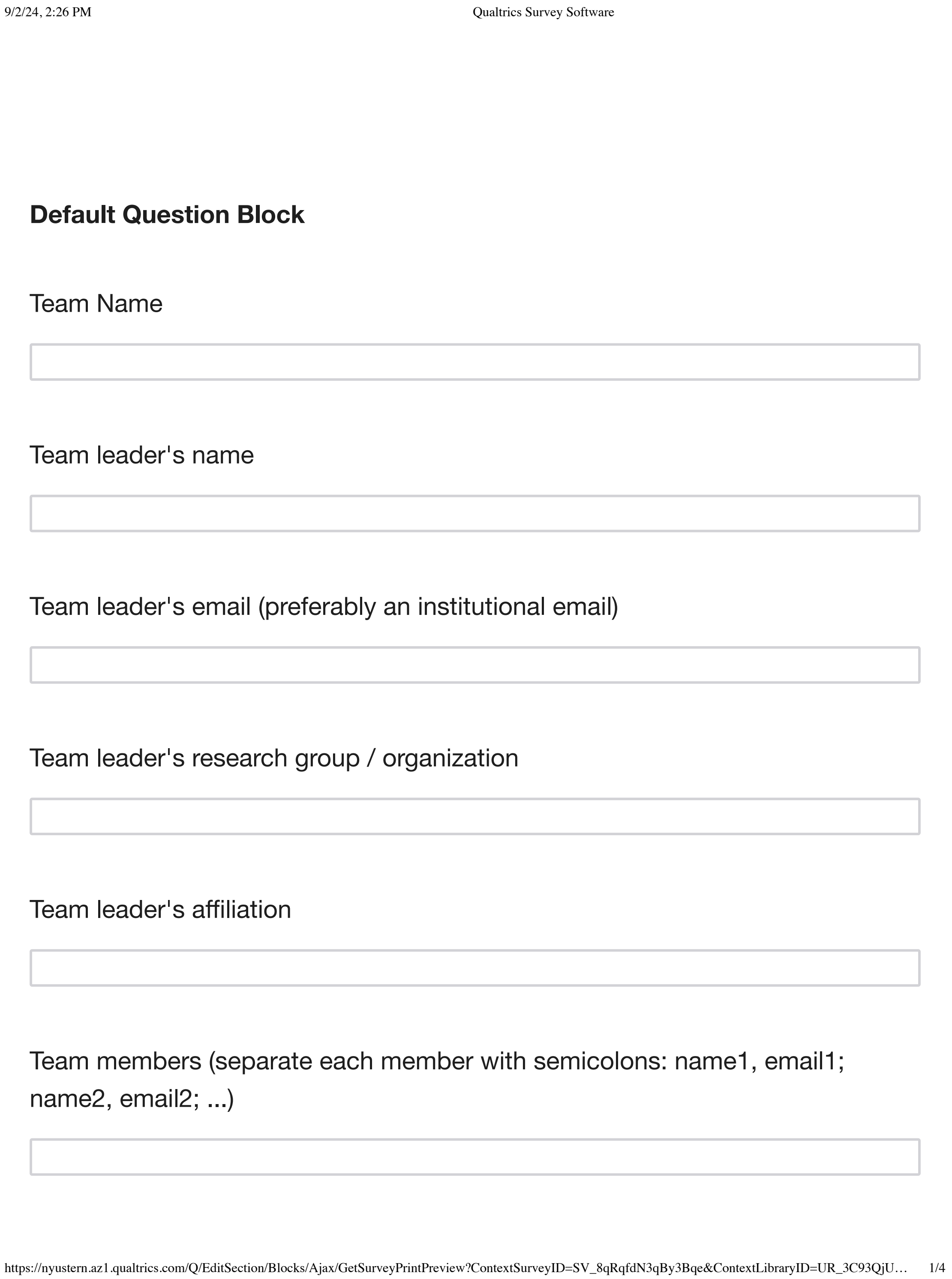


(a) BLEU (b) METEOR (c) BERTScore-F1

Figure 16: Metrics scores by input size on the D2T-2-FI English task (1 AMT reference text per data point)

E Pre-registration Form

For details about the pre-registration form, please see the file below.



Default Question Block

Team Name

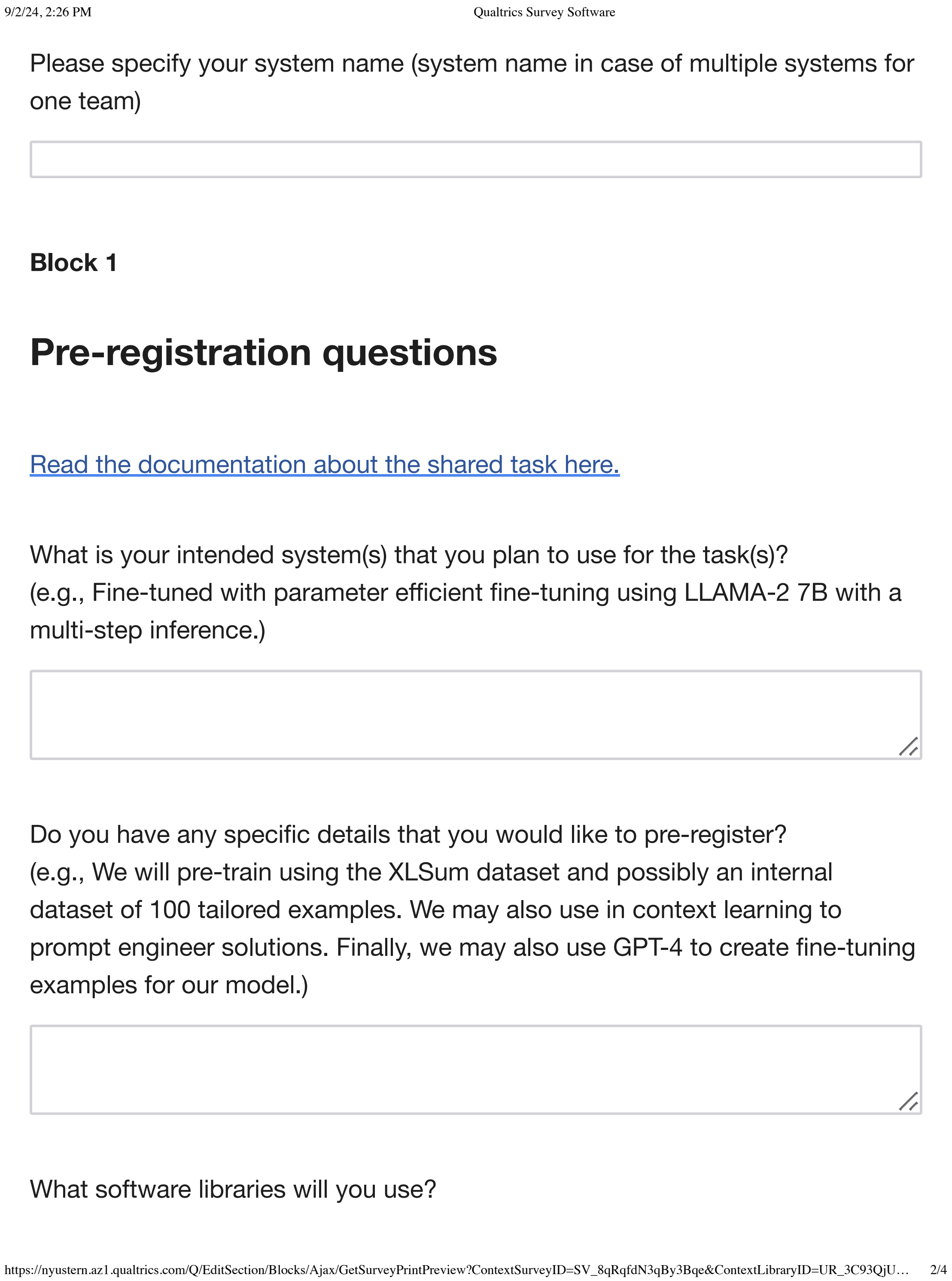
Team leader's name

Team leader's email (preferably an institutional email)

Team leader's research group / organization

Team leader's affiliation

Team members (separate each member with semicolons: name1, email1; name2, email2; ...)



Please specify your system name (system name in case of multiple systems for one team)

Block 1

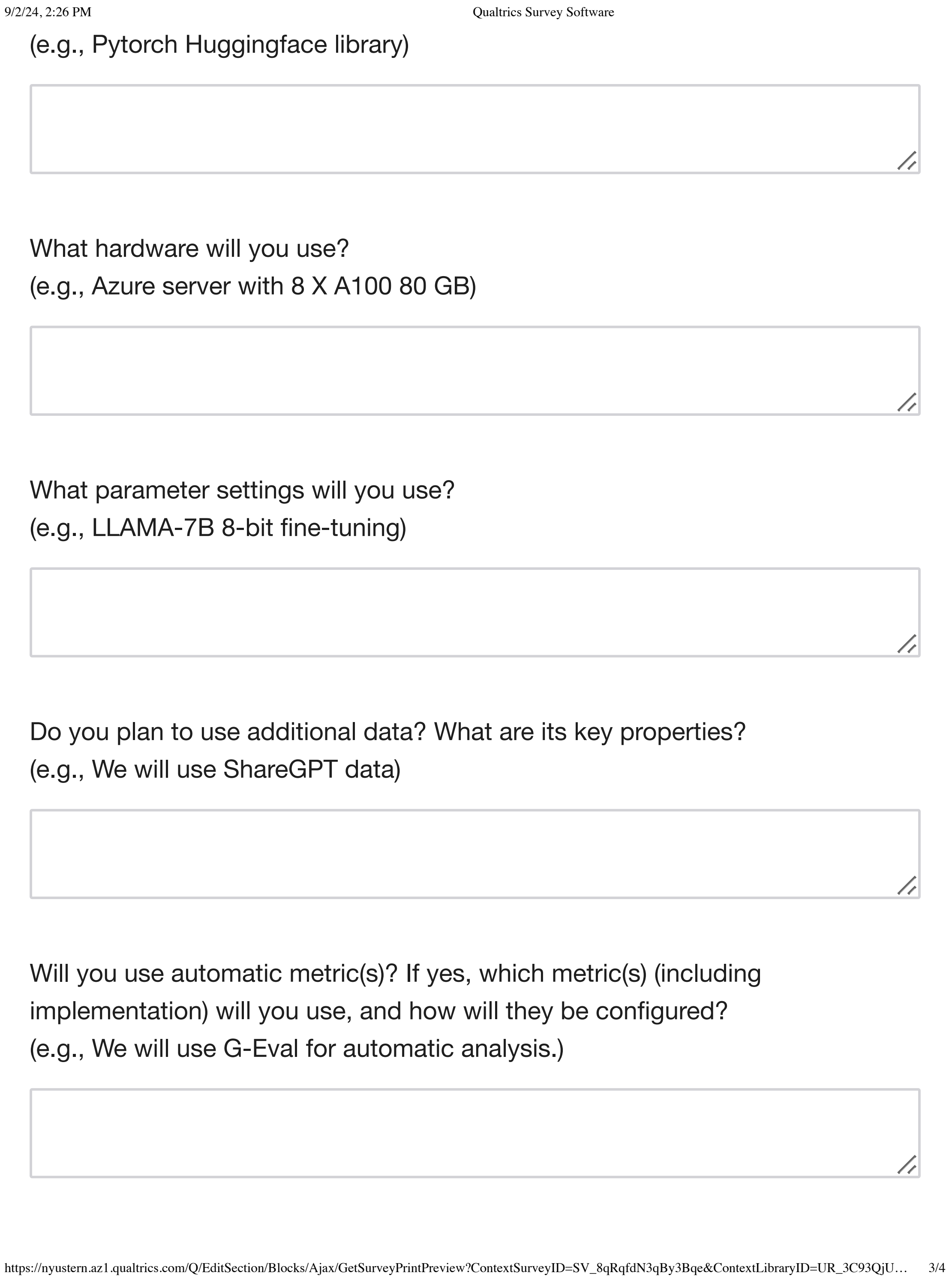
Pre-registration questions

Read the documentation about the shared task here.

What is your intended system(s) that you plan to use for the task(s)?
(e.g., Fine-tuned with parameter efficient fine-tuning using LLAMA-2 7B with a multi-step inference.)

Do you have any specific details that you would like to pre-register?
(e.g., We will pre-train using the XLSum dataset and possibly an internal dataset of 100 tailored examples. We may also use in context learning to prompt engineer solutions. Finally, we may also use GPT-4 to create fine-tuning examples for our model.)

What software libraries will you use?



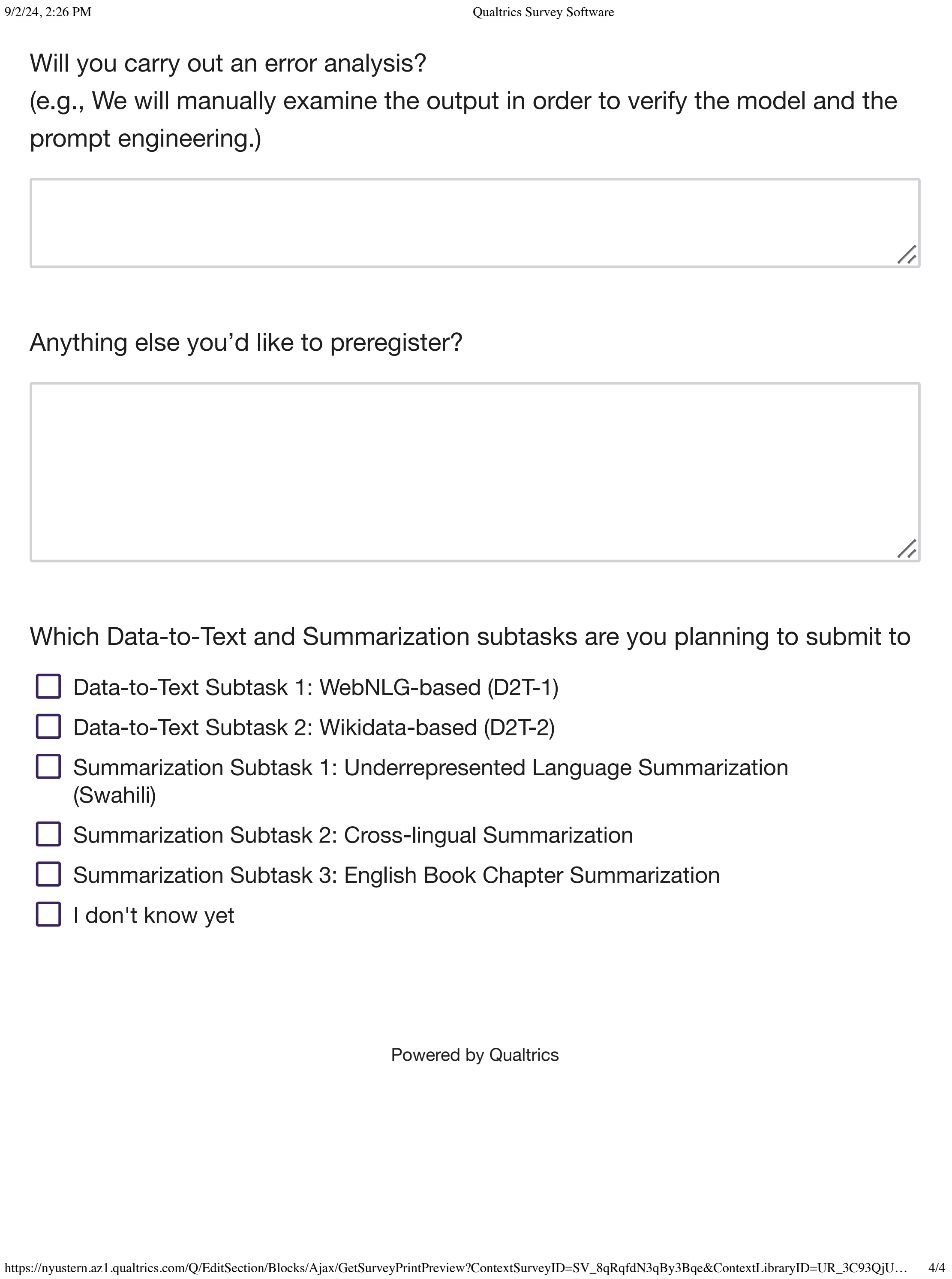
(e.g., Pytorch Huggingface library)

What hardware will you use?
(e.g., Azure server with 8 X A100 80 GB)

What parameter settings will you use?
(e.g., LLAMA-7B 8-bit fine-tuning)

Do you plan to use additional data? What are its key properties?
(e.g., We will use ShareGPT data)

Will you use automatic metric(s)? If yes, which metric(s) (including implementation) will you use, and how will they be configured?
(e.g., We will use G-Eval for automatic analysis.)



9/2/24, 2:26 PM Qualtrics Survey Software

Will you carry out an error analysis?
(e.g., We will manually examine the output in order to verify the model and the prompt engineering.)

Anything else you’d like to preregister?

Which Data-to-Text and Summarization subtasks are you planning to submit to

□ Data-to-Text Subtask 1: WebNLG-based (D2T-1)
□ Data-to-Text Subtask 2: Wikidata-based (D2T-2)
□ Summarization Subtask 1: Underrepresented Language Summarization (Swahili)
□ Summarization Subtask 2: Cross-lingual Summarization
□ Summarization Subtask 3: English Book Chapter Summarization
□ I don't know yet

Powered by Qualtrics

https://nyustern.az1.qualtrics.com/Q/EditSection/Blocks/Ajax/GetSurveyPrintPreview?ContextSurveyID=SV_8qRqfdN3qBy3Bqe&ContextLibraryID=UR_3C93QjU… 4/4