Fine-grained Gender Control in Machine Translation with Large Language Models

Anonymous ACL submission

Abstract

In machine translation, the problem of ambiguously gendered input has been pointed out, where the gender of an entity is not available in the source sentence. To address this ambiguity issue, the task of controlled translation that takes the gender of the ambiguous entity as additional input have been proposed. However, most existing works have only considered a simplified setup of one target gender for input. In this paper, we tackle controlled translation in a more realistic setting of inputs with multiple entities and propose **Gender-of-Entity (GoE) prompting** method for LLMs. Our proposed method instructs the model with fine-grained entity-level gender information to translate with correct gender inflections. By utilizing four evaluation benchmarks, we investigate the controlled translation capability of LLMs in multiple dimensions and find that LLMs reach state-of-the-art performance in controlled translation. Furthermore, we discover an emergence of *gender interference* phenomenon when controlling the gender of multiple entities. Finally, we address the limitations of existing gender accuracy evaluation metrics and propose leveraging LLMs as an evaluator for gender inflection in machine translation.

1 Introduction

In machine translation (MT) research, many efforts have been made to improve the gender accuracy of NMT systems, which have shown to exhibit gender bias (Savoldi et al., 2021; Piazzolla et al., 2023). This research includes the task of handling ambiguously gendered entities in text, which arises from differences in gender markings across different languages (Bentivogli et al., 2020). Without consideration for these ambiguities, existing MT systems default to masculine translations or use stereotypically associated gender, reflecting the bias in training data (Cho et al., 2019).

To address the gender ambiguity issue, multiple approaches have been proposed, such as gendered translation rewriting (Rarrick et al., 2023), generating gender-neutral translations (Piergentili et al., 2023b), and controlled translation (Bentivogli et al., 2020). Specifically, controlled translation methods take the gender of the ambiguous entity as additional input along with the source text, and generate a translation matching the given gender. However, most previous works have only considered gender control of a single entity for each input, and a gap still remains between this simplified experimental setup and texts found in real-world contexts where multiple entities are often mentioned within the same context.

In this work, we investigate the task of controlled translation in a more realistic, *fine-grained* setting where the given text has multiple entities with different gender assignments. To this end, we employ LLMs and propose **Gender-of-Entity (GoE) prompting** for fine-grained gender control in machine translation. Our method utilizes the powerful instruction-following and translation capabilties of LLMs for a more accurate translation aligned with the target gender inflections. LLMs are explicitly instructed to translate the source text with additional entity-level gender information given in natural language statements.

For a comprehensive assessment, we employ four existing benchmarks on gender bias evaluation and investigate the LLM's capabilities in scenarios, ranging from sentences with multiple ambiguously gendered entities to sentences containing both unambiguously gendered and ambiguously gendered entities. From our experiments, we find that the GoE prompting on the LLMs scores up to an average of 95.4% gender accuracy on the Must-SHE dataset, significantly outperforming previous methods based on fine-tuning. Furthermore, we identify a problematic phenomenon of *gender interference* in fine-grained controlled translation, where controlling the gender of one entity adversely affects the gender inflection of other entities. These find-

ings emphasize the necessity of fine-grained assessment of gendered entities in gender bias evaluation.

Finally, we find that conventional metrics used in gender bias evaluation are based on lexical matching, making it challenging to capture synonyms or paraphrases. We thus propose leveraging LLMs as a reference-free evaluator that checks the gender inflections and agreements of the translation. Experimental results show the validity of our proposed evaluator from the high correlation with human judgements and with the automated metrics as well.

To summarize our work, in Section 2, we formulate the fine-grained controlled machine translation task and introduce the four benchmark datasets used in our paper. Next, we report our controlled translation experiments on the four evaluation settings and share our findings in Section 3. In Section 4, we investigate using LLMs as a gender evaluator. In Section 5, we share related works and conclude our paper in Section 6.

2 Gender Control in Machine Translation

We formalize the controlled translation task of gender attributes in machine translation and introduce four gender control scenarios based on existing evaluation benchmarks. We then introduce our proposed LLM-based controlled translation methodology.

2.1 Task definition

In our study, we consider the controlled translation task where one or more entities in the source text are directed to have a specified gender inflection in the target translation output. We approach the task in a *fine-grained* setting, where we control the gender inflection of each entity in the text separately.

We formalize the controlled translation task as follows: given a source sentence src_i and a mapping that assigns a specific gender to each entity in src_i, produce a target translation with the correct gender inflections matching the given mapping. We will refer to the set of controlled entities E_i, set of target genders $\mathcal{G}$, entity-gender mapping $f_i : E_i \rightarrow \mathcal{G}$, and the target translation with matching gender inflections $tgt_i^{f_i}$:

$$controlled_translation(src_i, f_i) \rightarrow tgt_i^{f_i}.$$

In our study, we limit the set of target genders $\mathcal{G}$ to masculine and feminine supported by the evaluation datasets.

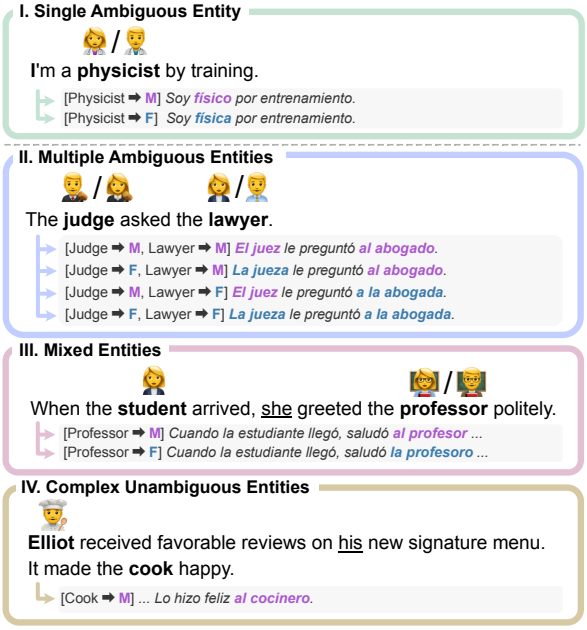

Figure 1: Four gender control scenarios in machine translation investigated in our study.

2.2 Evaluation benchmarks

In order for a comprehensive assessment of fine-grained controlled translation, we employ four existing gender evaluation datasets in our work. The evaluation benchmarks have been constructed with different objectives, enabling multi-faceted analysis of controlled translation, which we categorize into four scenarios, as shown in Figure 1.

Single Ambiguous Entity We first evaluate the controlled translation of sentences with a single ambiguously gendered entity using the MuST-SHE dataset (Bentivogli et al., 2020), constructed from parallel transcripts from TED talks. We specifically use the *Category 1* subset, which consists of sentence pairs that require knowledge of the speaker's gender for the correct translation.

Multiple Ambiguous Entities Next, we evaluate the controlled translation of sentences with multiple ambiguous entities via the recently released GATE dataset (Rarrick et al., 2023). The dataset consists of linguistically diverse sentences with multiple alternative target language translations constructed with the help of bilingual linguists.

Mixed Entities Thirdly, we evaluate controlled translation of sentences where both ambiguously gendered entities and unambiguously gendered entities co-exist via the widely used WinoMT benchmark (Stanovsky et al., 2019). The dataset consists of synthetically constructed sentences containing exactly two entities, of which only one is unambiguously gendered. While most works that utilize

Gender-of-Entity Prompting Template
SYSTEM: You are a professional [TGT_LANG] translator that especially considers translating gender inflections correctly. USER: Translate the following sentence into [TGT_LANG] ([GENDER_ANNOTATION]): [SRC]
Gender Annotation
for [ENT_1], use [GENDER_1]; ...; for [ENT_n], use [GENDER_n]

Table 1: Instruction template for our proposed Gender-of-Entity (GoE) prompting.

the dataset usually consider only the unambiguous entity, we adopt the extension for evaluating the ambiguous entity by Saunders et al. (2020).

Complex Unambiguous Entities Finally, we evaluate controlled translation where the entity is unambiguously gendered but hard to disambiguate due to the complex structure of the source text. For this scenario, we employ the *Contextual* subset in the MT-GenEval dataset (Currey et al., 2022). The samples in this subset consist of two sentences, where the gender of the entity in the second sentence can only be inferred via the first sentence, as illustrated in Figure 1.

For evaluation, we experiment controlled translation in three language directions supported by all four benchmarks: English to Spanish, English to French, and English to Italian. For dataset statistics and preprocessing details, refer to Appendix A.

2.3 Evaluation metrics

We use the term-level coverage and accuracy defined by Bentivogli et al. (2020) for evaluating gender accuracy on all benchmarks, excluding WinoMT, which does not have the target gender annotations required for this metric. **Coverage** is defined by the proportion of (either correct or incorrect) gendered terms that are lexically matched in the generated translation. **Accuracy** is subsequently defined by the proportion of correct terms out of all covered terms in the corpus.

Alternatively, the gender accuracy metric defined by Stanovsky et al. (2019) is used for the WinoMT dataset. The metric is based on a source-target alignment-based algorithm used jointly with a language-specific gender morphology analyzer to check if the gendered terms are correctly inflected.

For evaluating translation quality, we utilize the BLEU score, an n-gram based lexical metric, and COMET score (Rei et al., 2022), a neural metric that has been shown to be closely aligned with human judgments. [1]

2.4 Gender-of-Entity prompting for LLMs

We propose **Gender-of-Entity (GoE) prompting** in our work for fine-grained controlled translation of gender using LLMs. Our zero-shot approach builds upon LLM's translation and instruction-following capabilities to direct the LLM to translate with the specified gender for each entity.

The template for Gender-of-Entity prompting is shown in Table 1, where `[TGT_LANG]` is the slot for the name of the target translation language, `[SRC]` is the slot for the source text, and `[GENDER_ANNOTATION]` is the slot where we specify the entity-level gender mappings in natural language. By default, we use an entity-level gender annotation scheme where we list the entities and their target gender, delimited by ";". More specifically, `[ENT_i]` is substituted with the entity name found in source text, and `[GENDER_i]` is substituted by either "*he/him*" or "*she/her*" for male and female gender inflections respectively.

We use two instruction-tuned LLMs, Llama 2 70B Chat and ChatGPT 3.5 (`gpt-3.5-turbo`) for applying GoE prompting to LLMs. The two models have shown to have competitive translation performance for the three language directions evaluated in our study (Zhu et al., 2023).

3 Main Experiments

We experiment on controlled translation of gender with our proposed method on the four evaluation benchmarks and compare them with existing approaches.

3.1 Gender Control of Single Ambiguous Entity

First, we consider the most straightforward setup where there is a single ambiguously gendered entity in the source sentence. We evaluate on the MuST-SHE benchmark (Bentivogli et al., 2020), where we control the ambiguous entity to the designated gender label provided by the annotation. Since the ambiguously gendered entity is always the speaker for this dataset, we use the gender annotation "*the speaker is male*" or "*the speaker is female*" depending on the designated gender.

[1] The sacrebleu id for computing bleu is: `s:1000|rs:12345|c:mixed|e:no|tok:13a|s:exp|v:2.3.1` and for COMET, we use the `Unbabel/wmt22-comet-da`.

Method	ES				FR				IT			
	Cov.	Acc.	BLEU	COMET	Cov.	Acc.	BLEU	COMET	Cov.	Acc.	BLEU	COMET
NLLB-200 600M D.												
Baseline	74.6	53.9	43.7	85.1	62.7	53.6	37.0	82.8	60.0	52.3	35.4	84.8
Gender prefixing	75.3	77.6	44.9	85.6	62.2	72.2	38.3	83.3	60.3	74.2	36.2	84.9
CG* (Liu and Niehues, 2023)	-	82.8	44.7	84.7	-	79.4	38.7	82.5	-	83.6	35.4	83.7
FT* (Liu and Niehues, 2023)	-	86.9	43.7	84.0	-	85.0	38.2	82.0	-	87.8	34.4	83.5
NLLB-200 1.3B D.												
Baseline	76.1	60.0	45.5	85.8	63.3	58.7	39.4	83.8	64.1	59.6	37.5	86.0
Gender prefixing	76.8	84.3	**47.3**	86.0	61.8	81.2	**40.5**	83.5	63.9	84.7	**38.1**	85.9
Llama 2 70B Chat												
Baseline	69.0	54.8	34.4	82.4	53.4	54.3	29.5	80.2	54.1	54.1	28.5	81.9
GoE prompting	71.0	94.9	37.6	83.5	57.6	94.0	31.9	81.6	54.5	89.8	30.0	82.4
ChatGPT 3.5												
Baseline	73.4	54.0	39.1	85.5	42.0	56.7	33.1	83.5	63.1	51.9	33.5	86.1
GoE prompting	77.1	**96.5**	42.7	**87.0**	64.9	**95.3**	37.9	**85.3**	63.9	**94.4**	35.4	**86.7**

Table 2: Results of controlled translation on the Must-SHE dataset. Gray text denote baseline results without the gender specified. *Results are taken from Liu and Niehues (2023)

3.1.1 Baseline methods

We compare our approach with three baseline methods developed for pre-trained NMT models: gender prefixing, gender-specific fine-tuning (FT), and inference-time classifier guidance (CG) (Liu and Niehues, 2023). Gender prefixing simply adds gendered prefixes "`MALE:`" and "`FEMALE:`" in front of the source text. Gender-specific fine-tuning (FT) fine-tunes separate NMT models on a gendered parallel corpus for each gender. Finally, inference-time classifier guidance (CG) utilizes a pre-trained gender attribute classifier module to modify the decoder activations of existing NMT models during inference. For gender prefixing, we share evaluation results on both NLLB-200 600M distilled and NLLB-200 1.3B distilled models, which are multilingual NMT models shown to have strong translation performance (Team et al., 2022). For the fine-tuning and classifier-guidance approaches, we report results by Liu and Niehues (2023) on the NLLB-200 600M distilled model.

3.1.2 Results

Experimental results, shown in Table 2, indicate that GoE prompting is highly effective at controlling the gender of a single entity, reaching very high gender accuracies on both Llama 2 and ChatGPT 3.5 models and for all three target languages. Especially for ChatGPT, the accuracies are in the range of 94% and 96%, reaching state-of-the-art performance. Furthermore, even though the baseline gender accuracy of NLLB-200 600M distilled model and LLMs have similar scores, the improvement from our zero-shot prompting exceeds the improvement from existing baseline approaches that require fine-tuning. This highlights the strong zero-shot instruction following capabilities of LLMs.

In terms of translation quality, NLLB-200 models have the highest BLEU scores, followed by ChatGPT and Llama 2 models. Based on the COMET scores, however, ChatGPT scores the highest, followed by NLLB-200 models and Llama 2. These findings suggest that ChatGPT 3.5 LLMs have competitive zero-shot translation performance compared to the evaluated NLLB-200 models, while Llama 2 trails behind slightly.

3.2 Gender Control of Multiple Ambiguous Entities

Next, we evaluate controlled translation on the GATE benchmark (Rarrick et al., 2023), which consists of sentences with up to three ambiguously gendered entities. The dataset also includes translations and annotations of all possible combinations of male/female gender mappings for each entity. This means a sentence with N ambiguous entities will have 2^N possible gender mappings and an equal number of translations. We evaluate Llama 2 and ChatGPT on controlled translation to all possible gender mappings using the default GoE prompting template described in Table 1.

3.2.1 Results

Gender accuracy results of the various subsets of the GATE test set are reported in Table 3. First, we find that the baseline translations of LLMs without gender control default to masculine inflections by a ratio of approximately 9:1. With GoE prompting, we observe over 95% accuracy on male entities for both Llama 2 and ChatGPT. However, the accuracy

Lang.	Model	Method	Cov.	Gender		#Ent	
				Acc_M	Acc_F	Acc_1	$Acc_{\geq 2}$
ES	Llama 2	Baseline	57.4	88.7	11.3	50.0	50.0
		GoE	62.2	97.9	68.1	84.9	81.3
	ChatGPT	Baseline	66.5	88.9	11.1	50.0	50.0
		GoE	67.0	**98.8**	**92.3**	**96.6**	**94.6**
FR	Llama 2	Baseline	65.3	95.4	4.6	50.0	50.0
		GoE	66.2	**97.5**	58.8	82.7	74.1
	ChatGPT	Baseline	71.7	88.9	11.1	50.0	50.0
		GoE	69.9	96.4	**81.0**	**91.3**	**86.4**
IT	Llama 2	Baseline	62.1	94.4	5.6	50.0	50.0
		GoE	61.2	**98.7**	49.7	71.3	75.6
	ChatGPT	Baseline	72.6	94.8	5.2	50.0	50.0
		GoE	71.8	98.2	**77.9**	**89.8**	**87.3**

Table 3: Gender accuracy of Llama 2 70B Chat model and ChatGPT 3.5 model on the GATE test set. Gray text denote baseline results without the gender specified.

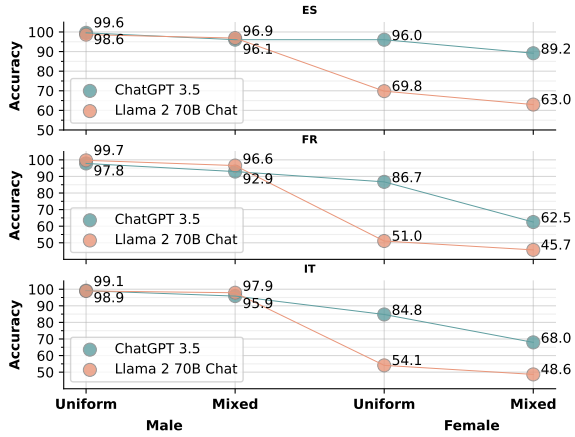

Figure 2: Gender accuracy of GoE prompting on the GATE subset with two ambiguous entities (#Ent=2). *Uniform* denotes translation with both entities mapped to the same gender, and *Mixed* denotes translation with entities mapped to different genders.

of female entities is lower in comparison, indicating room for improvement of controlled translation with LLMs.

Also, we generally find that the gender accuracy of sentences containing multiple ambiguous entities ($Acc_{\geq 2}$) is slightly lower than those containing a single ambiguous entity (Acc_1). This trend potentially suggests that LLMs find controlling gender inflections of multiple entities within a sentence more challenging. In order to investigate this further, we take the GATE subset with exactly two ambiguous entities and compare the gender accuracy of samples where the two genders are assigned the same gender with the samples where they are assigned differently.

We report our fine-grained analysis of sentences with two ambiguous entities in Figure 2. Results demonstrate that LLMs find it easier to translate sentences with the same gender inflection for all entities (*Uniform*) than those with mixed gender

Method	ES		FR		IT	
	Acc_U	Acc_A	Acc_U	Acc_A	Acc_U	Acc_A
NLLB-200						
Baseline	72.0	34.0	66.7	36.6	54.1	34.6
GACL	**91.2**	3.0	**85.0**	7.8	**72.3**	7.3
Llama 2						
Baseline	56.7	43.6	54.6	44.2	46.5	41.8
$GoE_{amb.}$	44.1	92.4	45.8	**89.9**	37.7	**85.7**
GoE_{full}	74.3	83.7	67.0	69.4	65.0	75.7
ChatGPT						
Baseline	62.4	41.6	58.0	41.8	49.6	38.5
$GoE_{amb.}$	39.2	**93.9**	42.7	84.5	37.0	76.8
GoE_{full}	84.3	91.1	76.9	82.7	63.2	74.5

Table 4: Results on the WinoMT dataset. Acc_U denotes the gender accuracy of the unambiguously gendered entity and Acc_A denotes the gender accuracy of the ambiguously gendered entity.

inflections (*Mixed*). This **interference behavior** in mixed settings adversely affects female gender mappings more strongly, with an absolute accuracy difference of up to 24.2% between uniform and mixed settings.

3.3 Gender Control of Mixed Entities

In this subsection, we evaluate on the controlled translation of sentences containing a mix of ambiguously gendered and unambiguously gendered entities via the WinoMT benchmark (Stanovsky et al., 2019). We evaluate controlled translation with the gender of the ambiguous entity specified to have a *different gender* from the existing unambiguous entity, as we observed in previous subsection 3.2 that having same genders for the multiple entities can artificially boost the performance.

3.3.1 Baseline methods

We compare our approach with a recent gender debiasing approach based on gender-aware contrastive learning (GACL) (Lee et al., 2023). While their approach was originally proposed for improving the gender accuracy of unambiguously gendered entities, we also evaluate its effect on the ambiguously gendered entitis. We evaluate the model based on NLLB-200 1.3B distilled model.

3.3.2 Results

From the results shown in Table 4, we first notice that almost all baseline models score over 50% on the gender accuracy of unambiguous entity (Acc_U) while scoring lower than 50% on the ambiguous entity (Acc_A). This indicates that instead of defaulting to masculine translation for ambiguous gender, models inflect it to the same gender as the

Method	ES					FR					IT				
	Cov.	Acc_M	Acc_F	BLEU	COMET	Cov.	Acc_M	Acc_F	BLEU	COMET	Cov.	Acc_M	Acc_F	BLEU	COMET
NLLB-200 1.3B D.															
Baseline	71.7	**99.3**	70.0	44.3	86.5	58.7	97.8	54.0	36.2	84.0	56.8	98.6	52.0	29.0	84.4
GACL (Lee et al., 2023)	72.1	99.1	**97.0**	39.8	85.6	58.8	**98.4**	87.5	33.8	83.6	56.7	**99.3**	**91.4**	22.7	82.5
Llama 2 70B Chat															
Baseline	66.5	98.8	67.9	43.7	86.7	56.0	97.0	55.7	35.3	84.4	55.4	98.6	49.5	29.0	86.1
GoE	67.8	99.0	78.6	44.1	87.1	58.2	97.3	70.0	37.1	85.3	54.9	**99.3**	62.0	29.6	86.5
ChatGPT 3.5															
Baseline	71.6	97.1	86.2	48.1	**89.3**	61.7	97.6	81.4	41.8	**87.8**	59.0	98.7	68.7	33.5	88.7
GoE	71.6	98.7	94.4	48.8	88.8	61.2	97.1	**89.7**	41.6	87.6	58.9	98.9	81.2	33.6	88.9
I-GoE (few-shot)	71.8	98.6	90.4	**50.5**	**89.3**	61.9	98.0	88.0	**43.8**	**87.8**	59.4	98.5	76.3	**35.0**	**89.0**

Table 5: Results of controlled translation on the MT-GenEval Contextual test set.

unambiguous entity. This gender interference is amplified by the GACL method, where Acc_A simultaneously drops by over 20% as Acc_U improves by 20% compared to baseline.

Next, we find that explicitly controlling the gender of the ambiguous entity with GoE ($GoE_{amb.}$) significantly improves the accuracy of Acc_A for both Llama 2 and ChatGPT 3.5 models. However, this time, we observe that Acc_U is lower by at least 10% compared to the baseline. These findings denote that both fine-tuning and GoE prompting methods interfere with the gender of other entities in the sentence. Also, on manual inspection, we find that a few of the WinoMT evaluation samples are inherently ambiguously phrased so that either of the entities could be referred by the gendered pronoun.

Lastly, we experiment with controlling the gender of both entities with gold annotations using GoE prompting (GoE_{full}). Results show that explicitly specifying both entities leads to the best balanced accuracy improvement of both entities for both LLMs. These results suggest the usefulness of controlled translation for facilitating the correct translation of unambiguous entities, even if it can be inferred via coreference resolution.

3.4 Gender Control of Complex Unambiguous Entities

In our fourth evaluation task, we evaluate controlling the gender of complex unambiguous entities using the *Contextual* subset of the MT-GenEval dataset (Currey et al., 2022). We experiment controlled translation by specifying the gender of the unambiguous entity from the second sentence in our prompt. However, the dataset does not provide annotations on the unambiguous entity nor its gender. We thus obtain pseudo-gold entity annotation by using the Spacy[2] dependency parser to extract the noun phrase of the second sentence while using the gendered word list (Zhao et al., 2017) to extract the gender of the entity in the first sentence.

3.4.1 Results

In the results shown in Table 5, we first note that baseline models show a relatively high gender accuracy compared to other evaluation datasets, as the evaluated entities are unambiguous and their gender can be inferred from the first sentence. Next, we find that explicitly specifying the pseudo-gold gender via GoE prompting improves the gender accuracy further, especially for the female gender with an improvement of 12.4% and 9.5% across evaluated language directions for Llama 2 and ChatGPT respectively. Translation quality remains within similar range before and after prompting, suggesting gender prompting does not harm the translation quality.

3.4.2 Additional results on end-to-end translation

Unlike ambiguously gendered entities, the gender of unambiguous entities can be inferred from the given text. Thus, we additionally experiment whether LLMs could be instructed to infer the entity and its gender from the given sentence and subsequently translate the sentence, in an end-to-end setup. This idea is adopted from recent findings that generating intermediate reasoning steps improve performance of LLMs on complex reasoning tasks, since identifying the gender of entities could be seen as an intermediate reasoning step to generating translation with correct gender inflection.

To instruct LLMs to **I**nfer the entity's gender and subsequently translate, we additionally add few-shot examples to the GoE prompt, which we refer to as **I-GoE prompting**. The few-shot examples are sampled from the MT-GenEval dev set, and the output translations start with the following pretext: "*From the given source text, we can infer that* [ENT]

[2] https://spacy.io

uses [GENDER]. *Therefore, the* [LANG] *translation with correct gender inflection is:*".

Results of I-GoE prompting on ChatGPT shown in Table 5 show a meaningful improvement from the baseline, with an average of 5.8% absolute improvement in female gender accuracy. However, the original GoE prompting based on pseudo-gold annotations still hold the highest gender accuracy overall, suggesting rooms for improvement in I-GoE prompting.

4 LLM as Gender Evaluators

In Section 3, our methodology exhibits significant performance based on automated gender accuracy metrics. However, the employed coverage-based metrics are dependent on the annotated gender terms. Such dependence poses a challenge in assessing gender terms that do not match the annotations due to the use of synonyms or different grammatical structures. For example, the English term "professor" can be translated into either "profesor/profesora" or "maestro/maestra" in Spanish. As shown in Tables 2, 3, and 5, at least 20% of the samples for each evaluation benchmark remain unevaluated due to missing coverage from the provided gender annotations. As a result, it is necessary to address these issues for a more complete and accurate assessment of gender accuracy.

To address such complexity, we propose LLMs as Gender Evaluators (LGE). We provide multilingual LLMs with instructions as specified in Table 11 in Appendix B. As input, the LLM is given the source sentence, the model prediction, the controlled entity and its designated gender. It is then prompted to evaluate whether the given entity is inflected to the designated gender as a binary judgement of either ACCURATE or INACCURATE. Our evaluation method does not require the reference translation nor any gendered term annotations, allowing evaluation of samples previously not covered by automated metrics.

We explore the viability of LGE by first performing a sanity check with evaluation of gold human-provided translations. Then, we collect human expert annotations to assess the correlation of LGE with human judgements. Finally, we re-assess controlled translation with LGE, including samples previously omitted by the coverage-based metrics.

Dataset	Lang.	F1-score	Precision	Recall
Must-SHE	ES	95.6	94.8	96.3
	FR	93.2	89.7	97.0
	IT	96.0	94.8	97.3
GATE	ES	95.5	93.3	97.9
	FR	86.5	81.4	92.3
	IT	92.0	87.1	97.5

Table 6: Sanity check results of LGE gender accuracy evaluation on Must-SHE and GATE test sets.

	Agreement (%)	Cohen's κ
LGE ⇔ Human	93.0	0.691
LGE ⇔ Ref-based*	87.0	0.688

Table 7: Agreement and Cohen's Kappa Coefficient between the LGE, human annotators, and the reference-based metric. *The metrics are computed with the subset measurable by the reference-based metric.

4.1 Sanity Check with Reference Translations

Initially, we conduct a sanity check to determine whether LLMs possess the capability to function as gender accuracy evaluators. The Must-SHE and the GATE dataset provide a valuable resource for this purpose, as they contain the possible variants of translations based on the gender of entities in the source sentences. Therefore, we conduct an experiment using these reference translations. In scenarios where the provided reference aligns with the specified gender condition, the LLMs should evaluate it as ACCURATE. Such correct references are considered positive samples. On the other hand, in cases where the provided reference is incorrect, the response should be INACCURATE, and these incorrect references are categorized as negative samples. We calculate the F1 score, precision, and recall based on this categorization. The results of these experiments are presented in Table 6. When correct references are provided, the LLMs predominantly evaluate them as accurate. Conversely, when incorrect references are given, the models mostly evaluate them inaccurate. This results show the effectiveness of LLMs as gender accuracy evaluators. Experimental details are in Appendix B.

4.2 Gender Accuracy Evaluation with LGE

Subsequently, we assess the validity of our LGE utilizing outputs of ChatGPT and NLLB-200 given MT-GenEval dataset. We sample 100 outputs covered by reference gender terms and another 100 not covered. These are then compared with evaluations from human annotators and those based on gender terms. For outputs not covered by gender terms, we

		Translator Model	
		ChatGPT + GoE	Llama 2 + GoE
Cov.		67.0	62.2
Ref-based	Acc_C	96.6	84.9
	$Acc_{N.C}$	N/A	N/A
LGE	Acc_C	94.7	82.6
	$Acc_{N.C}$	79.9	64.8
	Acc_{All}	90.6	76.6

Table 8: Re-evaluation results of our gender-controlled translation with LGE on the GATE dataset. Acc_C represents gender accuracy on sentences covered by reference gender terms, and $Acc_{N.C}$ represents gender accuracy on sentences not covered.

rely exclusively on human annotator evaluations. Details of the human annotation are in Appendix B.1. In Table 7, we observe a substantial agreement with human evaluations. This indicates the feasibility of using LGE to effectively evaluate outputs, regardless of whether they are covered by gender terms or not. Also, in cases covered by reference gender terms, there is a high concordance between term-based accuracy metrics and LGE.

After ensuring the reliability for LGE, we further examine the sentence-level gender accuracy of gender controlled translations generated through our GoE prompting. In Table 11, in situations where the translation output is covered by the reference gender terms, it shows a level of accuracy similar to that of traditional metrics. However, for sentences not covered by the reference gender terms, a tendency towards lower gender accuracy is observed. Our evaluation, being reference-free, allows us to uncover such situations. In cases of non-coverage, there is a higher likelihood that gender translation has not been accurately rendered. Therefore, metrics calculated only in cases of coverage should be interpreted as relative comparisons and not absolute values, as they might slightly overestimate the actual performance.

5 Related Works

5.1 Ambiguous Gender in Machine Translation

The problem of handling ambiguous gender in machine translation has been pointed out by multiple studies, providing benchmarks for evaluation (Cho et al., 2019; Bentivogli et al., 2020; Rarrick et al., 2023).

Multiple approaches have been proposed to handle ambiguous gender bias in machine translation, including rewriting a translation to another gender (Rarrick et al., 2023), generating gender-neutral translations (Piergentili et al., 2023a,b), and controlled translation (Bentivogli et al., 2020). However, they do not consider fine-grained gender control of multiple entities.

A recent work also proposed gender-specific machine translation with LLMs (S'anchez et al., 2023). However, they also only consider two gendered variations for each sentence, and use LLMs to translate both variations without control.

5.2 Machine Translation with LLMs

As LLMs are widely adopted to various fields, recent studies have explored usage of LLMs for machine translation (Herold et al., 2023; Garcia et al., 2023; Vilar et al., 2023). The adoption of LLMs in MT has been shown to contribute to addressing diverse gender biases including pronoun genders and name entities (Saunders and Olsen, 2023; Wang et al., 2022; Petrick et al., 2023; Zhang et al., 2023; Attanasio et al., 2023).

5.3 LLM-based evaluation

Traditionally, semantic-based metrics employ neural networks through encoder models such as BERTScore (Zhang et al., 2020). Recently LLM-Eval (Lin and Chen, 2023) utilized decoder-based models as metrics, and demonstrates a higher correlation with human evaluation. In MT tasks, Kocmi and Federmann (2023) shows GPT evaluation is better than BLEU.

6 Conclusion

In this paper, we tackled fine-grained gender control in machine translation. To solve this task, we proposed Gender-of-Entity prompting method for LLMs, where we instruct LLMs to translate with additional entity-level gender information. Results on four evaluation benchmark show promising capabilities of LLMs as controlled translator of gender, with up to 95% average accuracy on the MuST-SHE dataset. We also observe a new phenomena of performance degradation when translating sentences with multiple gendered sentences with different target genders. Finally, we addressed the limitations of existing automated gender evaluation metrics by proposing LLMs as Gender Evaluators (LGE). Based on experimental results, LGE evaluations was shown to have high correlation with human judgements.

7 Limitations

Our study has been focused on Spanish, French, and Italian, which fall within the Romance language family. As a result, languages outside the Romance family were not covered in our analysis. Nonetheless, given the gradual enhancement in the multilingual capabilities of LLMs, we anticipate that our methodology could be sufficiently applicable to other languages.

Additionally, the utilization of GoE prompting and its evaluation requires the gender-annotated dataset. Particularly, if the annotations contain errors, there is a possibility that it could actually lead to a degradation in performance. To address such problem in our research, we make evaluation methods extending the setting of existing studies. However, given the inherent complexity and intricate nature of languages, there may still be instances where our approach fails to adequately address scenarios involving sentences that lack explicit entities or where both ambiguous and unambiguous entities are intricately intertwined.

Finally, even though our methodology demonstrates great performance compared to baselines, there is much room for improvement. Some possible future directions include improving translation with few-shot examples, constructing a more sophisticated instrution prompt, and incorporating reinforcement learning.

8 Ethical Considerations

Since our research is concentrated on gender bias related to ambiguous entities, the applicability of our study to other demographic biases beyond gender remains under-explored. Therefore, any extension of our methodology to encompass demographic biases would require thorough consideration and additional research.

Furthermore, since annotations in existing datasets are framed within a binary setting, we have limited results only on the binary gender, difficult to evaluate performance on gender-netural or non-binary genders in our studies. However, as Multilingual Large Language Models (LLMs) have shown to well-adapt to tasks with instructions, we believe that, given the availability of relevant datasets, our methodology could also be applicable to non-binary genders.

References

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the must-she corpus. *ArXiv*, abs/2006.05754.

Won Ik Cho, Jiwon Kim, Seokhwan Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. *ArXiv*, abs/1905.11684.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, pages 10867–10878. PMLR.

Christian Herold, Yingbo Gao, Mohammad Zeineldeen, and Hermann Ney. 2023. Improving language model integration for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7114–7123, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Minwoo Lee, Hyukhun Koh, Kang il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2023. How transferable are attribute controllers on pretrained multilingual translation models? *ArXiv*, abs/2309.08565.

Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2023. Document-level language models for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 375–391, Singapore. Association for Computational Linguistics.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *ArXiv*, abs/2306.05882.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. From inclusive language to gender-neutral machine translation. *ArXiv*, abs/2301.10075.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus. *ArXiv*, abs/2310.05294.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eduardo S'anchez, Pierre Yves Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta Ruiz Costa-jussà. 2023. Gender-specific machine translation with large language models. *ArXiv*, abs/2309.03175.

Danielle Saunders and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. *ArXiv*, abs/2306.04573.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Experimental Details for Controlled Translation

A.1 Dataset Statistics

We report the dataset statistics of the four evaluation benchmarks used in this paper in Table 9. For the GATE dataset, we evaluate all entity-level gender mapping combinations for each sample. Hence, the number of evaluated translations is equal to the number of dataset samples multiplied by the number of possible mappings, which is 2^N, where N is the number of ambiguous entities.

For the GATE dataset, we exclude samples with incorrect annotations, where the number of entities does not match the annotations. For the MT-GenEval dataset, we exclude samples with incorrect annotations, where the first sentence is either blank or does not contain gendered terms based on the word list. We manually went over the excluded samples to verify that the annotation was incorrect. For computing the term-based accuracy, we use the gold annotated gender terms and entity terms for MuST-SHE and GATE datasets. For the MT-GenEval dataset, we obtain gendered terms by comparing and extracting the differing terms between the male and female gold reference translations provided by the dataset. For the diff tool, we use the `difflib.SequenceMatcher` algorithm in the Python 3 -build-in library.

A.2 LLM output post-processing

In our translation experiments with LLMs, we found LLMs often generate additional comments either before or after the translations. Thus, we apply a basic rule-based post-processing to extract the translated sentence from the LLM generation output. First, we split the output into sentences based on the newline character `\n`, and filter out sentences that contain any of the following tokens: "gender", "translat", "sentence", and "note". Out of the remaining sentences, we take the first sentence as the translation output.

B Experimental details for LLMs as Gender Evaluators

We utilize a state-of-the-art large language model, gpt-4-turbo (`gpt-4-1106-preview`) for the role of gender accuracy evaluator. Our initial sanity check experiments revealed that gpt-3.5-turbo model showed less satisfactory performance for this role, as shown in Table 10. The example prompts are in Table 11.

Dataset	Subset	ES	FR	IT
MuST-SHE	1M	287	292	282
	1F	284	315	278
GATE test	#Ent=1	751	775	564
	#Ent=2	150	222	259
	#Ent=3	12	0	20
WinoMT		3,888	3,888	3,888
MT-GenEval test	Contextual	1,096	1,099	1,094

Table 9: Dataset statistics of the four evaluated benchmarks.

		gpt-4-turbo			gpt-3.5-turbo		
Dataset	Lang.	F1-score	Precision	Recall	F1-score	Precision	Recall
Must-SHE	ES	**95.6**	94.8	96.3	71.8	56.4	98.8
GATE	ES	**95.5**	93.3	97.9	57.4	40.4	98.8

Table 10: Sanity Check Results for ChatGPT models.

B.1 Human Annotation Process

As described in Section 4.2, we employ three Spanish-English bilingual annotators to evaluate the gender accuracy of ChatGPT and NLLB outputs based on the MT-GenEval dataset. Native Spanish speakers from the author's local communities proficient in both Spanish and English are recruited as annotators. They were informed with the research objective of this annotation and obtained consent on the use of the dataset. We instruct the annotators to conduct a more detailed evaluation by assessing additional aspects beyond just gender accuracy. An example of the detailed annotation page can be found in Figure 3. The final assessment of gender accuracy is achieved by adopting the majority decision among the three expert annotators. In the gender accuracy evaluation conducted by three annotators, there are 182 instances where all three agree, and 18 instances where only two agree. The agreement among these three, calculated using the Fleiss kappa coefficient, is 0.709, indicating that the opinions of the three annotators substantially agree. Additionally, the comparison between the final human label, determined by the majority voting of these three annotators, and the original reference-based metric, is presented in Table 12. Most of the discrepancies stem from errors by the annotators or label errors within the dataset itself.

B.2 Example Outputs of LGE

Example outputs of LGE can be found in Table 13.

Evaluator Prompting Template
SYSTEM: You are evaluating a gender-conditioned translation. Please specifically focus on whether the translation accurately reflects the gender representation of the provided entities. Check if the words related to the entities are translated in a way that is consistent with the entities' specified genders. After reviewing the input, provide your evaluation in the following format: Comment: [Your explanation regarding the gender representation in relation to the entities in the translation.] Gender Accuracy: [ACCURATE or INACCURATE]. USER: Source [EN]: `[SRC]` Condition: Entity "`[ENT]`" should be translated as "`[GEN]`" Translation `[LANG]`: `[PRED]`

Table 11: Instruction template for evaluation using LLMs.

	Agreement(%)	κ
Human ⇔ Ref-based*	84.0	0.607

Table 12: Agreement and Cohen's Kappa Coefficient between the human annotators and the reference-based metric. Comparisons are calculated only for 100 outputs that are covered by reference gender terms.

`User Prompt:`
Source [EN]: At that time, the Russian ambassador was not there.
Condition: Entity "ambassador" should be translated as feminine.
Translation [ES]: En ese momento, la embajadora rusa no estaba allí.
 	
`Output:`
Comment: The translation accurately reflects the gender representation of the entity "ambassador" by using the feminine form "la embajadora" in Spanish, which aligns with the condition provided.
Gender Accuracy: ACCURATE

`User Prompt:`
Source [EN]: This was an uneasy alliance at best, and the allies would periodically break off the assault to fight each other.
Condition: Entity "allies" should be translated as masculine. Entity "each other" should be translated as masculine.
Translation [ES]: Aquella fue una alianza frágil, en el mejor de los casos, pues los aliados interrumpieron varias veces el ataque para luchar entre ellas.
 	
`Output:`
Comment: The translation of "allies" as "aliados" is correct and masculine, which is consistent with the specified gender. However, the phrase 'para luchar entre ellas' is incorrect because "ellas" is feminine and should be "ellos" to reflect the masculine gender specified for "each other".
Gender Accuracy: INACCURATE.

`User Prompt:`
Source [EN]: In 1668, plays by women began to be staged in London. The performance ran for six nights, which was regarded as a good run for an unknown author.
Condition: Entity "author" should be translated as feminine.
Translation [ES]: En 1668, las obras de teatro de mujeres comenzaron a ser puestas en escena en Londres. La actuación duró seis noches, lo que se consideró una buena carrera para un autor desconocido.
 	
`Output:`
Comment: The translation does not accurately reflect the gender of the entity "author" as feminine. The Spanish word for author, "autor" should have been translated as "autora" to indicate a female author.
Gender Accuracy: INACCURATE

Table 13: Example Outputs of LGE

English-Spanish Translation Proofreading

☑ Show Instructions

You are evaluating an English to Spanish translation. Please specifically focus on whether the translation accurately reflects the gender representation of the provided entities. Check if the words related to the entities are translated in a way that is consistent with the entities' specified genders. Comments are not mandatory.

English Text:

The topic of her prize-winning essay was same-sex marriage. Darville is reported to have worked variously as a graphic designer, property law lecturer, and physical education teacher.

Spanish Translation:

El tema de su ensayo premiado fue el matrimonio entre personas del mismo sexo. Darville trabajó como diseñadora gráfica, profesora de derecho de propiedad y profesora de educación física.

Rate the overall translation quality from 1 to 5 (Very Bad 1, Bad 2, Fair 3, Good 4, Very Good 5)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Is the entity "['**Darville**']" correctly inflected to have **female** gender in the Spanish translation?

○ Accurate ○ Inaccurate

If there are other entities in the sentence with specified gender, do they have correct gender inflections?

○ No_other_gender_specified_entity ○ Accurate ○ Inaccurate

What are the other entities in the sentence with specified gender?

Any additional comments on the translation? (Optional)

Item 1 of 200

Submit

📝 **Note**: Click 'Submit' to save your evaluation. 'Next' will not save the current state.

Figure 3: Example of human annotation pages