Analysing Reference Production of Large Language Models

Chengzhao Wu[♣], **Guanyi Chen**[♣*], **Fahime Same**[♡], **Tingting He**[♣]
[♣]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University
[♡]Trivago N.V.
wcz@mails.ccnu.edu.cn, {g.chen, tthe}@ccnu.edu.cn, fahime.same@trivago.com

Abstract

This study investigates how large language models (LLMs) produce referring expressions (REs) and to what extent their behaviour aligns with human patterns. We evaluate LLM performance in two settings: slot filling, where REs are generated within a fixed context, and language generation, where REs are analysed within fully generated texts. Using the WebNLG corpus, we assess how well LLMs capture human variation in reference production and analyse their behaviour by examining the influence of several factors known to affect human reference production, including referential form, syntactic position, recency, and discourse status. Our findings show that (1) task framing significantly affects LLMs' reference production; (2) while LLMs are sensitive to some of these factors, their referential behaviour consistently diverges from human use; and (3) larger model size does not necessarily yield more human-like variation. These results underscore key limitations in current LLMs' ability to replicate human referential choices.

1 Introduction

Referring Expression Generation (REG) is a fundamental sub-task of Natural Language Generation (NLG, Reiter and Dale, 2000; Krahmer and van Deemter, 2012; Gatt and Krahmer, 2018). At a given point in a discourse, REG generates a Referring Expression (RE) for an intended referent based on the surrounding context (Belz and Varges, 2007). REG has significant practical value in commercial NLG applications (Reiter, 2017) and remains an active area of research in theoretical linguistics and psycholinguistics (van Deemter, 2016).

Amid the recent surge of interest in Large Language Models (LLMs), several corpus-based studies have evaluated LLMs on the task of REG (Gautam et al., 2024; Ellison and Same, 2024). These studies have shown that LLMs differ significantly from humans, particularly in their inability to capture the variability of human reference use.

In this study, we propose a different corpus-based evaluation grounded in the observation that the standard REG task lacks ecological validity: in real-world communication, humans do not first construct surrounding contexts and then produce referring expressions (REs) accordingly. They formulate REs progressively while producing the text. In this experiment, we want to replicate this behaviour. Concretely, building on WebNLG (Gardent et al., 2017a,b), a widely used NLG dataset, we asked LLMs to first perform end-to-end NLG and then annotate the REs in the texts they generated. We subsequently compared the variation in reference use between the LLM-generated texts and the original human-authored corpus.

We hope this corpus-based evaluation will shed light not only on how LLMs' reference use diverges from human language production but also on how task design may influence the referential behaviour of LLMs.

To further investigate the source of these differences, we examine whether factors known to influence human reference production similarly affect LLMs. Specifically, we focus on three factors grounded in linguistic theories of reference: referential status (Chafe, 1976; Gundel et al., 1993), syntactic position (Brennan, 1995; Arnold, 2010), and recency (Greenbacker and McCoy, 2009; Kibrik et al., 2016). We annotated linguistic information related to these factors in both the LLM-generated texts and the human-authored WebNLG corpus, and compared their respective impacts.

2 Background

In this section, we first provide a brief overview of the REG task and models, followed by an introduction to the WebNLG dataset.

*Corresponding Author

Proceedings of the 18th International Natural Language Generation Conference, pages 182–194
October 29–November 2, 2025.

2.1 Referring Expression Generation

Referring Expression Generation (REG) is one of the main stages of the classic Natural Language Generation (NLG) pipeline (Reiter and Dale, 2000; Krahmer and van Deemter, 2012; van Deemter, 2016). Research on REG typically distinguishes between two tasks. The first, classic REG (also known as one-shot REG), aims to identify a set of attributes that uniquely distinguish a referent from a set of competitors. The second, discourse REG, concerns the generation of referring expressions (REs) within a discourse context. This paper focuses on the latter. As Belz and Varges (2007) put it: *Given an intended referent and a discourse context, how do we generate appropriate referring expressions (REs) to refer to the referent at different points in the discourse?*

In earlier works, REG was often approached as a two-step process (Henschel et al., 2000; Krahmer and Theune, 2002). The first step determines the form of a referring expression (RE), for instance, whether the reference should be a proper name ("*Marie Skłodowska-Curie*"), a description ("*the physicist*"), or a pronoun ("*she*") at a given point in the discourse. The second step concerns content selection, that is, choosing among alternative ways of realising a referential form. For example, to generate a description of *Marie Curie*, the REG system decides whether it is sufficient to mention her profession (i.e., "*the physicist*") or whether it is better to mention her nationality as well (i.e., "*a Polish-French physicist*"). With the rise of deep learning, neural REG models were developed that generate REs in an end-to-end manner, jointly handling both form and content selection (Castro Ferreira et al., 2018a; Cao and Cheung, 2019; Cunha et al., 2020; Chen et al., 2023, 2024).

More recently, motivated by the success of large language models (LLMs), researchers have begun investigating to what extent LLMs can produce human-like REs. However, consistent with broader findings on the limitations of LLMs in pragmatic reasoning (Chang and Bergen, 2024; Beuls and Van Eecke, 2024; Gautam et al., 2024), Ellison and Same (2024) report that LLMs differ markedly from humans, particularly in their inability to capture the variability of human reference use.

2.2 The WebNLG Dataset

The WebNLG corpus was originally developed to evaluate the performance of natural language generation (NLG) systems (Gardent et al., 2017a). Each sample in the corpus consists of a knowledge base represented as a set of RDF triples (see Table 1). To adapt the dataset for the REG task, Castro Ferreira et al. (2018a) and Castro Ferreira et al. (2018b) enriched and delexicalised the corpus. Table 1 illustrates an example of a text generated from an RDF triple along with its corresponding delexicalised version.

Triples: (AWH_Engineering_College, country, India), (Kerala, leaderName, Kochi), (AWH_Engineering_College, academicStaffSize, 250), (AWH_Engineering_College, state, Kerala), (AWH_Engineering_College, city, "Kuttikkattoor"), (India, river, Ganges)
Text: AWH Engineering College is in Kuttikkattoor, India in the state of Kerala. The school has 250 employees and Kerala is ruled by Kochi. The Ganges River is also found in India.
Delexicialised Text: **Pre-context**: AWH_Engineering_College is in "Kuttikkattoor" , India in the state of Kerala . **Target Entity**: AWH_Engineering_College **Pos-context**: has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India .

Table 1: An example data from the WebNLG corpus. In the delexicalised text, every entity is underlined.

Taking the delexicalised text in Table 1 as an example, given the entity "*AWH_Engineering_College*", REG chooses a RE based on that entity and its pre-context ("*AWH_Engineering_College is in "Kuttikkattoor" , India in the state of Kerala .* ") and its pos-context ("*has 250 employees and Kerala is ruled by Kochi . The Ganges River is also found in India .*").

3 Research Questions and Hypotheses

Focusing on our main research question *how do LLMs use REs differently from human beings*, the current study has the following three research questions and hypotheses.

First, we refer to the REG task that requires models to generate RE given a fixed context as Slot Filling REG (SF-REG). In contrast, we define a variant of this task in which models perform full language generation (and REs are subsequently annotated for analysis) as Language Generation REG (LG-REG). Intuitively, LG-REG is more ecologically valid than SF-REG, as it more closely reflects how humans use language in natural contexts.

In this study, we are particularly interested in *how well LLMs capture the variation in human reference use*. Prior work on the coherence of LLM-

generated texts (Beyer et al., 2021) has shown that LLMs tend to overuse proper names in contexts where pronouns would be more natural, suggesting a lack of human-like referential variation. Based on this, we hypothesise that LLMs fail to reproduce human-like variation in both SF-REG and LG-REG settings. However, we expect their performance on LG-REG to be closer to humans due to the greater ecological validity of the task. We denote this as hypothesis $\mathcal{H}_1$.

Furthermore, we expect that the task setting itself significantly affects LLMs' reference use. That is, the variation in REs produced by LLMs in SF-REG will differ significantly from that in LG-REG. We refer to this as hypothesis $\mathcal{H}_2$.

Second, we investigate *whether the factors known to influence human reference production affect LLMs in the same ways*. Specifically, we examine three well-established factors: referential status, syntactic position, and recency. According to linguistic theories of reference, using pronouns is more likely the shorter the distance between a referent and its antecedent (recency; Greenbacker and McCoy (2009)), when the referent is in subject rather than object position (syntactic position; Brennan (1995)), and when it follows a prior mention rather than being introduced for the first time (referential status; Chafe (1976)). We hypothesise that all three would significantly influence LLMs' reference production (hypothesis $\mathcal{H}_3$). However, to account for the observed failure of LLMs to fully capture human-like variation, we further hypothesise that the effects of these factors on LLMs would differ significantly from their effects on human-produced references (hypothesis $\mathcal{H}_4$).

Finally, we are curious about *the impact of model size on reference production*. According to the scaling law (Kaplan et al., 2020), LLMs with more parameters tend to exhibit greater capabilities. We therefore hypothesise that larger LLMs would more effectively capture the variation in human reference use (hypothesis $\mathcal{H}_5$).

4 Method

To test the hypotheses in Section 3, we prompted LLMs to perform both SF-REG and LG-REG on the test set of WebNLG, which is a dataset of NLG from Resource Description Framework (RDF) triples (see Section 2.2 for more information and examples). Unlike previous REG studies using WebNLG (e.g., Castro Ferreira et al. (2018a)), we focus exclusively on REs for the protagonist in each discourse, since we are more interested in the use of REs within a referential chain rather than in single referential mentions. This decision is motivated by the structure of WebNLG: since the dataset is built on DBpedia, most REs refer to a single main referent. In contrast, REs for other referents often occur only once within a discourse and would therefore introduce noise when analysing variation in reference use.

For LG-REG, we instructed LLMs to perform RDF-to-Text generation using the prompt shown in Figure 4 in Appendix A. The example outputs of each model are shown in Table 7 in Appendix B. We then employed an LLM-based protagonist RE annotator (described below) to identify the referring expressions that denote the protagonist in each generated discourse.

For SF-REG, although Castro Ferreira et al. (2018a) provides gold-standard annotations of protagonist REs in WebNLG, we used our protagonist RE annotator to mark these REs as well, ensuring a fair comparison across tasks. After identifying the REs, we removed them from the original texts and prompted the LLMs to regenerate them using the instruction shown in Figure 6 in Appendix A.

Protagonist RE Annotator. We developed the Protagonist RE Annotator using DeepSeek-V3-0324, guided by the prompt shown in Figure 5 in Appendix A. To evaluate its performance, we compared its output against the gold-standard annotations provided by Castro Ferreira et al. (2018a). The annotator demonstrated strong performance, achieving an F1 score of 87.05. Further details of this evaluation are provided in Appendix C.

Extracting Linguistic Features. For each identified RE, we extract linguistic information using predefined rules and the spaCy toolkit. First, we determine the *referential form* —whether the RE is a description, a pronoun, or a proper name—which is used to test the first two hypotheses regarding LLMs' ability to capture variation in reference use. To test hypotheses $\mathcal{H}_3$ and $\mathcal{H}_4$, we extract additional features known to influence human reference production. These include *referential status*, defined at both the discourse and sentence levels: discourse-level status (**DisStat**) distinguishes between discourse-old entities (already mentioned in the previous discourse) and discourse-new entities (not yet mentioned), while sentence-level status (**SenStat**) differentiates between sentence-new and

sentence-old mentions within a sentence. We also identify the *syntactic position* (**Syn**) of each RE, whether it appears in subject or object position, and its *recency* (**Recency**), defined as the number of sentences between the RE and its antecedent. Given that texts in WebNLG contain at most three sentences, recency is categorised as: (a) same sentence, (b) one sentence away, or (c) more than one sentence away.

5 Results

We first introduce the settings of our experiment and report on our testing of the hypotheses from Section 3.

5.1 Experimental Settings

The Choice of LLMs. We used three mainstream LLMs in our experiments: DeepSeek-V3-0326 (Liu et al., 2024), GPT-3.5, and GPT-4o-mini. To investigate the impact of model size on reference production, we also included the full Qwen 2.5 model family (Qwen et al., 2025), which comprises Qwen-2.5-0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B.

Evaluation Metrics. To analyse variation in reference use, we report the distribution of referential forms produced by both LLMs and humans in Table 2. Following prior work (Castro Ferreira et al., 2016a,b; Ellison and Same, 2024), we quantify the divergence between LLM and human distributions using the Jensen-Shannon Divergence (JSD). As complementary metrics, based on the human-authored texts in WebNLG, we also report the BLEU score and the RE accuracy of each LLM on both the SF-REG and LG-REG tasks in Table 8 in Appendix D.

5.2 Testing the Hypotheses

Hypothesis $\mathcal{H}_1$ posits that LLMs would better capture human-like variation in reference use on LG-REG than on SF-REG, due to the greater ecological validity of the former. We observed that, with the exception of GPT-3.5, all models used more pronouns in LG-REG than in SF-REG, which suggests that models may employ referring expressions more naturally in a more ecologically valid setting. However, the JSD scores in Table 2 reveal no consistent trend indicating that either task yields more human-like reference use. Approximately half of the models achieved lower JSD scores on LG-REG, while the other half performed better on SF-REG.

Model	Task	D	PN	P	JSD
Human	-	8.05	77.74	14.22	-
DeepSeek-V3	SF	17.17	69.81	13.01	0.014
	LG	14.63	66.32	19.05	0.013
GPT-3.5	SF	11.23	69.89	18.89	0.006
	LG	11.64	73.46	14.90	0.003
GPT-4o-mini	SF	7.10	73.85	19.06	0.003
	LG	13.57	62.43	24.00	0.020
Qwen-0.5B	SF	11.70	77.41	10.89	0.004
	LG	18.71	59.21	22.08	0.031
Qwen-1.5B	SF	8.41	87.78	3.81	0.025
	LG	12.27	66.06	21.67	0.012
Qwen-3B	SF	4.98	80.28	14.74	0.003
	LG	8.89	74.44	16.67	**0.001**
Qwen-7B	SF	7.15	72.72	20.13	0.004
	LG	11.57	66.06	22.37	0.012
Qwen-14B	SF	11.75	73.94	14.31	0.003
	LG	12.94	71.84	15.23	0.005
Qwen-32B	SF	12.88	72.94	14.17	0.005
	LG	12.45	70.68	16.88	0.005
Qwen-72B	SF	10.47	74.07	15.45	0.002
	LG	13.25	70.86	15.89	0.006

Table 2: The distribution of referential forms of each LLM and its JSD with Human. D, PN, and P mean description, proper name and pronoun, respectively. NB: JSD is the lower the better.

These findings also indicate that hypothesis $\mathcal{H}_5$ does not hold: larger models do not necessarily capture human-like variation in reference use more effectively.

We tested whether the differences between LLM and human referential form distributions were statistically significant. Chi-square tests revealed that all LLMs produced distributions that differed significantly from those of humans ($p < .01$), regardless of the task setting. This divergence may partly stem from the characteristics of the WebNLG texts, which are typically short, formal, and unlike everyday language (Same et al., 2022, 2023). Consequently, human-authored texts in the dataset tend to rely more heavily on proper names and use fewer descriptions or pronouns—patterns that contrast with those observed in most LLM outputs and deviate from our initial expectations.

To evaluate whether task setting significantly influences each LLM's referential behaviour ($\mathcal{H}_2$), we conducted chi-square tests comparing the referential form distributions produced by each model on SF-REG and LG-REG. The results show that, with the exception of Qwen-14B and Qwen-32B, all LLMs exhibit significant differences in reference use across the two settings as expected. Overall, LLMs tend to use more proper names in SF-REG than in LG-REG, though this does not neces-

Model	Task	DisStat	SenStat	Syn	Recency
Human	-	✓	✓	✓	✓
DeepSeek	SF	✓	✓	✓	✓
	LG	✓	✓	×	×
GPT-3.5	SF	✓	✓	✓	✓
	LG	✓	✓	Δ	Δ
GPT-4o-mini	SF	✓	✓	✓	Δ
	LG	✓	✓	Δ	Δ
Qwen-0.5B	SF	✓	✓	✓	Δ
	LG	✓	✓	Δ	Δ
Qwen-1.5B	SF	✓	✓	×	×
	LG	✓	✓	Δ	Δ
Qwen-3B	SF	✓	✓	✓	✓
	LG	✓	✓	Δ	Δ
Qwen-7B	SF	✓	✓	✓	✓
	LG	✓	✓	Δ	Δ
Qwen-14B	SF	✓	✓	✓	✓
	LG	✓	✓	Δ	Δ
Qwen-32B	SF	✓	✓	✓	✓
	LG	✓	✓	Δ	Δ
Qwen-72B	SF	✓	✓	✓	✓
	LG	✓	✓	Δ	Δ

Table 3: Results of testing whether each feature significantly influences pronominalisation in each LLM. A ✓ indicates a significant impact consistent with human ($p<.001$), a × indicates no significant impact, and a Δ indicates a significant impact that differs from the human pattern.

sarily bring them closer to human patterns.

To address hypotheses $\mathcal{H}_3$ and $\mathcal{H}_4$ concerning the factors that influence reference production, we examined how the four features that were extracted (see Section 4) affect pronominalisation (i.e., the use of pronominal versus non-pronominal noun phrases). Before testing these hypotheses on LLMs, we first analysed how these features influence pronominalisation in human-authored texts within the WebNLG dataset. Chi-square tests revealed that humans are significantly more likely to pronominalise when the RE is (1) discourse-old, (2) sentence-old, (3) in subject position, and (4) one sentence away from its antecedent. These findings perfectly align with previous findings in psycholinguistic studies (Gundel et al., 1993; Brennan, 1995; Greenbacker and McCoy, 2009).

We then used chi-square tests to examine whether each factor influences pronominalisation in each LLM (hypothesis $\mathcal{H}_3$). The results are summarised in Table 3. On SF-REG, the pronominalisation of nearly all LLMs is significantly influenced by all four features. The only exceptions are the two smallest Qwen models—Qwen-0.5B and Qwen-1.5B—and, unexpectedly, GPT-4o-mini. In contrast, when LLMs are asked to produce language directly (LG-REG), their pronominalisation remains significantly influenced by the four factors; however, the influence of Syntactic Position (Syn) and Recency diverges from human patterns. For instance, LLMs tend to use more pronouns in object position rather than in subject position, contrary to human tendencies.

Lastly, we conducted two-sample chi-squared tests to examine whether the influence of each factor on LLM pronominalisation significantly differs from that observed in human data, as predicted by hypothesis $\mathcal{H}_4$. The results confirm that, for all LLMs, the effects of these factors differ significantly from those observed in human behaviour. Even in cases where an LLM exhibits the same directional pattern as humans, the degree of influence typically varies.

In sum, regarding the research questions in Section 3, we found that: (1) LG-REG does not lead LLMs to better capture human-like variation in reference production compared to SF-REG; (2) LLMs exhibit different patterns of variation in LG-REG and SF-REG; (3) factors known to impact human reference production also affect LLMs, though always to different degrees and sometimes in different patterns; and (4) larger models do not capture human-like variation more effectively.

6 Conclusion

This study examined how large language models produce referring expressions under different task framings and whether their behaviour aligns with human patterns. We found that LLMs' referential behaviour consistently diverges from human use, both in the overall distribution of referential form choices and in response to factors known to affect human reference production. Notably, task framing significantly affects LLM reference production, but using a more ecologically valid task framing does not lead to better alignment with human patterns. These findings highlight persistent gaps between human and LLM reference production.

Limitations

One key limitation of this study lies in the choice of the dataset. All experiments were conducted using the WebNLG corpus, which, while widely used in NLG research, has known limitations in reflecting everyday language use. WebNLG texts are typically short, structured, and formal, and they are generated from RDF triples, which constrain both content and style. Prior studies (e.g., Same et al. (2022, 2023)) have shown that human-authored

texts in WebNLG differ from naturally occurring discourse, particularly in their higher reliance on proper names and reduced use of pronouns or descriptive references. This domain-specific and stylised nature may limit the generalisability of our findings to broader or more conversational contexts. Future work should address this by evaluating LLM reference production on corpora that more closely reflect spontaneous human communication, such as dialogue or narrative datasets, and by extending evaluation to languages beyond English (Chen, 2022), in order to further validate and strengthen the conclusions drawn here.

Acknowledgments

We are grateful for the comments from reviewers. Guanyi Chen is supported by the start-up funds of Central China Normal University (No.31101232053) and Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning (No.2025AISL002).

References

Jennifer E Arnold. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4(4):187–203.

Anja Belz and Sebastian Varges. 2007. Generation of repeated references to discourse entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 9–16, Saarbrücken, Germany. DFKI GmbH.

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(3):1277–1311.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Susan E Brennan. 1995. Centering attention in discourse. *Language and Cognitive processes*, 10(2):137–167.

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3163–3172, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016a. Individual variation in the choice of referential form. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 423–427, San Diego, California. Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016b. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018a. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018b. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic*.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Guanyi Chen. 2022. *Computational generation of Chinese noun phrases*. Ph.D. thesis, Utrecht University.

Guanyi Chen, Fahime Same, and Kees van Deemter. 2023. Neural referential form selection: Generalisability and interpretability. *Computer Speech & Language*, 79:101466.

Guanyi Chen, Fahime Same, and Kees Van Deemter. 2024. Intrinsic task-based evaluation for referring expression generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7220–7231, Bangkok, Thailand. Association for Computational Linguistics.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2261–2272, Barcelona, Spain (Online). International Committee on Computational Linguistics.

T. Mark Ellison and Fahime Same. 2024. Experimental versus in-corpus variation in referring expression choice. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6838–6848, Torino, Italia. ELRA and ICCL.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Charles Greenbacker and Kathleen McCoy. 2009. UDel: generating referring expressions guided by psycholinguistic findings. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 101–102. Association for Computational Linguistics.

Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Renate Henschel, Hua Cheng, and Massimo Poesio. 2000. Pronominalization revisited. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 306–312. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Andrej A. Kibrik, Mariya V. Khudyakova, Grigory B. Dobrov, Anastasia Linnik, and Dmitrij A. Zalmanov. 2016. Referential choice: Predictability and its limits. *Frontiers in Psychology*, 7:1429.

Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. *Information sharing: Reference and presupposition in language generation and interpretation*, 143:223–263.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ehud Reiter. 2017. A commercial perspective on reference. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 134–138, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

Fahime Same, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural models matter: a re-evaluation of neural referring expression generation systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5554–5567, Dublin, Ireland. Association for Computational Linguistics.

Fahime Same, Guanyi Chen, and Kees van Deemter. 2023. Models of reference production: How do they withstand the test of time? In *Proceedings of the 16th International Natural Language Generation Conference*, pages 93–105, Prague, Czechia. Association for Computational Linguistics.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

A Prompts

We hereby list the prompts we used in this study:

1. Figure 4 shows the prompt for RDF-to-text generation;
2. Figure 5 is the prompt for protagonist referring expression annotation.
3. Figure 6 is the prompt for SF-REG.

B Example Outputs

Table 7 shows the example outputs of LLMs based on the same set of Triples.

C Protagonist RE Annotator

In Section 4, we briefly introduced the construction of our LLM-based Protagonist RE Annotator. Here, we provide details on how its effectiveness was evaluated. Using the delexicalised WebNLG corpus (see Table 1 for an example), we extracted the boundaries of all REs referring to the protagonist in each discourse. We then compared these gold-standard boundaries with those predicted by our annotator. The evaluation yielded a precision of 85.23, a recall of 88.96, and an F1 score of 87.05.

D Complementary Results of LLMs

Table 8 presents complementary results to those reported in Table 2, including the following: (1) BLEU scores, (2) RE accuracy, (3) whether the variation distribution of each LLM differs significantly from that of humans, and (4) whether the variation distributions produced by each LLM differ significantly across the two task framings (SF-REG vs. LG-REG).

=== Goal === Your task is to generate a logically coherent and descriptive paragraph based on the given set of triplets. You should seamlessly integrate all the relationships expressed in the triplets into a natural-flowing text, avoiding rigid or list-like structures.
=== Return Format === Provide only the descriptive text generated from the triplets. Do not include any additional explanations or notes.
=== Warnings === Return plain and concise text. This means you should include all relevant information from the triplets, while strictly avoiding overly long or verbose descriptions.
=== Context === A triplet is a structured way to represent relationships between entities. It consists of three parts: the subject, the relation, and the object. The subject is the main entity that we want to start the description from. It can be a person, a thing, an event, etc. The relation describes how the subject and the object are connected. It might not always follow a strict subject - verb - object pattern, like in 'operator' relation, which shows who is in charge. The object is the entity that is associated with the subject through the specified relation. In each set of triplets, there is usually a main subject that we want to focus on for the overall description.
=== Example === Triplets: (Alan Bean, mission, Apollo 12) (Alan Bean, nationality, United States) (Apollo 12, operator, NASA) (Alan Bean, occupation, Test pilot) (Alan Bean, birth place, Wheeler, Texas) (Alan Bean, status, Retired) (Alan Bean, birth date, 1932 - 03 - 15) Text: Alan Bean was born on March 15, 1932 in Wheeler, Texas and is American. He worked as a test pilot and was a member of Apollo 12, which was run by NASA. Bean is retired.
=== Your Task === Now, please generate a descriptive text for the following triplets:

Table 4: Prompt for RDF-to-Text generation.

=== Goal ===
Your task is to annotate all references to the given entity in the text by wrapping them in angle brackets < >.

=== Return Format ===
Return the result in plain text.

=== Warnings ===
1. Annotate every word/phrase that refers to the given entity
2. Wrap complete references in < > (don't split words)
3. Include articles if they're part of the reference (e.g., <a/an/the astronaut>)
4. Handle all forms:
   - Full names (<Alan Bean>)
   - Pronouns (<he>, <his>)
   - Nicknames (<Bean>)
   - Possessive forms (<Bean's>)
5. Do not generate any content other than the text annotated with < >, such as explanations, descriptions, or instructions.

=== Example ===
Entity: Alan Bean
Original Text: Alan Bean was born in Wheeler, Texas, and earned a Bachelor of Science degree from UT Austin in 1955. He worked as a test pilot and was a member of the Apollo 12 mission, which was operated by NASA. Among his fellow crew members was David Scott. Bean is now retired. Bean's wife is also retired.
Annotated Text: <Alan Bean> was born in Wheeler, Texas, and earned a Bachelor of Science degree from UT Austin in 1955. <He> worked as a test pilot and was a member of the Apollo 12 mission, which was operated by NASA. Among <his> fellow crew members was David Scott. <Bean> is now retired. <Bean's> wife is also retired.

=== Your Task ===
Entity: `{main_entity}`
Original Text: `{input_text}`
Annotated Text:

Table 5: Prompt for LLM-based protagonist referring expression annotation.

=== Goal ===
Your task is to fill in the appropriate entity references in "< >" based on the given entity names and the context to be filled in.

=== Return Format ===
Return the filled content in plain text format.

=== Warnings ===
1. Keep the sentence structure unchanged:
   - Only fill in the blanks within "< >".
   - Preserve all original words/punctuation exactly
   - Maintain original capitalization and grammatical number
2. Sometimes you need to include articles like "a," "an," or "the" in the entity references within "< >".
3. Only return the text with "< >" filled in using references, without generating any other content.

=== Examples ===
[Person]
Entity: Alan Bean
Original Text: < > was born in Texas. < > became an astronaut in 1963. < > wife is an astronaut, too.
Restored Text: <Alan Bean> was born in Texas. <He> became an astronaut in 1963. <Bean's> wife is an astronaut, too.

[Object]
Entity: The Eiffel Tower
Original Text: < > is a famous landmark. Many tourists visit < > every year.
Restored Text: <The Eiffel Tower> is a famous landmark. Many tourists visit <it> every year.

[Organization]
Entity: NASA
Original Text: < > was founded in 1958. Many scientists work at < >.
Restored Text: <NASA> was founded in 1958. Many scientists work at <the agency>.

[Location]
Entity: Paris
Original Text: < > is the capital of France. < > is known for its beautiful architecture.
Restored Text: <Paris> is the capital of France. <It> is known for its beautiful architecture.

=== Your Task ===
Entity: `{entity}`
Original Text: `{masked_text}`
Restored Text:

Table 6: Prompt for slot filling REG.

Model	Text
Human	AWH Engineering College in Kuttikkattoor, India is in the state of Kerala which has Mahe to its northwest. The school has 250 employees. India is home to the Ganges River and its largest city is Mumbai.
DeepSeek-V3	AWH Engineering College is located in Kuttikkattoor, a city in the state of Kerala, India. Kerala lies northwest of Mahé, India. The college has a faculty size of 250 and is part of India, which is home to Mumbai, its largest city, and the Ganges River.
GPT-3.5	India, located in Asia, houses its largest city, Mumbai, renowned for its vibrant culture and bustling activity. AWH Engineering College, situated in the state of Kerala, India, boasts a sizeable faculty of 250 members and is based in the city of Kuttikkattoor. Kerala, in the southwest of India, borders Mahé in the northwest. Moreover, the sacred Ganges river flows through the heart of India, enriching the land with its cultural and historical significance.
GPT-4o-mini	AWH Engineering College, located in Kuttikkattoor, Kerala, India, boasts a faculty size of 250 dedicated educators. Kerala is situated in the northwest region of India, near Mahé, with Mumbai recognized as the country's largest city. The Ganges River, a significant feature of India, further emphasizes the rich cultural and geographical diversity of the nation.
Qwen-0.5B	AWH Engineering College is located in the state of Kerala, which is situated in the northwest part of India. The college has a faculty size of 250 students and serves as the headquarters of AWH Engineering College. The Ganges River flows through the city of Kuttikkattoor, making it one of the most important water bodies in the region.
Qwen-1.5B	Mumbai, the largest city in India, serves as the capital of Maharashtra state, located in the western part of the subcontinent. AWH Engineering College, situated in Kerala's state capital, Kochi, has a faculty size of 250 students. The college is also located in Kuttikkattoor, another city within the same state. The region of Kerala borders Mahé, another Indian territory. The Ganges River flows through northern India, including the states of Uttar Pradesh and Bihar.
Qwen-3B	India's largest city is Mumbai, located in the state of Maharashtra. AWH Engineering College, situated in Kuttikkattoor, Kerala, has a faculty size of 250 and is part of India, whose longest river is the Ganges.
Qwen-7B	India, a vast country with the largest city being Mumbai, is home to AWH Engineering College located in the state of Kerala. The college, situated in the city of Kuttikkattoor, boasts a faculty size of 250 members. To the northwest of Kerala lies Mahé, another region within India. The Ganges, one of India's most significant rivers, flows through various parts of the country, adding to its rich cultural and geographical heritage.
Qwen-14B	India, home to the Ganges river, has Mumbai as its largest city. AWH Engineering College is located in Kuttikkattoor, a city in the state of Kerala, which is situated northwest of Mahé, India. The college has a faculty size of 250.
Qwen-32B	AWH Engineering College is located in Kuttikkattoor, a city in the state of Kerala, India. The college has a faculty size of 250 and is situated in a country where Mumbai stands as the largest city. Kerala borders Mahé to the northwest, and the Ganges River flows through India, adding to its rich geographical diversity.
Qwen-72B	India, a country known for its vast cultural diversity and the mighty Ganges River, has Mumbai as its largest city. In the state of Kerala, located partly northwest of Mahé, lies AWH Engineering College in the city of Kuttikkattoor. The college, which is part of the Indian education system, boasts a faculty size of 250 members dedicated to providing quality engineering education.

Table 7: Texts generated by humans and LLMs based on the following set of Triples: (AWH_Engineering_College, country, India), (Kerala, leaderName, Kochi), (AWH_Engineering_College, academicStaffSize, 250),(AWH_Engineering_College, state, Kerala), (AWH_Engineering_College, city, "Kuttikkattoor"),(India, river, Ganges).

Model	Task	BLEU	Acc.	D	PN	P	JSD	Human Sig. Diff.	Task Sig. Diff.
Human	-	-	-	8.05	77.74	14.22	-	-	-
DeepSeek	SF	86.85	64.33	17.17	69.81	13.01	0.014	✓	✓
	LG	17.17	-	14.63	66.32	19.05	0.013	✓	
GPT-3.5	SF	82.04	62.91	11.23	69.89	18.89	0.006	✓	✓
	LG	15.50	-	11.64	73.46	14.90	0.003	✓	
GPT-4o-mini	SF	85.02	62.43	7.10	73.85	19.06	0.003	✓	✓
	LG	10.64	-	13.57	62.43	24.00	0.020	✓	
Qwen-0.5B	SF	79.22	44.27	11.70	77.41	10.89	0.004	✓	✓
	LG	11.25	-	18.71	59.21	22.08	0.031	✓	
Qwen-1.5B	SF	76.86	58.79	8.41	87.78	3.81	0.025	✓	✓
	LG	11.59	-	12.27	66.06	21.67	0.012	✓	
Qwen-3B	SF	86.76	66.12	4.98	80.28	14.74	0.003	✓	✓
	LG	13.47	-	8.89	74.44	16.67	**0.001**	✓	
Qwen-7B	SF	82.73	54.49	7.15	72.72	20.13	0.004	✓	✓
	LG	11.67	-	11.57	66.06	22.37	0.012	✓	
Qwen-14B	SF	85.07	66.12	11.75	73.94	14.31	0.003	✓	×
	LG	17.71	-	12.94	71.84	15.23	0.005	✓	
Qwen-32B	SF	86.37	64.33	12.88	72.94	14.17	0.005	✓	×
	LG	14.76	-	12.45	70.68	16.88	0.005	✓	
Qwen-72B	SF	87.32	65.86	10.47	74.07	15.45	0.002	✓	✓
	LG	13.82	-	13.25	70.86	15.89	0.006	✓	

Table 8: Complementary results to those in Table 2, including (1) BLEU scores, (2) RE accuracy, (3) whether the referential form distribution of each LLM differs significantly from that of humans, and (4) whether the referential form distributions differ significantly between the two task settings (SF-REG vs. LG-REG) for each LLM.