# Polish Coreference Corpus as an LLM Testbed: Evaluating Coreference Resolution within Instruction-Following Language Models by Instruction–Answer Alignment

**Karol Saputa** and **Angelika Peljak-Łapińska** and **Maciej Ogrodniczuk**

Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland

karolsaputa@gmail.com,
{angelika.peljak,maciej.ogrodniczuk}@ipipan.waw.pl

## Abstract

In this article, we analyse coreference resolution in encoder- and decoder-based approaches in the Polish language. We convert the Polish Coreference Corpus into the instructions suitable for training language models and create supplementary data based on examples that are difficult for encoder-based models, analyse them and create additional questions for more precise mention boundary detection and other ambiguities found.

We propose an evaluation framework for our instructions. The best closed model, Claude 3 Sonnet, achieves 44.52 CoNLL $F_1$ in instruction following, zero-shot setting, which is surpassed by the fine-tuned Llama 3.1 8B model, which achieves 46.54 $F_1$.

## 1 Introduction

Coreference resolution (CR) is traditionally a part of classical natural language processing (NLP) pipeline tasks, treated as a discriminative problem. Until recently, most of the solutions were encoder-based architectures (Liu et al., 2023; Martinelli et al., 2024). Generative approach has been discussed as an alternative, starting with the formulation of coreference resolution as a question answering task (Wu et al., 2020) and the advancements in language models. Thus, a comparison between these two approaches is needed.

A broad focus on large language models with a high number of parameters (Touvron et al., 2023; Dubey et al., 2024), which can be easily trained using human-readable formats of training data, provides an opportunity to reframe the CR problem and improve results. Improvements in encoder-based solutions, which are in exchange much faster (thanks to smaller models), may lead to easier applicability of CR in NLP pipelines.

In this article, we analyse coreference resolution in encoder- and decoder-based approaches and discuss the possible advantages of generative modelling in coreference resolution. Our research case is the Polish language.

For this task, both groups of models are evaluated, an error analysis is conducted, and the potential of providing supplemental Winograd-like fine-tuning for LLMs is explored.

Smaller LLMs, such as Llama 3.1 8B, fine-tuned on our instructions achieve results comparable to bigger, commercial, closed models such as Claude 3 Opus. However, these results are far below the levels of custom architectures. These results support the focus of further research on building of new training resources for the Polish language.

## 2 Related Work

The following Section analyses elements of coreference resolution evaluation related to comparing encoder and decoder approaches.

### 2.1 Coreference Evaluation

The main resource for CR in the Polish language is the Polish Coreference Corpus (PCC) (Ogrodniczuk et al., 2016) which has been included in the multilingual coreference dataset, CorefUD (Nedoluzhko et al., 2022). The most commonly compared CR metric is the CoNLL $F_1$ score. This metric, along with others, can be calculated by the coreference scorer (Yu et al., 2023) which evaluates coreference predictions in the CorefUD format and has been used in CR challenges (Žabokrtský et al., 2022, 2023).



*Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2024)*, pages 23–32
November 15, 2024. 

### 2.2 Language Models in Coreference Resolution

There have been multiple LLM-based coreference resolution systems proposed recently that can be grouped into two categories: (1) LLMs usage is limited to annotating texts in a specific format as in (Hicke and Mimno, 2024; Le and Ritter, 2023; Gan et al., 2024), (2) LLM is incorporated into processing framework as a part of an algorithm e.g. controlling the incremental input to LLM and decoding it (Bohnet et al., 2023), extracting mentions via LM (Skachkova et al., 2023). This system is considered the best known to us solution for the English language.

The first approach (1) requires fewer steps of work. There is no custom data modelling, architecture, or optimisation needed, only supervised fine-tuning of a language model. The annotation schema in this approach can be not expressive enough. For example, the approach of Hicke and Mimno (2024) does not include any texts with minor text alterations in the evaluation, only evaluates exact match scores and requires strict matching of index clusters. Gan et al. (2024) does not analyse the detection of mentions and uses gold mentions instead.

The second approach (2) gives state of the art results thanks to language models' great common sense reasoning about language and world knowledge. Bigger pre-trained models tend to score higher in CR benchmarks (Hicke and Mimno, 2024), as in other tasks. However, in this second approach, there is still a custom architecture needed and coreference reasoning cannot be used directly to improve the general LM performance.

### 2.3 Encoder-based Solutions

Best-performing solutions for coreference resolutions have moved to an end-to-end, encoder-based approach (Lee et al., 2017), which has been further improved (Kirstain et al., 2021). The Maverick system (Martinelli et al., 2024) presents several improvements to the state-of-the-art encoder-based end-to-end architecture for the English language. Most importantly, it sets the maximum mention span length as a sentence level parameter based on sentence length[1].

These improvements lead Maverick to achieve scores comparable to decoder-based solutions but with a much shorter inference time. However, the benchmark results for coreference resolution plateaued at slightly above 80% CoNLL $F_1$ score. An encoder-based approach requires modelling of all edge cases in data structures and model architecture. The gains from corrections and inclusion of new edge-cases are small. For example, the CAW system (D'Oosterlinck et al., 2023) improves the score of the earlier model for the 0.9 CoNLL $F_1$ score.

### 2.4 Polish Language

Previous attempts to evaluate coreference resolution in the Polish language have been outlined by Saputa (2022) who compares the transformer-based end-to-end approaches with previous systems and discusses dataset-specific modelling for Polish. The performance of models in the Polish language was also discussed as a part of multilingual systems in recent Shared Tasks on Multilingual Coreference Resolution (Žabokrtský et al., 2022, 2023).

## 3 Challenges for Current Coreference Resolution Systems

### 3.1 Beyond Annotation in Coreference Resolution

Due to the typical formulation of the task, a prediction of a set of clusters of coreferential mentions, the error analysis of the models is difficult in both qualitative and quantitative way. This was addressed by developing different CR metrics and tools for error analysis, e.g., the taxonomy of errors (Kummerfeld and Klein, 2013). Most importantly, the score of coreference resolution (the correct grouping of mentions into coreferential clusters) cannot be higher than the mention detection score (the correct recognition of all mentions in the text with their proper span limits). This means that mention detection (and the definition of a mention) has a strong impact on the overall coreference resolution score.

[1]It should be noted that the Polish Coreference Corpus contains multi-sentence mentions which are not detected by this architectural approach. The inclusion of longer mentions in the training set, which are more numerous (372 mentions with more than 35 tokens), could yield comparable advantages as in the case of multi-sentence mentions (223 mentions) from a modelling perspective. There is a 91-mention overlap between these two categories: multi-sentence, very long mentions. However, reducing memory overhead is of substantial benefit to the training process.

### 3.2 Sentence-level Reasoning

One of the frequently occurring errors in mention recognition involves subject clauses, both subordinate (Example 1) and coordinate (Example 2):

(1) *Zresztą fundacje musiałyby rozbudowywać do tych celów jakieś specjalne aparaty urzędniczo-śledcze, co jest absurdem.*
'Besides, NGOs would have to develop some special clerical and investigative apparatus for these purposes, which is absurd.'

(2) *Wydłużyła się droga dzieci do szkół i to także budzi powszechne niezadowolenie.*
'The journey of children to school has lengthened and this, too, is causing widespread dissatisfaction.'

Mentions were often not detected in similar contexts where mention coreferentiality answers the questions of *who?* or *what?*, as in the examples above. The effectiveness of the algorithm is similarly low in the case of mentions in adverbial clauses. Thus, these types of problems were addressed in Section 4.2.

## 4 Dataset

We convert the Polish Coreference Corpus (Ogrodniczuk et al., 2016) into the instruction format for the evaluation of language models that is suitable for training coreference resolution in the generative approach. The dataset consists of the converted, annotated texts, and two types of supplementary data. The additional data taken from the original collection that is inspired by Winograd-like challenge and post-training approaches to language models: (1) question-answering datasets of examples that are difficult for encoder-based CR model to answer correctly, (2) preferences for answer style and reasoning between models. These supplementary data are motivated by the problems described in Section 3.

### 4.1 Conversions of PCC into Instructions

The instructions use two formats: bracket-style and list-style. In brackets format, the answer should include the original text of the prompts with mentions annotated in brackets referring to the cluster id (Appendix B.1) e.g.: `[Man]:1`. In list format, the model is asked to construct in its answer a list of clusters with all mentions listed for each cluster (Appendix B.2). The second format is resembling a chain-of-thought, incrementally focusing on next entity in the text. Table 1 presents the number of instructions of each type.

| Instructions | Examples | | |
|---|---|---|---|
| | Train | Dev | Test |
| Brackets-style | 1463 | 183 | 182 |
| List-style | 1463 | 183 | 182 |
| QA-style | 59 | 7 | 8 |
| Preferences | 59 | 7 | 8 |

Table 1: Details of the instructions provided for LM training. Examples are the entire texts (Brackets/List-style) or sentences (QA-style and Preferences).

### 4.2 Extracting Difficult Examples from the Corpus

Difficult examples for encoder-based models were selected from the dataset after evaluating the encoder-based model at the sentence level. The sentences with the lowest CoNLL $F_1$ score were analysed and used to create additional questions for more precise mention boundary detection and more context for other ambiguities found.

QA-style supplemental data (1) is aimed at improving the detection of correct mention boundaries and reasoning about unclear examples in the style of Winograd questions, which require the model to behave as if it was performing commonsense reasoning and possessed knowledge (Cozman and Munhoz, 2020). Preference-style supplementary data (2) is meant to improve the reasoning and explanatory coherence of the model answers, especially when there are multiple possible interpretations, that are resolved by annotators agreement, about which there is no information in a dataset used by encoder-based models. In this context, we refer to the discussion of examples in Section 3.1. In Appendix B.4, an example question is shown with a gold answer and GPT-4o answer that shows both the importance of correct mention boundary detection and coreference reasoning.

### 4.3 Generating Artificial Examples with LLMs

We used the available language models, GPT 3.5 and LLama 3.1 8B, to generate answers for the prepared questions and assess the preferences between models in terms of correct answer, justification of the answer, precision of citation, and use of appropriate vocabulary.

| System | Open | FT | IF rate | MD $F_1$ | CoNLL $F_1$ partial match | Precision exact match |
|---|---|---|---|---|---|---|
| GPT-4o | ✗ | ✗ | 89.80 | 32.00 | 24.60 | 51.06 |
| GPT-4o-mini | ✗ | ✗ | 64.00 | 19.32 | 14.68 | 23.85 |
| Claude 3 Sonnet | ✗ | ✗ | 100.00 | 47.88 | 44.52 | 62.22 |
| Claude 3 Opus | ✗ | ✗ | 100.00 | 48.30 | 36.70 | 69.13 |
| Claude 3 Haiku | ✗ | ✗ | 84.62 | 25.94 | 30.36 | 38.12 |
| Llama-3.1-70B | ✓ | ✗ | 26.32 | 2.99 | 0.81 | 3.72 |
| Llama-3.1-8B | ✓ | ✗ | 1.78 | 0.56 | 0.34 | 0.00 |
| Llama-3.1-8B-FT | ✓ | ✓ | 100.00 | 57.80 | 46.54 | 62.89 |
| s2e-herbert-large | ✓ | ✓ | — | 78.40 | 69.91 | 73.21 |
| s2e-herbert-base | ✓ | ✓ | — | 75.53 | 62.85 | 70.27 |

Table 2: Instruction following results of coreference resolution evaluation: Instruction Following (IF) rate, mention detection $F_1$ (MD), CoNLL $F_1$ measure on the PCC development set. The following instruction does not apply to `s2e` models as the correct output is asserted by their custom architecture. Evaluation concerns commercial models, open models, and fine-tuned (FT) open models.

## 5 Evaluation

### 5.1 Generative Answers Parsing and Alignment

We first tested several prompts on a small development set and then chose one instruction (Appendix B.3) that produced the highest prompt follow-up rate in the tests. This prompt was used in the evaluation of language models in generative coreference resolution.

The text alignment technique (Boyd et al., 2024) was used to match the fragment of each model's generative response to text from the PCC dataset. This is an effective algorithm that allows the modified text (answer) to be matched with the original on the level of individual tokens. Thus, allowing for different tokenization and modifications. Even if the generative answer has a modified version of the texts, the mentions, provided they are intact, should be matched with the original text tokens. This makes it possible to evaluate coreference resolution in general not fine-tuned models whose answers typically include other comments and reasoning in addition to machine-annotated text and have error-prone and alterations-prone evaluation pipelines. This approach also takes into account all possible comments from the model at the beginning and at the end of the text.

The annotation format (Appendix B.1) presents a bracketed format to annotate coreference relations. Such annotated spans are extracted using a regular expression and grouped by cluster id (`[mention_span]:cluster_id`). Text alignment allows for comparison of span indices in each cluster with indices in gold clusters in the dataset, and it also enables writing the prediction back to the conllu file, preserving the original tokenisation. Such conllu files are then evaluated using the coreference scorer.

### 5.2 Instruction Following

The following instruction is a type of task that does not involve fine-tuning of a model, with only the prompt instructing the model about the task (Zhou et al., 2023). The prompt does not include an example of a complete solution, so it can be described as a zero-shot setting.

The instruction following (IF) rate is a measure of the compliance of a language model with the instructional requirements. We measure IF rate as the correct use of the annotation schema, i.e. non-zero results in the mention detection score. This allows for errors, but reflects at least one correct application of the schema described in the instruction.

In Table 2. we present the results of the evaluation. The IF rate ranges from 26.32% for Llama 3.1 70B to 100% for Claude Opus. Precision scores have been included to demonstrate that the models typically annotate a smaller number of coreference relations than the gold standard annotations, but the predictions are more accurate than the CoNLL $F_1$ score would suggest. This reflects the issue of task modelling discussed in Section

3.1, which considers the challenge of annotating a large number of relations for each text.

### 5.3 Instruction Fine-Tuning

We tested the smallest LLama 3.1 model (8B parameters) with supervised fine-tuning for 4 epochs using `SFTTrainer`[2] from the Huggingface ecosystem accustomed to the training infrastructure (see Acknowledgements).

The training used the following default parameters: BF16 precision, batch size of 1, AdamW optimiser, WarmupDecayLR scheduler, maximum sequence length of 8192 tokens, and automatic gradient accumulation. We did not perform any kind of hyper-parameter optimisation apart from tests of prompt instruction formulation (Appendix B.3) that were evaluated on not-tuned models for only a few texts from the training part of the dataset.

In Table 2, we describe results from the development part of the CorefUD Polish dataset, as there is a publicly available gold standard for this part. The fine-tuned model performed better than the best non-tuned model, Claude 3 Sonnet. However, its results are much lower than our reproduction of the results of the start-to-end architecture (Kirstain et al., 2021) that was adapted for the Polish language by Saputa (2022). Table 2 shows scores of non-tuned models, fine-tuned Llama and s2e results[3].

## 6 Conclusions and Future Work

We proposed a conversion of the Polish Coreference Corpus (PCC) into instructions suitable for generative training, as an adaptation of the coreference resolution for generative models, as well as the evaluation framework for bracket-style answers. There potential for further ablation studies and interaction studies of the proposed resources; for example, we did not provide here an extensive analysis of the difference between training on bracket- and list-style instructions and training on the preferences data. These resources are aimed at reformulation of the coreference resolution dataset format and going beyond standard annotations to handle more fuzziness than is possible using existing available resources.

The first results of the fine-tuning are better than the available commercial and open-source models. The differences in results between open models, commercial models, and the fine-tuned model indicate that commercial models may have been trained on similar types of instructions. Thus, it is important to develop non-commercial datasets and models as alternatives for further advancements of natural language processing in the Polish language.

However, the highest score is much lower than the encoder-based approach discussed for Polish and the decoder-based approaches discussed for English. It means that: (1) custom encoder architectures should be used in specific applications that require coreference resolution, and (2) solving multiple coreference chains during text generation is difficult in the setting proposed in our research.

[2] https://huggingface.co/docs/trl/sft_trainer

[3] It is worth to note that the encoder-based results obtained here are slightly lower than the Shared Tasks state-of-the-art results for Polish. However, since the difference between the performance of the generative modelling is more than 20 points, we did not focus on the improvements.

## Acknowledgements

Work completed as part of a project funded by the Polish Minister of Digital Affairs under a special purpose subsidy No. 1/WI/DBiI/202 "Responsible development of the open large language model, PLLuM (Polish Large Language Universal Model) aimed at supporting breakthrough technologies in the public and economic sectors, including an open, Polish-language intelligent assistant for public administration clients."

We would like to thank Piotr Pęzik and Konrad Kaczyński for providing the code for fine-tuning the models that they designed for the LLM training infrastructure used within the PLLuM project.

## References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference Resolution through a seq2seq Transition-Based System. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Adriane Boyd, Daniël de Kok, Matthew Honnibal, and Basile Dura. 2024. spacy-alignments. Original-date: 2020-12-08T08:07:25Z.

Fábio Cozman and Hugo Munhoz. 2020. The Winograd Schemas from Hell. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 531–542. SBC. ISSN: 2763-9061.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and

Chris Develder. 2023. CAW-coref: Conjunction-Aware Word-level Coreference Resolution. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Laven-

der A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs].

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the Capabilities of Large Language Models in Coreference: An Evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Rebecca Hicke and David Mimno. 2024. [Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference Resolution without Span Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-Driven Analysis of Challenges in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Nghia T. Le and Alan Ritter. 2023. Are Large Language Models Robust Coreference Resolvers? arXiv:2305.14489 [cs].

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and Accurate Coreference Resolution Defying Recent Trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference Meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2016. Polish Coreference Corpus. In *Human Language Technology. Challenges for Computer*

*Science and Linguistics*, pages 215–226, Cham. Springer International Publishing.

Karol Saputa. 2022. Coreference Resolution for Polish: Improvements within the CRAC 2022 Shared Task. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 18–22, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Natalia Skachkova, Tatiana Anikina, and Anna Mokhova. 2023. Multilingual coreference resolution: Adapt and Generate. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 19–33, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs].

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference Resolution as Query-based Span Prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Juntao Yu, Michal Novák, Abdulrahman Aloraini, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2023. The Universal Anaphora Scorer 2.0. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 183–194, Nancy, France. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the Second Shared Task on Multilingual Coreference Resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the Shared Task on Multilingual Coreference Resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

# A Other methods

## A.1 Adaptation of English Winograd Schema

Translating the English Winograd Schema into Polish proved unsuccessful in most respects due to structural differences between the languages. Those differences do not concern the English-Polish pair exclusively. Emelin and Sennrich (2021), working with German, French, and Russian, found "that not all WinoGrande samples are suitable for the inclusion in Wino-X, as replacing the "gap" [token in place of an ambiguous pronoun in each schema, which can be filled by one of two preceding nouns] with "it" can yield ungrammatical or disfluent sequences" (p. 8518)

Emelin and Sennrich also used certain heuristics to filter out cases that would be difficult to translate, but most of those heuristics, however, do not apply to the Polish language. Moreover, the WinoMT dataset was quality checked with the use of Python grammar checker, also known as OpenOffice spellchecker, and it proved to be insensitive to syntax and stylistic errors, which usually disqualify most Polish translations of Winograd Schema Challenge examples.

Translation attempts revealed that only a handful of ambiguous structures present in the original schema are in fact ambiguous and both grammatically and stylistically correct in Polish.

In our search for difficult examples of coreference, we also carried out a literature review, aimed specifically at finding sentences and texts containing mentions that should be ambiguous for the language model but should not pose a challenge for a human. This method also gave unsatisfactory results.

## A.2 Creating New Examples based on Samples Found in Previous Efforts

We got a handful of examples that proved difficult for an existing model, but there was no apparent pattern connecting those instances.

# B Instruction Details

## B.1 Generative Answer Schema

This is a fragment of text with id 307 (with original punctuation). Mentions with cluster index appearing only once appear later in the text. Singletons (mentions appearing only once, not coreferential) are omitted.

> kompletnie nie [zgadzam]:0 się z tą interpretacją że ruch. To wskazanie [Marka Belki]:1 jest obliczone na pozyskanie przez [Bronisława Komorowskiego]:2 [elektoratu centrolewicowego]:3 wszystkie badania pokazują że [ten elektorat centrolewicowy]:3 jest zdecydowany głosować na. [Komorowskiego]:2. SLD z całym szacunkiem ma te między pięć a siedem procent twardego elektor elektoratu lewicowego a nie [centrolewicowego]:3 to po pierwsze po drugie wydaje [mi]:0 się że. akurat [mam]:0. prawo bronić [decyzji. [marszałka Komorowskiego]:2 żeby już teraz zgłaszać [kandydata na [prezesa [banku]:7]:6]:5]:4 bo po pierwsze od początku [mówiłem]:0 że akurat [ta instytucja]:7 w przeciwieństwie do niektórych innych.

Following English translation of the above fragment:

> [I]:0 completely disagree with this interpretation that the movement. This indication of [Marek Belka]:1 is calculated to win over [Bronisław Komorowski]:2 [the centre-left electorate]:3 all polls show that [this centre-left electorate]:3 is determined to vote for. [Komorowski]:2. The SLD with all due respect has those between five and seven percent of the hard left electorate and not [centre-left electorate]:3 this is first of all, secondly it seems to [me]:0 that. just [I]:0 have. the right to defend [the decision of. [Speaker Komorowski]:2 to announce [a candidate for [the president of the [bank]:7]:6]:5]:4 already now because firstly from the beginning [I]:0 said that exactly [this institution]:7 unlike some others.

## B.2 List-style Instruction

This is fragment of the list-style answer generated for text 307. Singletons (mentions appearing only once, not coreferential) are omitted.

> grupa (1): *zgadzam*, *mi*, *mam*, *mówiłem*, *ja*, *moja*, *ja*, *przyjmowałem*, *mi*
>
> grupa (2): *Marka Belki*, *Marek Belka*
>
> grupa (3): *Bronisława Komorowskiego*,

*Komorowskiego*, *marszałka Komorowskiego*, *marszałka Komorowskiego*, *marszałek*

grupa (4): *elektoratu centrolewicowego*, *ten elektorat centrolewicowy*, *centrolewicowego*

grupa (5): *decyzji marszałka Komorowskiego żeby już teraz zgłaszać kandydata na prezesa banku*, *ta decyzja*

grupa (6): *kandydata na prezesa banku*, *jakiejś kandydatury*

grupa (7): *prezesa banku*, *tym prezesem*

grupa (8): *banku*, *ta instytucja*, *ta instytucja*, *bank*, *on*

Following English translation of the above fragment:

group (1): *I*, *me*, *I have*, *I said*, *I*, *my*, *I*, *I accepted*, *me*

group (2): *Mark Belka*, *Marek Belka*

group (3): *Bronislaw Komorowski*, *Komorowski*, *marshal Komorowski*, *marshal Komorowski*, *marshal*

group (4): *centre-left electorate*, *this centre-left electorate*, *centre-left*

group (5): *the decision of Marshal Komorowski to already put forward a candidate for bank president*, *this decision*

group (6): *a candidate for bank president*, *some candidacy*

group (7): *the bank president*, *this president*

group (8): *the bank*, *this institution*, *this institution*, *the bank*, *it*

### B.3 Instruction Following Prompt

Zaznacz relacje koreferencji w poniższym tekście za pomocą nawiasów kwadratowych i indeksów wspólnej referencji - `[zakres wzmianki]:indeks_grupy` np. `[syn [jednej z [Polek]:3]:2]:1`. Zwróć uwagę na dokładne granice wzmianek i ich kolejność. Tekst:

Following English translation of the above fragment:

Mark the coreference relations in the following text using square brackets and subscripts of the common reference - `[mention range]:index_group` e.g. `[son of [one of [Poles]:3]:2]:1`. Note the exact boundaries of the mentions and their order. Text:

### B.4 Winograd-like Questions

Below we include one exemplary sentence-level question in the Winograd style from the development part of the QA-style dataset that has a wrong answer from the GPT-4o model.

**Question:** Odpowiedz na poniższe pytanie. Napisz wyłącznie samą odpowiedź lub przynajmniej powtórz dokładną odpowiedź osobno w ostatniej linii. Zacytuj dokładny fragment, do którego odnosi się 'to' w zdaniu: "Wprawdzie już zapoznał się z naszymi broszurami, ale to mu nie wystarcza, chciałby przeprowadzić wywiady z dostojnikami, przyjrzeć się naszemu życiu z bliska". Odpowiedz wyłącznie cytatem z tekstu.

**GPT-4o answer:** ...ale to mu nie wystarcza...

**Gold answer:** zapoznał się z naszymi broszurami

Following English translation of the above fragment:

**Question:**. Answer the following question. Write only the answer itself or at least repeat the exact answer separately on the last line. Quote the exact passage to which 'it' refers in the sentence: 'Although he has already familiarised himself with our brochures, but this is not enough for him, he would like to interview the dignitaries, take a close look at our life'. Respond with a quote from the text only.

**GPT-4o answer:** ...but that is not enough for him...

**Gold answer:** has familiarised himself with our brochures