Optimizing LLMs for Polish Reading Comprehension

A Comparative Study of Ensemble and Unified Approaches

Krzysztof Wróbel

Enelpol

>_ SpeakLeash

/'spix.lɛʂ/ a.k.a Spichlerz

BIELIK

PoQuAD - Reading Comprehension

"**title**": "Karnawał",
"**summary**": "Karnawał, zapusty – okres zimowych balów, maskarad, pochodów i zabaw. Rozpoczyna się najczęściej w dniu Trzech Króli, a kończy we wtorek przed Środą Popielcową, która oznacza początek wielkiego postu i oczekiwania na Wielkanoc.",
"**url**": "https://pl.wikipedia.org/wiki/Karnawał",
"**context**": "Dawniej w zapusty jedzono dużo tłustych dań, bardzo popularne były placki ziemniaczane. Jedzono wiele, jakby chciano zaspokoić głód przed zbliżającym się postem. Ludzie bardzo biedni starali się w tym dniu zjeść choć trochę mięsa wierząc w to, że kto w zapusty nie je mięsa, tego komary przez lato zjedzą. W tym czasie smażono też pączki, faworki i bliny. Taniec, zabawa, śpiewy, to nieodzowne elementy zapustne. Do dzisiaj zachował się tam zwyczaj chodzenia w zapusty przebierańców. Spotkać można wśród nich postacie z kolędowania, niedźwiedzia, konia, bociana, cygana czy żandarma.",
{"**question**": "Dlaczego w zapusty ludzie spożywali więcej jedzenia?",
"**generative_answer**": "chciano zaspokoić głód przed zbliżającym się postem"
"**is_impossible**": false}
{"**question**": "Jakie danie na zapusty było najpopularniejsze w bogatych domach?",
"**is_impossible**": true}

Dataset statistics

Table 1: Dataset statistics showing total number of examples and number of nonanswerable questions per split. The average and maximum lengths are measured in tokens.

Split	Total	Nonanswerable	Average token length	Max token length
train	56,618	10,431 (18.42%)	409.24	3,427
dev	7,060	1,296 (18.36%)	412.40	1,598
test-A	3,501	-	399.26	2,036
test-B	3,585	-	410.20	2,237

- Average context length: ~410 tokens
- Maximum context length: 3,427 tokens

Evaluation

Answerability Score - whether a question can be answered based on the given context; binary F1 score, where the positive class represents questions that are not answerable

Levenshtein Score - Levenshtein edit distance between the predicted and ground truth answers, after converting both to lowercase. The distance is normalized by the length of the longer sequence

Final - arithmetic mean of the Answerability and Levenshtein scores

Methods

LLM - long-context decoder model, LoRA, SFT

CausalLM head for joint prediction of both answerability and answer generation through fine-tuning.

SequenceClassification head specifically for the answerability task

1. Single model
2. Ensemble system

Experiments - Prompts

1. A binary classification prompt to determine if the context contains an answer to the question:
Tytuł: {title}
Kontekst: {context}
Pytanie: {question}
Czy kontekst jest relewantny dla pytania? Odpowiedź:

2. A direct question-answering prompt that generates an answer based on the provided context:
Kontekst: {context}
Pytanie: {question}
Odpowiedz krótko i zwięźle na powyższe pytanie. Odpowiedź:

3. A conditional answering prompt that generates an answer only if the context contains relevant information:
Tytuł: {title}
Kontekst: {context}
Pytanie: {question}
Jeśli kontekst zawiera odpowiedź na powyższe pytanie to odpowiedz krótko i zwięźle, a jeśli kontekst nie zawiera odpowiedzi to napisz: "Brak informacji". Odpowiedź:

5-shot Answerability

Model	Prompt 1			Prompt 3		
	Precision	Recall	F1	Precision	Recall	F1
Meta-Llama-3.1-405B-Instruct-FP8	87.08%	50.46%	63.90%	78.26%	64.74%	70.86%
Mistral-Large-Instruct-2407	80.67%	52.16%	63.36%	70.42%	67.05%	68.70%
Qwen2.5-72B-Instruct	77.58%	50.46%	61.15%	75.93%	60.11%	67.10%
Qwen2-72B-Instruct	59.74%	57.72%	58.71%	64.96%	56.94%	60.69%
Mixtral-8x22B-Instruct-v0.1	60.13%	49.23%	54.14%	78.53%	48.53%	59.99%
Meta-Llama-3-70B-Instruct	80.56%	44.14%	57.03%	81.47%	43.75%	56.93%
Qwen2-72B	56.51%	58.26%	57.37%	64.00%	43.75%	51.97%
openchat-3.5-0106-gemma	79.02%	25.00%	37.98%	65.87%	42.75%	51.85%
Meta-Llama-3.1-70B-Instruct	75.48%	51.77%	61.42%	83.99%	34.41%	48.82%
Bielik-11B-v2.0-Instruct	56.84%	53.86%	55.31%	71.65%	21.45%	33.02%
Meta-Llama-3-70B	72.17%	36.42%	48.41%	59.50%	14.74%	23.62%
Bielik-7B-Instruct-v0.1	35.00%	24.85%	29.06%	28.10%	9.80%	14.53%

5-shot Levenshtein similarity

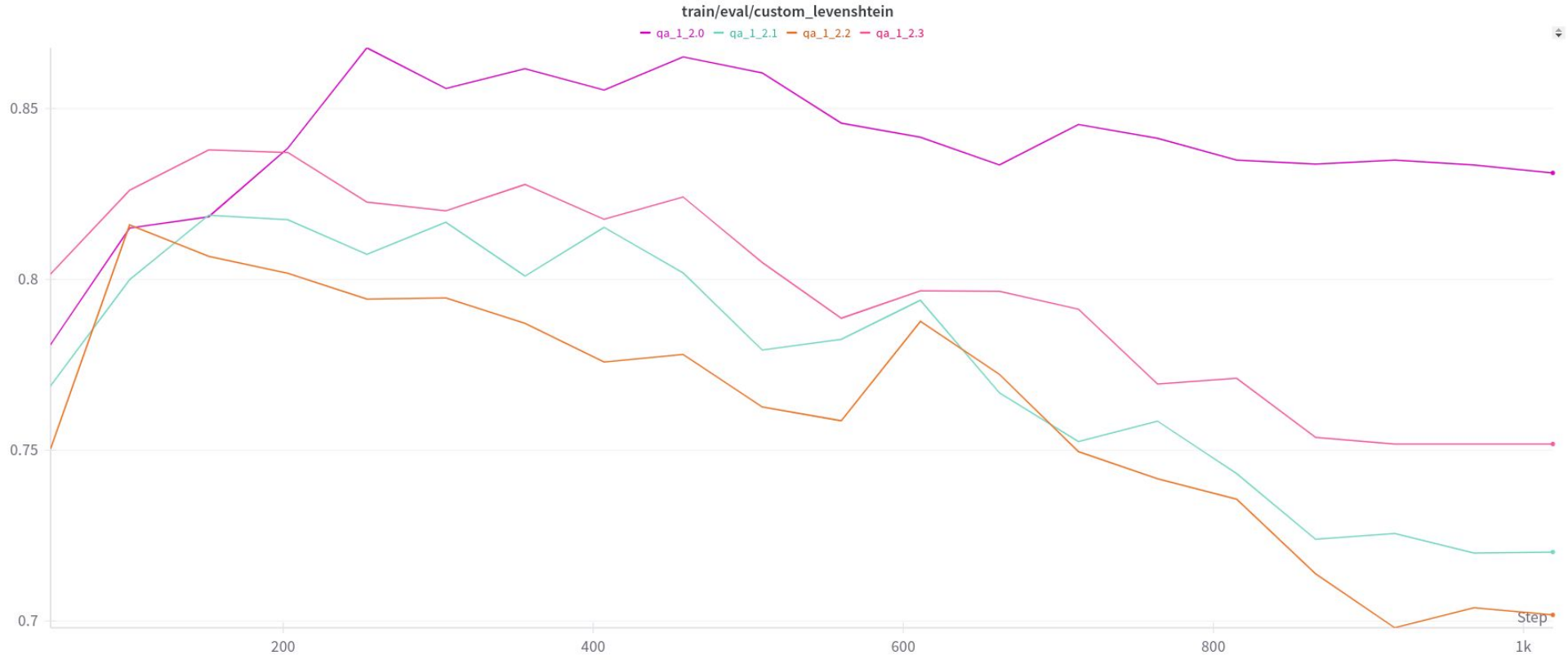
Model	Params (B)	Prompt 2 with Oracle	Prompt 3
Meta-Llama-3.1-405B-Instruct-FP8	405	83.87%	81.26%
Qwen2-72B	72	83.59%	78.69%
Mistral-Large-Instruct-2407	123	82.04%	76.84%
Qwen2-72B-Instruct	72	81.83%	75.50%
Mixtral-8x22B-Instruct-v0.1	141	81.64%	79.25%
Meta-Llama-3-70B	70	81.44%	80.29%
Bielik-11B-v2.0-Instruct	11	81.16%	78.77%
Qwen2.5-72B-Instruct	72	80.91%	77.88%
Meta-Llama-3-70B-Instruct	70	80.83%	79.33%
Meta-Llama-3.1-70B-Instruct	70	80.51%	79.89%
openchat-3.5-0106-gemma	7	78.98%	75.37%
Bielik-7B-Instruct-v0.1	7	70.62%	63.07%

3 models based on Bielik-11B-v2

Table 4: Comparison of training parameters for the three model variants. Model 1 focuses on answer generation for answerable questions only, Model 2 specializes in answerability classification, and Model 3 represents a balanced approach handling both tasks.

	Prompt 2	Prompt 1	Prompt 3
Parameter	**Model 1**	**Model 2**	**Model 3**
Learning rate	2e-4	2e-4	1e-4
Scheduler type	Linear	Linear	Linear
Warmup	None	None	None
Batch size	16	16	16
Max steps	1,000	7,000	3,500
Eval steps	100	500	100
LoRA r	16	16	16
LoRA alpha	16	16	16
Max sequence length	1,024	1,024	1,024
Head type	CausalLM	SequenceClassification	CausalLM
Data subset	Answerable only	Original	Original
Title included	No	Yes	Yes

Levenshtein similarity by Bielik versions



Results

87.15 for oracle

	dev-0			test-A			test-B		
	Ans.	**Lev.**	**Score**	**Ans.**	**Lev.**	**Score**	**Ans.**	**Lev.**	**Score**
2 models	81.73	84.42	**83.07**	81.44	86.08	**83.76**	82.33	83.52	**82.92**
1 model	77.44	83.42	80.43	79.09	85.64	82.36	77.94	82.24	80.09
GPT 3.5 few shot	48.20	67.25	57.73	-	-	-	-	-	-
plT5 baseline	57.67	83.25	70.46	-	-	-	-	-	-

Llama 405B: Ans. 71%, Lev. 81%

Finetuned Bielik-1.5B-Instruct: Ans. 68%, Lev. 81%

Error Analysis - Answerability

- Question Error: "**Czy kiedykolwiek ORP Zwinny przeszedł remont grawitacyjny?**" - should have been "gwarancyjny" instead of "grawitacyjny".
- Unclear Question: "**Jakie wyłącznie gazety funkcjonują na terenie Szczecina?**"
- Imprecise Question: "**Jak nazywał się prezydent Syrii, z którym się przyjaźnił?**" - about who or what was being referred to.
- Different Assessment: "**Po publikacji 11 maja 2019 filmu Tylko nie mów nikomu, pod adresem księdza Makulskiego padły oskarżenia o kontakty seksualne z osobami małoletnimi. Bohaterem jakiego filmu został Eugeniusz Makulski?**".

Error Analysis - Levenshtein similarity

- Grammatical inflection variations in Polish words
- Presence or absence of prepositions
- Different number formats (numerical vs written form)
- Use of Roman vs Arabic numerals
- Incomplete personal names (surnames without given first names)
- Presence or absence of quotation marks and other punctuation

Examples of errors 1/3

Lev.	Question	Reference answer	Predicted answer
0.98	Dlaczego młode tego gatunku ...?	gdyż **mają** za słabo wykształconą warstwę tłuszczową, by przeżyć w wodach Arktyki lub Antarktyki	gdyż **ma** za słabo wykształconą warstwę tłuszczową, by przeżyć w wodach Arktyki lub Antarktyki
0.97	Na co wskazuje ... kryzys religii ...?	religia nie przetrwa jeśli nie współgra z elementarnym rozumnym postrzeganiem świata	**że** religia nie przetrwa jeśli nie współgra z elementarnym rozumnym postrzeganiem świata
0.96	W jakich krajach ...?	**w** Austrii, Finlandii, Izraelu, Holandii i Hiszpanii	Austrii, Finlandii, Izraelu, Holandii i Hiszpanii
0.95	W jakich miejscach ...?	**w** organizacjach, instytucjach i firmach	organizacjach, instytucjach i firmach
0.94	Jaką teorię na temat tkanki ... posiadał ...?	powstaje ona z metaplazji nabłonka otrzewnej	**że** powstaje ona z metaplazji nabłonka otrzewnej

Examples of errors 2/3

0.91	Czego wyrazem było ...?	**podkreślić rolę** Alego w swoich wierzeniach	**podkreślenia roli** Alego w swoich wierzeniach
0.91	Jakie przebudowy ... miały miejsce ...?	obmurowano krużganki i zmieniono wnętrza zamku, rozebrano część budynków (m.in. kościoły św. Jerzego i św. Michała)	obmurowano krużganki i zmieniono wnętrza zamku, rozebrano część budynków (m.in. kościoły św. Jerzego i św **[reached max tokens]**)
0.90	W meczu z jakim klubem ...?	**z** Rakowem Częstochowa	Rakowem Częstochowa
0.90	W jakim kraju ...?	**W** Jordanii	**w** Jordanii
0.80	Jaki statek...?	Civilian	„Civilian"
0.87	Od czego zależy ...?	**od** występowania deszczy	występowania deszczy
0.80	Z jakiego powodu ...?	**Powodem wojny miała być** polityka Ozeasza, króla Izraela, który odmówił płacenia trybutu Asyrii i sprzymierzył się z Egiptem	polityka Ozeasza, króla Izraela, który odmówił płacenia trybutu Asyrii i sprzymierzył się z Egiptem

Examples of errors 3/3

0.71	O ile zmniejsza się ...?	**o** 18,4%	18,40%
0.66	Jakim typem państw były ...?	teokratyczny**ch** islamski**ch państw**	teokratyczny**mi** islamski**mi**
0.57	Kto kierował ...?	**lejtnant** W. Moczulski	W. Moczulski
0.55	Kiedy Polska ...?	30 września 1938 **o godz 23:45**	30 września 1938
0.53	W jakim celu ...?	**Na cele** wystawowe	wystawowe
0.48	Ile razy dziennie ...?	dwa lub trzy **razy na dobę**	dwa lub trzy
0.29	W jaki sposób ...?	rozpad budynku nie postąpił dalej **i reszta wieżowca, choć też uszkodzona, „osiadła" na gruzach zmiażdżonych pięter**	rozpad budynku nie postąpił dalej
0.11	W której lidze ...?	**w** I **lidze**	I

Key Findings

1. Ensemble approach consistently outperforms unified model
2. Trade-off between performance and computational efficiency
3. Levenshtein metric may be overly sensitive to linguistic variations
4. 80% of low Levenshtein score answers were semantically correct

Future Work

1. Modified output format for unified model
2. Joint training of classification and generation
3. Enhanced few-shot learning capabilities
4. Dataset quality improvements:
    a. Better question quality control
    b. Semantic similarity metrics
5. Fine-tune other (smaller) models

Thank you

Source code and models are available at:

https://github.com/enelpol/poleval2024-task1

Contact:

contact@enelpol.com

https://www.linkedin.com/in/wrobelkrzysztof/

Supported by PLGrid ACK Cyfronet AGH (Grant PLG/2024/016951)

Enelpol

>_ SpeakLeash

/'spix.lɛʂ/ a.k.a Spichlerz

BIELIK