Polish linguistic and cultural competency benchmark

🗋 Paper 🗐 Cite

AUTHORS
Sławomir Dadas, Małgorzata Grębowiec, Michał Perełkiewicz, Rafał Poświata

AFFILIATION
National Information Processing Institute

UPDATED
20/01/2026

Large language models (LLMs) are becoming increasingly proficient in processing and generating multilingual texts, which allows them to address real-world problems more effectively. However, language understanding is a far more complex issue that goes beyond simple text analysis. It requires familiarity with cultural context, including references to everyday life, historical events, traditions, folklore, literature, and pop culture. A lack of such knowledge can lead to misinterpretations and subtle, hard-to-detect errors. To examine language models' knowledge of the Polish cultural context, we introduce the **Polish Linguistic and Cultural Competency Benchmark**, consisting of **600 manually crafted questions**. The benchmark is divided into six categories: history, geography, culture & tradition, art & entertainment, grammar, and vocabulary. This evaluation provides a new perspective on Polish competencies in language models, moving past traditional natural language processing tasks and general knowledge assessment.

Recently added models: GLM-4.7-Flash (20/01/2026), Bielik-11B-v3.0-Instruct (31/12/2025), GLM-4.7 (23/12/2025), Gemini-3-Flash-Preview (19/12/2025), GPT-5.2-2025-12-11 (xhigh reasoning) (14/12/2025)

Leaderboard

Filter models XLSX CSV HTML

Model	Provider	Average (6 categories) ▼	art & entertainment	culture & tradition	geography	grammar	history	vocab
Gemini-3.0-Pro-Preview	Google	95.83	95	99	100	91	95	95
Gemini-2.5-Pro-Preview-06-05	Google	92.17	91	96	98	86	92	90
Gemini-3-Flash-Preview	Google	91.67	91	98	96	85	92	88
GPT-5-Pro-2025-10-06 (high reasoning)	OpenAI	91.00	88	94	96	85	91	92
Grok-4	xAI	90.50	86	95	94	90	94	84
Gemini-2.5-Pro-Exp-03-25	Google	89.50	88	91	97	79	92	90
GPT-5-2025-08-07	OpenAI	89.50	85	89	97	84	91	91
GPT-5.2-2025-12-11 (xhigh reasoning)	OpenAI	89.33	79	93	94	89	94	87
O1-2024-12-17	OpenAI	89.17	86	92	95	84	90	88
O3-2025-04-16	OpenAI	89.17	83	91	97	85	89	90
GPT-5.1-2025-11-13 (high reasoning)	OpenAI	88.83	85	90	97	82	89	90
GPT-5.2-2025-12-11 (high reasoning)	OpenAI	87.17	78	87	95	87	90	86
GPT-4.5-preview-2025-02-27	OpenAI	86.50	90	92	90	74	90	83
GPT-5.2-2025-12-11 (medium reasoning)	OpenAI	85.00	74	84	94	82	90	86
Gemini-2.5-Flash-Preview-04-17	Google	83.50	78	85	94	77	86	81
Gemini-Exp-1206	Google	83.00	83	90	86	69	88	82
Claude-3.5-Sonnet-20241022	Anthropic	82.67	77	87	85	79	91	77
GPT-4o-2024-05-13	OpenAI	82.33	83	92	89	70	82	78
Claude-3.7-Sonnet-Thinking	Anthropic	82.17	77	82	87	80	92	75
Claude-3.7-Sonnet	Anthropic	81.50	80	83	87	74	90	75
GPT-4o-2024-08-06	OpenAI	81.33	82	89	88	66	86	77
GPT-4o-2024-11-20	OpenAI	81.33	82	89	86	67	84	80
DeepSeek-V3.2-Speciale	DeepSeek	81.00	71	76	94	84	90	71
Claude-3.5-Sonnet-20240620	Anthropic	80.67	73	85	86	75	89	76
Claude-Opus-4.5	Anthropic	80.33	74	82	84	79	87	76
GPT-4.1-2025-04-14	OpenAI	80.33	77	84	89	67	85	80
Claude-Opus-4.1	Anthropic	79.00	67	83	86	74	91	73
GPT-5.2-2025-12-11 (no reasoning)	OpenAI	78.83	70	86	86	69	85	77
Claude-Opus-4	Anthropic	78.67	72	81	83	76	87	73
DeepSeek-v3.1 (thinking)	DeepSeek	78.67	69	76	89	75	89	74

Model	Provider	Average (6 categories) ▼	art & entertainment	culture & tradition	geography	grammar	history	vocab
GPT-5.1-2025-11-13 (default reasoning)	OpenAI	77.83	72	82	86	70	82	75
GPT-5-mini-2025-08-07	OpenAI	77.50	62	74	94	82	83	70
Grok-3-Beta	xAI	77.17	71	90	83	65	85	69
DeepSeek-R1-0528	DeepSeek	76.17	65	75	85	73	91	68
DeepSeek-R1	DeepSeek	76.00	66	75	84	74	85	72
Gemini-2.0-Flash-Thinking-Exp-01-21	Google	74.83	72	76	84	68	80	69
Gemini-2.0-Flash-Experimental	Google	74.17	68	78	79	65	83	72
Claude-3-Opus	Anthropic	73.83	73	76	80	66	86	62
GLM-4.7	Zhipu AI	73.50	64	79	88	66	85	59
O4-Mini-2025-04-16	OpenAI	72.83	62	73	88	72	77	65
Grok-4.1-Fast	xAI	72.33	54	74	85	72	84	65
DeepSeek-V3.2	DeepSeek	71.67	61	78	78	66	82	65
Kimi-K2-Thinking	Moonshot.AI	71.67	63	71	84	73	80	59
Grok-3-Mini-Beta	xAI	71.33	61	67	84	71	84	61
Claude-Sonnet-4.5	Anthropic	71.00	61	72	79	68	85	61
DeepSeek-v3-0324	DeepSeek	71.00	64	76	78	64	82	62
DeepSeek-v3.1 (no thinking)	DeepSeek	71.00	63	69	82	64	86	62
Bielik-11B-v3.0-Instruct	SpeakLeash	70.67	69	78	75	57	78	67
GLM-4.6	Zhipu AI	70.67	59	76	82	63	87	57
Mistral-Large-2512	Mistral	70.67	63	75	76	67	79	64
Grok-4-Fast	xAI	70.17	59	71	79	72	81	59
DeepSeek-v3.2-Exp	DeepSeek	70.00	59	71	80	63	83	64
Gemini-Pro-1.5	Google	69.67	62	77	74	58	79	68
PLLuM-12B-nc-chat-250715	PLLuM	69.67	72	75	79	52	73	67
DeepSeek-v3	DeepSeek	69.17	61	73	79	62	77	63
Claude-Sonnet-4	Anthropic	68.17	55	72	77	63	81	61
PLLuM-8x7B-nc-chat	PLLuM	68.17	72	76	73	47	73	68
GPT-4-turbo	OpenAI	67.00	61	74	79	56	76	56
Mistral-Medium-3	Mistral	66.83	56	67	77	61	78	62
GLM-4.5	Zhipu AI	66.50	56	68	79	59	77	60
Grok-2-1212	xAI	66.00	57	67	77	64	74	57
Bielik-2.6	SpeakLeash	65.50	61	68	75	55	72	62
Llama-3.1-Tulu-3-405B	Meta	63.83	64	64	71	56	75	53
Bielik-2.2	SpeakLeash	63.00	54	60	72	53	77	62
GPT-5-nano-2025-08-07	OpenAI	62.50	47	59	80	69	73	47
Bielik-2.3	SpeakLeash	62.17	58	61	68	49	76	61
GPT-4.1-mini-2025-04-14	OpenAI	62.17	51	62	75	62	67	56
Bielik-2.5	SpeakLeash	62.00	52	61	72	51	75	61
Kimi-K2	Moonshot.AI	62.00	50	67	70	58	73	54
Qwen3-Max	Alibaba	61.33	50	57	75	58	74	54
Bielik-2.1	SpeakLeash	61.00	55	64	68	50	73	56

Model	Provider	Average (6 categories)	art & entertainment	culture & tradition	geography	grammar	history	vocab
Kimi-K2-0905	Moonshot.AI	61.00	54	63	67	59	70	53
Llama-3.1-405b	Meta	60.00	56	57	74	57	73	43
GPT-4	OpenAI	59.50	49	63	67	58	72	48
PLLuM-12B-nc-chat	PLLuM	59.50	59	65	70	41	70	52
O3-mini-2025-01-31	OpenAI	59.33	46	51	78	67	67	47
Llama-PLLuM-70B-chat	PLLuM	58.50	49	64	68	50	74	46
Llama-4-Maverick	Meta	58.17	46	52	71	59	76	45
Llama-PLLuM-70B-chat-250801	PLLuM	58.00	54	62	63	54	69	46
Claude-3.5-Haiku-20241022	Anthropic	57.83	43	62	72	57	61	52
GPT-4o-mini-2024-07-18	OpenAI	56.83	42	57	69	55	67	51
Claude-3.0-Sonnet	Anthropic	56.50	46	53	65	56	73	46
Command-A-03-2025	Cohere	56.17	44	55	67	49	73	49
Qwen3-235B-A22B	Alibaba	55.00	37	45	69	66	70	43
GLM-4.5-Air	Zhipu AI	54.67	48	51	64	52	66	47
GPT-OSS-120b	OpenAI	54.33	42	46	71	64	65	38
Qwen3-Next-80B-A3B-Thinking	Alibaba	54.33	43	45	64	65	72	37
Mistral-Large-2407	Mistral	54.17	48	52	63	51	71	40
PLLuM-8x7B-chat	PLLuM	54.17	45	60	66	42	68	44
Mistral-Large-2411	Mistral	52.00	39	52	61	54	64	42
O1-mini-2024-09-12	OpenAI	51.67	38	44	66	61	61	40
WizardLM-2-8x22b	Microsoft	51.50	45	50	60	49	67	38
Qwen-Max	Alibaba	50.83	43	50	53	51	63	45
Claude-Haiku-4.5	Anthropic	50.67	36	52	52	59	60	45
Command-R-Plus-08-2024	Cohere	50.17	44	49	61	43	61	43
Mixtral-8x22b	Mistral	49.83	45	41	59	50	69	35
Command-R-Plus-04-2024	Cohere	49.33	39	52	53	45	61	46
Llama-3.3-70B	Meta	48.83	43	40	59	49	65	37
Llama-3.1-70B	Meta	47.83	42	41	58	44	68	34
Gemma-3-27b	Google	47.33	43	55	51	46	52	37
PLLuM-12B-chat	PLLuM	47.00	48	49	54	37	61	33
Bielik-0.1	SpeakLeash	46.67	43	52	61	29	58	37
Gemini-Flash-1.5	Google	46.50	33	41	61	46	51	47
Mistral-Small-3.2-24B-2506	Mistral	46.17	38	39	51	53	61	35
GPT-4.1-nano-2025-04-14	OpenAI	43.67	30	40	59	45	50	38
GPT-3.5-turbo	OpenAI	43.33	39	38	55	41	51	36
Mistral-Small-3.1-24B-2503	Mistral	43.33	35	39	45	50	54	37
Llama-3.0-70B	Meta	43.00	40	38	49	45	64	22
Qwen3-Next-80B-A3B-Instruct	Alibaba	43.00	34	36	46	52	58	32
Gemma-2-27b	Google	42.67	32	41	47	46	53	37
Bielik-4.5B-v3.0-Instruct	SpeakLeash	42.33	28	44	53	35	55	39
GLM-4.7-Flash	Zhipu AI	42.33	31	40	55	44	54	30

Model	Provider	Average (6 categories)	art & entertainment	culture & tradition	geography	grammar	history	vocab
Llama-4-Scout	Meta	41.50	23	35	51	51	47	42
EuroLLM-9B	UTTER	41.00	30	40	54	39	49	34
Magistral-Small-2506	Mistral	39.33	30	29	45	47	54	31
Qwen-2.5-72b	Alibaba	39.17	25	30	45	45	54	36
Ministral-14b-2512	Mistral	39.00	25	29	45	44	52	39
Mistral-Small-24B-2501	Mistral	39.00	33	29	42	45	49	36
Llama-PLLuM-8B-chat	PLLuM	38.50	33	34	46	33	50	35
Qwen-Plus	Alibaba	38.50	26	32	42	47	46	38
Qwen3-32B	Alibaba	37.67	21	28	37	48	55	37
Mixtral-8x7b	Mistral	35.33	31	27	44	34	56	20
Ministral-8b-2512	Mistral	35.17	20	30	39	44	43	35
Qwen3-30B-A3B	Alibaba	33.00	19	30	31	49	42	27
GPT-OSS-20b	OpenAI	32.33	19	26	35	54	37	23
Qwen-2.5-32b	Alibaba	30.50	17	21	25	43	44	33
Qwen3-14B	Alibaba	30.33	14	16	30	46	42	34
Gemma-2-9b	Google	29.17	19	23	30	38	35	30
Phi-4	Microsoft	29.17	23	17	35	34	40	26
Qwen-Turbo-2024-11-01	Alibaba	28.50	15	20	30	33	42	31
Bielik-1.5B-v3.0-Instruct	SpeakLeash	27.50	27	25	35	23	32	23
Qwen-2.5-14b	Alibaba	26.67	21	17	23	34	37	28
Qwen3-8B	Alibaba	26.00	12	13	27	38	41	25
Mistral-Nemo	Mistral	23.00	20	13	26	31	28	20
Command-R7B	Cohere	22.83	14	18	33	23	27	22
Llama-3.1-8B	Meta	22.67	19	13	31	29	25	19
Ministral-3b-2512	Mistral	22.33	11	17	24	30	30	22
Mistral-7b-v0.3	Mistral	21.83	22	9	27	27	30	16
Ministral-8b	Mistral	20.67	14	12	19	24	33	22
Qwen-2.5-7b	Alibaba	17.67	5	11	17	29	23	21

Examples

The table below presents sample questions from our benchmark. We selected 10 questions from each category, aiming to ensure thematic and structural diversity. Click on the tabs to switch between categories.

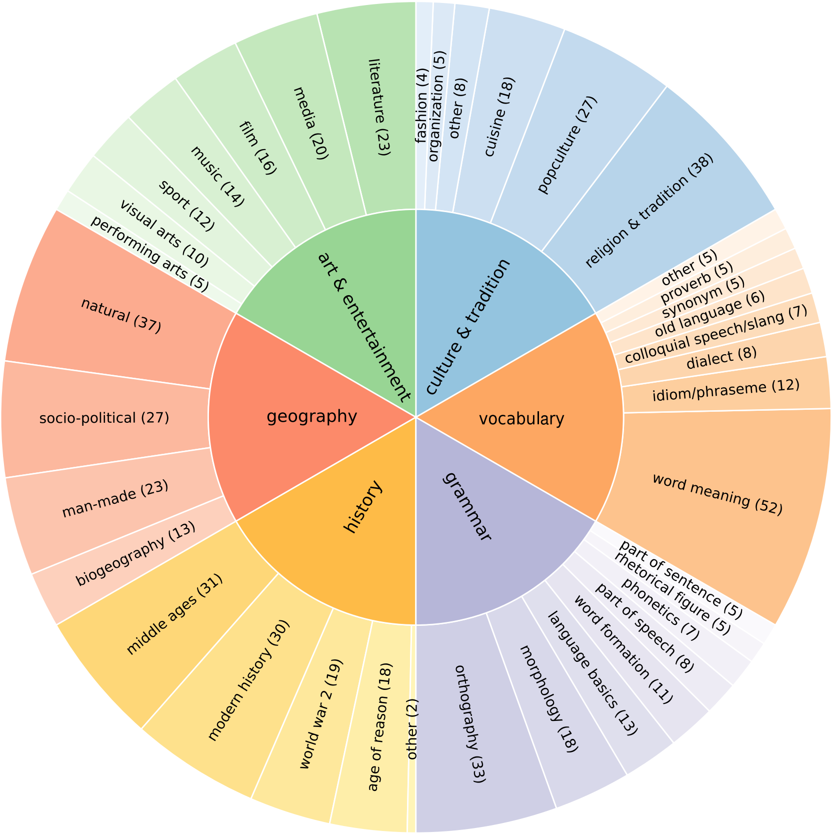
culture & tradition | art & entertainment | geography | history | grammar | vocab

ID	Question	Verification	Subcategory
6	Jak nazywała się subkultura młodzieżowa funkcjonująca w Polsce po II wojnie światowej, którą cechował bunt wobec narzuconych norm oraz fascynacja muzyką jazzową i kulturą amerykańską, za co jej przedstawiciele byli prześladowani przez władze PRL?	**include**: {bikiniarze, bikiniarzy, bikiniarstwo}	popculture
23	Jaki gatunek małpy stał się bohaterem polskich memów o "typowym Januszu"?	**include**: {nosacz sundajski, nosacza sundajskiego, nasalis larvatus}	popculture
38	Który z poniżej wymienionych klasztorów jest najstarszym z istniejących klasztorów w Polsce? 1. Opactwo benedyktynów w Tyńcu 2. Klasztor zakonu paulinów na Jasnej Górze 3. Opactwo benedyktynów na Świętym Krzyżu 4. Opactwo cystersów w Lubiążu	**include**: 1 **exclude**: 2, 3, 4	religion & tradition

ID	Question	Verification	Subcategory
	Podaj pojedynczą liczbę 1, 2, 3 lub 4 odpowiadającą poprawnej odpowiedzi. Nie dodawaj komentarza.		
42	Jakie produkty Polacy tradycyjnie wkładają do wielkanocnej święconki?	**include**: {jajko, jajka}, {chleb, pieczywo}, {mięso, wędlina, kiełbasa}, sól, baranek, chrzan, {ciasto, babka} **params:** include_min=5	religion & tradition
73	Który z poniższych wypieków nie ma owalnego kształtu z dziurką w środku? A. kołacz B. obwarzanek C. bajgiel D. donut E. kołocz śląski Odpowiedz tylko jedną literą, bez dodatkowego komentarza.	**include**: E **exclude**: A, B, C, D	cuisine
74	Jaki rodzaj ciasta stał się sławny dzięki papieżowi Janowi Pawłowi II?	**include**: kremówka	cuisine
77	Na początku lat 90-tych Polską wstrząsnęło zabójstwo znanego muzyka oraz jego kochanki dokonane przez jej męża, wówczas reżysera filmowego. Jak nazywały się ofiary?	**include**: {Andrzej Zaucha, Andrzeja Zauchy}, {Zuzanna Leśniak, Zuzanny Leśniak}	other
86	Czym, podczas bożonarodzeniowego zwyczaju zapraszania dzikich zwierząt na Podhalu, wabiony był wilk?	**include**: {grochem, grochu}	religion & tradition
96	Czym różni się żurek od barszczu białego?	**include**: {żytni, żytnia}, {pszenny, pszenna}, {zakwas, mąka}	cuisine
100	Których spośród wymienionych poniżej produktów nie stosuje się jako zakąsek do wódki? śledzie, jogurt, korniszony, awokado, grzybki, galareta, lukrecja Wypisz tylko listę produktów, bez dodatkowego komentarza.	**include**: jogurt, awokado, lukrecja **exclude**: śledzie, korniszony, grzybki, galareta	cuisine

Description

Polish linguistic and cultural competency benchmark comprises of hand-crafted questions designed to evaluate LLM's factual knowledge on Polish culture, tradition and language. The level of difficulty of the questions varies, from those that would be answered by the majority of Poles to detailed questions focusing on region-specific culture or ethnic minorities. The questions have been phrased in such a way that it is deterministically possible to verify their correctness. Approximately half of them are various forms of closed-ended questions such as single-choice, multiple-choice, matching two sets of concepts, or filtering concepts from a list. The rest of the questions usually require answers consisting of a single sentence containing a specific fact or a set of facts. Open-ended questions allow some freedom for the model to generate an answer, but the response should refer to specific entities such as people, dates, numbers, places, certain concepts or phrases. In addition, each question is typically supplemented with instructions for the model, imposing a specific form of answer - short, precise, without additional comments or elaborate explanations. The dataset has been divided into six categories, with 100 questions in each. These categories include:



- **Culture & tradition** - The category contains questions about beliefs and religion, which are derived from both Christian and folk traditions, as well as Slavic mythology. It also contains questions about pop culture, including characters and events that have had a significant impact on Polish society. Finally, we also included questions about everyday life, covering Polish customs, cuisine, and clothing, among other topics.
- **Art & entertainment** - This category focuses on fields of art such as literature, painting, sculpture, theater, music, dance, or film. The questions cover works and figures related to Polish art. In addition, this category also features questions related to entertainment, including sports, popular music, television, radio, and people associated with show business.
- **Geography** - The category covers Polish geography and is divided into four subcategories. Two of them deal with structures and phenomena of natural (e.g., mountains, rivers, climate) and man-made origin (e.g., cities, tourist attractions, industrial plants, mines). The third subcategory deals with socio-political geography, which includes questions on population, social problems, national borders, and administrative units. The last group is biogeography, which covers the fauna and flora of Poland.
- **History** - The category covers Polish history from the time of Mieszko I to the present day. It is divided into subcategories corresponding to historical periods. It contains questions about important historical events and figures of significance to Polish politics, science, and

humanities.

- **Grammar** - The category deals with the rules and principles that govern the structure of sentences in the Polish language, as well as the rules of spelling (orthography). The questions address both theoretical foundations and practical applications of grammar. In addition to orthography, the questions cover such topics as morphology, parts of a sentence, parts of speech, phonetics, word formation, or rhetorical figures.
- **Vocabulary** - The category verifies LLMs' ability to understand the meaning of words, idioms, sayings and proverbs. The questions mainly focus on less frequently used words and phrases. In addition, slang expressions, regionalisms, dialects, colloquial language, and youth language were also included. Apart from modern Polish, several questions also deal with archaisms and old language.

Grading

Each question in the benchmark defines one or more conditions that must be met for an answer to be accepted. The scores are binary, a model can receive one or zero points for an answer. Partial points are not possible. This means that an answer is considered correct if and only if all the conditions defined in the question are satisfied. If at least one condition is not met, the model gets zero points for that question. The final score in the benchmark is calculated as a percentage of correct answers.

The process of verifying a single question starts with sending it to the model and obtaining the answer. We do not use the system prompt in the evaluation, and the question is encoded using a chat template specific to each model as a single message with the user role. In addition, to ensure deterministic responses, the generation temperature parameter is set to 0. In the second step, we normalize the response. Normalization involves removing all characters except letters and numbers, making all letters lowercase, and then lemmatizing the text, that is, reducing all words to their base forms. Such normalization is necessary for languages with rich morphology like Polish, because only then we are able to match different forms of the same word between the model's response and the question's conditions. After normalization, we check each of the conditions defined in the question. The model scores a point only if all conditions have been verified successfully. Our benchmark supports the following condition types:

- **Include** - Checks whether the words or phrases defined in the condition occurred in the model's response. In the simplest case, we can provide a list of comma-delimited expressions and the condition will be satisfied only if all these expressions are found in the text. In practice, however, there may be more than one way to formulate the correct answer, so it is allowed to define alternative expressions for each item. In such a case, only one of the provided expressions needs to be in the answer. In other words, the **include** condition can verify any logical formula in conjunctive normal form (CNF), in which individual clauses check the occurrence of a word or phrase in the model's response. Moreover, the condition can be parameterized. Instead of the default behavior of matching all defined expressions, we can specify the minimum (**include_min**) and maximum (**include_max**) number of expressions that should be included in the response. In special cases, we can also disable lemmatization (**lemmatize**) if we want to verify the occurrence of a specific, and not just any, form of a word.
- **Exclude** - This is the inverse of the previous condition. It checks that none of the given words or phrases occur in the answer. The condition is verified in a similar way to **include**. We can also use the **lemmatize** parameter to disable lemmatization when matching words.
- **Order** - Checks whether the words or phrases defined in the condition occur in the response in the expected order. Verification of the condition is almost identical to **include**. The difference is that the positions of the first occurrence of the expression, or any of the alternative expressions, are saved. The condition is considered satisfied if and only if the positions are in ascending order. The condition can be used to verify questions involving sorting or matching information. The **lemmatize** parameter can also be applied.
- **Regex** - This is the most complex condition, which checks whether a defined regular expression occurs in the answer. If no additional parameters are specified, the condition is considered to be satisfied if at least one string matches the regular expression in the text. However, it is possible to control the acceptability criteria through parameters. For example, we can define the minimum (**regex_min**) and maximum (**regex_max**) number of occurrences allowed. Furthermore, it is possible to force all occurrences to have the same number of characters (**regex_match_length**). Finally, we can introduce an additional dictionary criterion (**regex_match_word**). If the parameter is set to **true**, each matched string must be a valid word in Polish, which is verified using an external dictionary. Unlike the other conditions, **regex** is always verified on the unnormalized version of the response, so the **lemmatize** parameter is not applicable.