Journal Pre-proof

The following article:

A scientific evaluation of the use of limited versions of AI tools as support in identifying and defining simple non-English lithological terms

Urszula STĘPIEŃ, Aleksandra FRONCZAK, Wiktor WITKOWSKI and Daniel ZASZEWSKI

is accepted, peer reviewed article assigned to issue 4 of volume 69 that is not yet appropriately edited, but is citable using DOI:

https://doi.org/10.7306/gq.1830

This version will undergo additional copyediting, typesetting and review before it is published in its final form.

A scientific evaluation of the use of limited versions of AI tools as support in identifying and defining simple non-English lithological terms

Urszula STĘPIEŃ[1,*], Aleksandra FRONCZAK[2], Wiktor WITKOWSKI[2]and Daniel ZASZEWSKI[2]

[1] Polish Geological Institute – National Research Institute, Rakowiecka 4, 00-975 Warszawa, Poland, ORCID: 0000-0003-3568-0459
[2] University of Warsaw, Faculty of Geology, Żwirki I Wigury 93, 02–089 Warszawa, Poland; ORCID: 0000-0003-0830-8547
* Corresponding author, e-mail: uste@pgi.gov.pl

Received: September 24, 2025; accepted: December 10, 2025

Abstract: This study was prompted by the need to examine how well AI tools and large language models (LLMs) handle geological issues, particularly lithological issues, in languages other than English. The study aimed to evaluate the quality of responses in Polish generated by free versions of AI tools accessible to non-geologists with limited technological expertise. The survey, which was conducted between February and May 2025, involved people with a background in geology and students of geosciences, whose task was to evaluate each of the responses received. The lithology questions were the same for all respondents. The study involved using ChatGPT, Claude, DeepSeek AI, Google Gemini, Microsoft Copilot, Perplexity AI, and Qwen2.5. Respondents were most likely to use ChatGPT, Microsoft Copilot and Perplexity. The assessment covered the factual accuracy of the responses, the reliability of the sources referenced, and the comprehensibility of the responses received. The study revealed that not all AI tools can process the Polish language effectively, and a lack of relevant publications in Polish hinders the improvement of response quality. It was shown that more complete and complex queries that delve deeper into substantive knowledge enable higher quality and more satisfactory results. These results indicate the need to adapt algorithms to regional scientific terminology specifics, which could enhance the quality, reliability and usefulness of the content.

Key words: large language models, lithology, geological terminology, chatbots

Introduction

The development of tools based on large language models (LLM) and their growing popularity have inspired research at the intersection of geology and popular science. Given that geological information is encountered not only in scientific publications, on national park and educational trails, and popular science publications, but also, incidentally, in stores such as garden centres and interior design shops, we decided to assess what information concerning lithology an ordinary citizen can obtain by asking questions to artificial intelligence (AI) systems. Given the prevalence of English in the development of AI tools, and also the number of publications in reputable journals in this language, one may expect that the quality of results in English would be higher. Queries in AI chatbots formulated in national languages can return less satisfactory answers (Zhang et al., 2023; Jędrzejczak and Kochanek, 2025). These assumptions led to the design and implementation of the survey described in this article.

Our main objective was to assess the quality and reliability of responses generated by AI tools on geological issues of varying complexity, developed in languages other than English, when the questions are asked by people without any geological background. Information pertaining to lithology can be accessed on taking many different routes through various media, including information boards, leaflets and popular science books. The term ‘lithology’ can be used to describe a variety of geographical and cultural features. For instance, it can be employed to characterize historical objects such as marble statues and slate roof tiles. Additionally, it can be used to describe natural attractions, including waterfalls that cut through limestone, as well as artistic creations such as ceramic decorations made of clay.

Achieving this research objective entailed several assumptions. The subject matter of the questions was narrowed down to issues related to lithology, assuming that among the many possible geological issues, terms related to rocks seem to be the most popular and understandable. An attempt was made to put oneself in the shoes of a user who is not a geologist and has at most basic geological knowledge, which does not guarantee a critical assessment of the results obtained. Additionally, it was assumed that knowledge of how to use AI tools is basic, and this was reflected in the way questions were formulated for the chat. This study was also conducted to test how commonly available AI tools handle relatively simple questions and answers not formulated in English. According to data from August 2025, the total number of Polish speakers worldwide does not exceed 45 million (h*ttps://www.gov.pl/web/nauka/jezyk-polski-za-granica*). An additional goal was to check how AI tools deal with terms that are ambiguous and whose meaning is clarified by the context of the sentence and by additional attributes. Differences between answers to the same question, but written in a simple or complex way, were also compared. The study involved mainly students and geologists of various specializations, who used a questionnaire to make a preliminary assessment of the substantive quality of the responses and references received.

Application of large language models in geosciences

LLMs are becoming increasingly important in the natural sciences, including sciences covering geological issues, especially as concerns data processing, document analysis and knowledge extraction. Historical studies, such as the work of Campbell and Roelofs from the 1980s, already postulated the use of AI for intelligent spatial data management and decision support in the Earth sciences—which today, thanks to LLMs, is taking on a new dimension in the form of integrated assistants capable of analysing text, numerical data, satellite images and seismic models within a single platform (Campbell and Roelofs, 1984).

Contemporary AI chatbots based on LLMs are increasingly used in geological research and Earth sciences, offering support in data analysis, interpretation of results, and scientific communication. Models such as ChatGPT, Claude and Copilot excel at spatial and geoinformatics tasks. They perform well in tasks related to spatial skills, GIS theory, and code and programming function interpretation, but show weaknesses in cartography, code writing and spatial reasoning (Hochmair et al., 2024). Additionally, significant differences in accuracy have been observed between individual chatbots. Research by Cahyana et al. (2024) involving soil scientists from the National Research and Innovation Agency, universities, and members of the Indonesian Soil Science Society (HITI), showed that the overall level of trust in ChatGPT was 55%. Additionally, 80% of respondents believed that ChatGPT could only be used as a tool to support research in soil science and should not replace the involvement of specialists.

Analyses of the potential application of ChatGPT in hydrology and earth sciences have shown that, although ChatGPT can be a useful tool to support scientists in creating content and accelerating scientific progress, there are also some substantial concerns. These include potential errors, model limitations, and the risk of unethical or inappropriate use. In addition, there is also the problem of excluding non-English-speaking users who use language models to edit and correct their own work (Wee and Reimer, 2023). Therefore, it is recommended that the academic community adapt its regulations and policies to maximize the benefits of using such models while minimizing the associated risks (Foroumandi et al., 2023).

Microsoft Copilot, thanks to its integration with seismic data visualization and processing tools, has become a platform that enables complex geoscientific analyses to be conducted using natural language. As part of the project described by Altynova et al. (2024), a multimodal extension was developed that allows for the analysis of borehole data, seismic profiles in SEG-Y format, and reservoir simulation results in an integrated AI environment.

Gemini and Perplexity AI offer high linguistic quality and accuracy of responses in text analysis – the former stands out for its linguistic accessibility, while the latter excels at effective information retrieval and organization (Reyhan et al., 2024). In a study by Salih et al. (2024), which evaluated the competence of Gemini and Perplexity AI in tasks related to scientific publications, both models showed comparable results to ChatGPT, achieving high response accuracy in tasks related to literature review and research methodology. These properties may be useful in the analysis of geological literature and the automation of scientific reviews. The application of these language models in the Earth sciences goes beyond classic text generation—they are increasingly used as tools to aid data analysis, predict natural phenomena, and support scientific communication. A particularly interesting example is the study by Mousavi et al. (2024), which showed that Gemini 1.5 Pro can estimate the intensity of seismic shocks based on unstructured data from social media. This model showed consistency with measurement data, suggesting that it can draw conclusions concerning physical phenomena such as earthquakes despite the lack of access to classical geophysical data.

Research conducted by Khanifar (2025) showed that LLM-based chatbots such as Claude 3.5 Sonnet, GPT-4o, GPT-4o mini, Gemini 1.5 Pro, and Gemini 1.5 Flash demonstrated varying degrees of effectiveness in answering questions related to soil science. The results indicate that Claude 3.5 Sonnet and GPT-4o achieved similar, promising results, although ~35% of their responses contained errors. This suggests that soil science questions are particularly challenging for chatbots. Furthermore, the performance of the GPT-4o model did not depend significantly on the input language, which means that language is not a limitation in the application of ChatGPT to this field.

Although DeepSeek AI and Qwen2.5 have not yet been analysed in detail in the context of geosciences, their open architecture and development in open-source environments suggest that they can be easily adapted to the geological domain, especially when integrated with local databases and GIS tools. These findings show that LLM chatbots, when properly tuned, can serve as intelligent assistants in Earth sciences, supporting data analysis and scientific communication, and also geological education.

At the same time, specialized LLM models dedicated to users involved in geosciences are emerging. Examples include GeoGalactica, which specializes in the analysis of geological terminology (Lin et al., 2024), the K2 model, which focuses on climate and Earth sciences (Deng et al., 2023; Hadid et al., 2024), and GeologyOracle, a GPT-3.5-based chatbot designed for geological education, GeoChat, presented by Kuckreja et al. (2024), which processes images in addition to natural language.

However, it should be noted that despite technological advances, there are also limitations and risks associated with the use of LLM in geosciences. Hawkins (2024) described the case of the Chinese GeoGPT model, which, despite its technological sophistication, censored terms and restricted freedom of scientific exploration. This case highlights the importance of algorithmic transparency and independence in the context of using artificial intelligence in geological research.

However, all of the solutions described above, developed for and with the participation of geologists, are tools developed for the needs of the industry. They are generally little known, and, likely, they will not be used to the same level as popular AI tools implemented in most browsers at a mass scale. Therefore, our study reached out towards those popular solutions that are used by the vast majority of the population.

Due to the growing popularity of AI tools, the European Commission at the legislative level (PE/24/2024/REV/1 2024) emphasizes that generating synthetic content using AI creates the risk of providing information that misleads the user. This requires system providers to take various measures to minimize these risks. The results of our study can be used by AI users and providers to identify possible errors and problems when using this technique.

Materials and methods

We aimed to assess the quality and reliability of AI-generated responses to basic questions about lithology formulated in Polish, of the kind asked by users who are not geologists and have only basic knowledge of how to

use AI tools, as well as those who can use AI tools in the same way as they previously used internet search engines. Issues related to rock terminology used, for example, in construction and interior design were not assessed here, as the terms used in these fields are, in most cases, incorrect from a geological point of view. An example is the commercial term “marble”, which, from a geological point of view, can be marble, limestone, or dolomite that has been ground and polished, making it attractive from a decorative point of view. When analysing the responses received, attention was paid to whether the sources were appropriately selected and based on geological knowledge.

Another risk that may arise among non-geologists using the AI chatbot is the uncritical acceptance of the answers generated by the tool. The study involved students, doctoral students, and researchers associated with the field of geology. The spectrum of chats we examined required a large research sample, ensured by a wide range of respondents. At the same time, focusing on people associated with geology gave us the opportunity to more reliably assess the content of the responses obtained.The study involved employees of the Polish Geological Institute – National Research Institute (PIG-PIB) and lecturers and students from the Faculty of Geology at the University of Warsaw (WG UW).

The study was divided into two stages. The first stage was a survey. Microsoft Forms (MSForms) was used for this purpose. The survey was anonymous, but it was designed so that the first part determined the profile of the respondent. Then, the respondents chose an AI tool, which they used to complete the main part of the survey. A total of 202 completed questionnaires were collected for further analysis.

The second part of the survey contained seven questions that had to be copied and pasted into the chat window. For the results to be comparable, the set of questions was formulated as one would expect from a person who is not a geologist and is not proficient in the use of AI tools, which should be understood as, among other things, a lack of knowledge and experience in the field of mathematical algorithms, and a lack of ability to correctly formulate queries in AI-based tools. It should be emphasized that knowledge of lithology is crucial in order to construct prompts in such a way as to increase the accuracy of the response received. Therefore, even a proficient knowledge of AI tools would not guarantee a substantively correct result. In addition, one of the questions was designed to test how AI copes with the ambiguity of lithological terms that the average citizen may encounter. It was assumed that there was a risk that it might search for definitions of terms whose full meaning is derived from context, so it was decided to test how AI would deal with this issue when no additional explanation was available, because a potential interested party would not have sufficient substantive knowledge. It has often been stated that, during the development of dictionaries for international projects and initiatives (OJ L 108, 2007; *https://eurogeosurveys.org/projects/onegeology-europe/*; Asch, 2010; Asch et al., 2025; Stępień et al., 2025), that ambiguity can cause problems. It commonly happens that one English term corresponds to many more precise terms in national languages and vice versa. All AI responses to queries and sources were pasted into the survey by respondents to enable later verification and evaluation.

AI Tools

For security reasons, it is recommended to use AI tools that do not require logging in. Some institutions, in accordance with their internal security policies, do not recommend, or even prohibit, logging into external systems. Caution is advised, as it can be argued that the data sent to the chatbot for analysis is a kind of payment for the free use of AI tools. This means that this data could be further used by other users (*https://www.gov.pl/web/baza-wiedzy/pulapki-zwiazane-z-wykorzystywaniem-sztucznej-inteligencji--jak-unikac-zagrozen*). The restrictions introduced are also intended to minimize the risk of phishing and data leaks. Public entities are particularly vulnerable to data theft.

The list of models used for testing included seven LLM-based tools (Tab. 1). The selection criterion was the possibility of using the tool free of charge. Respondents were informed about the conditions for free access to the AI tool, i.e., the necessity or lack of obligation to log in. This information is important for network security reasons. Some users do not log in of their own accord, while others, using work devices, cannot log in to external platforms, including AI tools, for security reasons. Therefore, respondents were asked about the frequency and scope of their use of AI tools to date. The rest have their own dedicated engines. The most important features of the tools proposed to respondents for use in the study are discussed below.

ChatGPT, developed by OpenAI, stands out for its wide availability without the need to log in and the ability to use the internet browsing feature in selected versions. This model is widely used to generate definitions, classifications, and explanations in many fields, and its intuitive interface makes it easy to use for users of all skill levels. ChatGPT is good at synthesizing information from various sources, allowing it to provide up-to-date and comprehensive answers. (*https://openai.com/index/chatgpt/*)

Claude, developed by Anthropic, is a model focused on security and handling large data sets. It stands out for its ability to analyse very long documents, summarize them, classify them, and draw conclusions. Claude requires logging in, which allows for personalization and access to conversation history. This model is particularly useful for tasks requiring in-depth text analysis and working with extensive source materials. (*https://www.anthropic.com/claude*).

DeepSeek AI uses Transformer architecture with Mixture of Experts (MoE) elements, which enables it to effectively analyse long and complex texts. This model is particularly effective in processing large documents, making it suitable for tasks requiring accurate analysis and synthesis of information. The free version requires logging in, which allows for personalization and access to conversation history. DeepSeek AI stands out for its speed and ability to analyse text at multiple levels of detail (*https://www.deepseek.com/en*)

Google Gemini is a multimodal model that can analyse both text and images, making it unique among language models. Gemini uses tools such as a web search engine and file analysis to search and synthesize

information from various data formats. The model was developed by Google DeepMind and is integrated with Google services, allowing for seamless collaboration with other Google applications (*https://gemini.google.com*).

Microsoft Copilot is an AI assistant integrated with Microsoft 365 applications such as Word, Excel and PowerPoint, allowing it to directly analyse, generate, and visualize data within these tools. Copilot uses OpenAI technology (including GPT-4o), which gives it access to the latest language models. This model is particularly useful for tasks involving documents, spreadsheets and presentations, and its integration with the Microsoft ecosystem enables seamless collaboration with other Microsoft services. Compared to the Pro version, the free version has limited functionality (*https://copilot.microsoft.com*).

Perplexity AI is an AI-based search engine that automatically cites sources in its generated responses, which sets it apart from other language models. This model is particularly useful in scientific research, as it allows you to quickly obtain and verify information from many reliable sources, such as scientific articles, news and forums. Perplexity AI is available without logging in and allows for easy comparison of information (*https://www.perplexity.ai*).

Qwen2.5, developed by Alibaba Cloud, is a model that supports advanced data classification and tasks related to coding and mathematical data analysis. The model is available after logging in and is distinguished by its high performance and ability to handle long contexts. Qwen2.5 is particularly useful for document analysis and tasks requiring the processing of large amounts of text and numerical data (*https://Qwen.ai*).

Due to the pace and intensity of change and development of AI tools, it should be noted that the research was conducted between February 3 and May 31, 2025.

Surveys as a research method

The primary research tool was a survey developed in MS Forms, comprising several parts: an introduction that defined the respondent's profile and the tool used to complete the main part of the survey, a set of substantive questions, and a subjective assessment.

The introduction to the survey served to determine the respondent's profile; the first questions concerned education. There were six possible answers to choose from: first-cycle student, bachelor's/engineer's degree, second-cycle student, master's/master's engineer's degree, doctoral student, doctor and above. Although the survey was addressed to employees of the Faculty of Geology at the University of Warsaw and the Polish Geological Institute – National Research Institute, questions were also asked about the field of study to ensure that, as intended, the survey was completed mainly by people with an education that would allow them to critically evaluate the answers generated by AI. This was a multiple-choice question, and in addition to the answers suggested (geology, geography, geophysics, geoinformation, computer science), it was also possible to provide other answers not included on the list. Another question defining the profile of the respondent was a subjective assessment of their knowledge of artificial intelligence (no knowledge, basic, intermediate, advanced, expert). Respondents were also asked about the frequency of their use of AI models (daily, at least once a week, once a month, less often, never) and the purposes for which they use them (professional, educational, entertainment, other).

The next part of the survey asked respondents to select the AI tool they used to develop their answers to the study's main questions. Seven globally known free AI tools were suggested for completing the questionnaire: ChatGPT, DeepSeek AI, Google Gemini, Perplexity AI, Claude, Qwen2.5 and Microsoft Copilot. Each tool was accompanied by a direct link and information on whether logging in was required. The Polish AI tools were purposely excluded from the research. BIELIK was only available to logged-in users, and PLLuM was launched after the survey had already begun. After selecting a tool, respondents were asked to indicate whether they had previously used it to search for geological content: 'Yes, several times a week'; 'Yes, several times a month'; 'Yes, but I do it occasionally, several times a year'; or 'Never'. Only after this introduction did the main part of the study begin.

As previously mentioned, the main part of the survey comprised seven questions which had to be copied and pasted into the chosen chat window. These questions varied greatly in complexity. The order in which the questions were asked was also important. Question 6 was a more specific version of question 5. Some questions requested the sources of the definitions provided, while others deliberately omitted this information to see if it would be returned by the chatbot. Below is a set of questions in their original wording, alongside an English translation and commentary to help you understand the linguistic nuances that played a special role in the study (Tab. 2).

The next step after receiving the response was to paste it into the survey window and enter information about possible sources provided by AI.

The main part of the survey began with a question asking respondents to define lithology, the field to which the subsequent questions related. Respondents were not asked to cite sources; the question was simple and did not define the substantive needs or knowledge of the respondent.

Bearing in mind that for AI tools, the order of questions in a single session is important; if the user used the same session for the purposes of the study, one question appeared twice, in both simple and complex form (questions 5 and 6). First, a simple question was asked, and then a differently worded question, with more detail. The purpose of this question was to check whether the model contradicted itself, and how refining the prompt would affect the quality of the response.

The survey ended with a simple question about the definition of a rock with the ambiguous Polish name “łupek,” which corresponds to several English terms: slate, schist, shale. The question was intended to test the problem of ambiguity of terms, translation, and precision of answers. In Polish, this term is used to describe sedimentary and metamorphic rocks, and its meaning is derived from the context or clarified by an adjective referring to a specific group of rocks. The study was designed to check whether the model would provide all possible options without additional guidance.

Another issue considered in the survey was the level of complexity of the questions. Questions 1, 3, 4, and 7 were based on basic knowledge, while questions 2, 5, and 6 required reference to scientific sources. The entire question session began and ended with simpler questions that did not require the respondent to be so cautious in

assessing the result obtained. In addition, three questions included a request to provide sources for the answer generated (Table 2, questions 1, 3, and 6).

Each question in the survey contained a subjective assessment of the answers provided by AI. Respondents answered the following set of questions:

1. Does the tool provide sources for the information provided?
2. What are the types of sources provided in the answer? There was a set of source types to choose from (scientific, popular science, academic textbooks, school textbooks, websites run by scientific institutions, and websites), and the additional option to add a missing source type, as well as the statement that “the source seems unreliable”.
3. Do you find the information you received understandable?
4. Are you satisfied with AI's response?
5. Why are you not satisfied with AI's response?
6. Overall satisfaction rating for working with the selected AI tool?
7. As noted at the beginning, the respondents had sufficient geological knowledge to assess the accuracy of their answers, so their responses to the above set of questions provided a preliminary assessment of AI in the survey itself. This is a very important part of the study, because in the case of simple information provided by AI that does not require qualification, it is easy to recognize its inaccuracy (Fig. 1A). In advanced matters, it is relatively easy to lose vigilance and, due to a lack of sufficient knowledge and critical thinking, accept incomplete or not entirely reliable information as fact, or even unknowingly become its promoter (Fig. 1B).

.

The next stage of the assessment was based on the AI responses gathered from the questionnaire. The survey also enabled us to identify the sources of the responses, allowing us to compare and verify them.

Analysis of sources used by chatbots

An automated analytical pipeline, implemented using dedicated Python scripts, was used to analyse the data obtained from the surveys. The research process consisted of four main stages: data extraction and cleaning, automatic source classification, result aggregation, and visualization.

In the first stage, data from the MS Excel survey file was pre-processed. Internet domain names were extracted from the URLs contained therein. For example, for the address *https://pl.m.wikipedia.org/wiki/Torf*, the domain pl.m.wikipedia.org was obtained. Records containing links to PDF files were deliberately excluded from further analysis, as this type of resource was not subject to automated verification under the adopted methodology.

The cleaned domains were automatically categorized. Two complementary approaches were used in this process:

1. **Classification using an AI model:** The main classification tool was the Perplexity AI PRO model. This process, supervised by humans, involved assigning each domain to one of four predefined categories: scientific publications, popular science sources, scientific institution websites, and other websites.
2. **Heuristic classification:** In addition, an algorithm was used to identify scientific publications and academic textbooks. It operated based on rules that took into account the presence of keywords and dates in the records analysed. The results obtained using this method were subjected to a general review by us, but due to the number of records (~1,500), they were not subjected to detailed verification in each case.

All data classified was aggregated in a summary table. For each record, the type of domain, the survey question number from which it originated, and the AI model that indicated the source data, were specified.

The effectiveness of the automatic procedure was estimated at ~90%. This indicator determines the percentage of cells in which the system recognized at least one source (link or scientific reference) among all non-empty records. However, this is an estimate. The group of unaccounted cells included, among others, cells containing only references to PDF files that were not recorded by the model, which means that the actual detection efficiency could have been higher. In addition, cells were considered correctly included if some of the sources they contained were recognized, even if not all of them were recorded. Despite these limitations, the mechanism adopted was considered the most appropriate for the characteristics of the material collected. In the final stage, the aggregated results were also visualized.

Ranking evaluation of selected chatbots

The final stage of the analysis assessed the effectiveness of four selected versions of artificial intelligence tools (ChatGPT, Perplexity AI, Microsoft Copilot, Qwen2.5) as support in identifying and defining simple, non-English lithological terms. The study took into account three key aspects of the tools' performance: the presence of references to information sources in the responses generated, the comprehensibility of the messages conveyed, and the level of user satisfaction with the responses received. Based on this data, a qualitative ranking of the individual tools was developed.

The criteria were assigned point values (Tab. 3). Given the uneven number of data sets for the selected chats, the response rates within a given criterion were calculated for each of the seven questions. These shares were then multiplied by the point value. This allowed for the quality of responses for each question to be rated on a five-point scale. The average rating for all questions gave the overall score for the chatbots analysed.

Results and disscusion

Survey results: characteristics of the respondents

During the period between 3 February and 31 May 2025, a total of 202 survey responses were collected. Thirty-three per cent of respondents held a master's degree. The same percentage (33%) were respondents with doctoral degrees and higher. The remaining 34% of the sample were first- and second-cycle students, as well as individuals in possession of a bachelor's or engineering degree (Fig. 2A). This information is crucial for the preliminary substantive assessment of the responses received. The objective of the study was to verify the content of the responses, a task which is not always feasible for individuals lacking a background in geology.

The survey revealed that almost 70% of respondents indicated that they possessed either fundamental knowledge of, or lacked any prior experience in, utilising AI tools (Fig. 2B). A mere 10% of respondents claimed to possess advanced or expert knowledge. Figure 2C illustrates that 23% of respondents utilise AI tools on a daily basis, while 27% do not employ them. Furthermore, 66% of respondents reported utilising AI tools for professional and educational purposes (Fig. 2D). An additional query posed to respondents pertained to the utilisation of the tool designated in the survey, enquiring whether the respondents had previously employed this tool in the pursuit of geological content. As illustrated in Figure 2E, a significant proportion of respondents, amounting to ~50%, selected an AI tool with which they had no prior experience.

ChatGPT emerged as the predominant platform of choice for conducting the survey, garnering the selection of 88 respondents. Microsoft Copilot and Perplexity AI also received significant attention, with 41 and 16 responses, respectively. The least popular of the subjects was Claude and the Chinese DeepSeek (Fig. 2F). In view of the marked discrepancy in the number of responses for each tool, the four with the highest number were selected for further analysis and evaluation. The following examples are provided for illustration: ChatGPT, Perplexity AI, Microsoft Copilot, and Qwen2.5.

The analysis of the available data indicates a clear correlation between the frequency of AI tool utilisation and satisfaction levels. Individuals who do not utilise AI tend to exhibit lower, albeit more diversified, levels of satisfaction (Fig. 3). It is evident that among users who employ AI on a sporadic basis, there is a substantial increase in satisfaction levels, with the 3–5 range demonstrating a predominant presence. The highest levels of satisfaction are reported by individuals who frequently utilise AI tools, with the majority of ratings falling within the 4–5 range, indicating a narrow distribution.

The level of education – irrespective of whether the participants are students, doctoral students, or individuals in possession of a PhD or higher – does not differentiate satisfaction as strongly as the frequency of AI use. The same pattern is observed in each educational group: the more frequent the use of tools, the higher the satisfaction.

Survey results: responses to questions

After all the responses had been collected, the analysis was carried out in several stages. First, all the responses were compiled regardless of the AI tool used, and then they were divided into seven categories. The level of satisfaction with the answers received was analysed. After obtaining information on whether the AI tool provided sources in addition to answers to the query, the respondents were asked about the type of sources provided. Scientific publications and academic textbooks were considered scientific sources, while school textbooks and websites run by scientific institutions were considered popular science sources. The 'other sources' group included websites and other sources whose credibility was difficult to verify quickly. A total of 1414 responses were collected, almost 40% of which did not contain any references to sources of information.

The first question of the study aimed to define the term 'lithology' and provide references. Despite the clear request to provide sources for the definitions, 5% of the responses did not include them (Fig. 4A). No pattern was found here, as the sources were not provided by Chat GPT or Microsoft Copilot, which gave complete answers to most questions. The remaining 94.6% of responses included citations, most of which referred to popular science and other non-scientific sources (Fig. 4B). While 87% of respondents found the answer understandable (Fig. 4C), only 64% considered it to meet their requirements (Fig. 4D).

The comments shared by dissatisfied respondents are important for evaluating the answers. They added that the answers generated by AI did not always capture the full meaning of the term ”lithology”, and that sometimes the answers were vague or even incorrect. For instance, one respondent defined lithology as a branch of geology. Respondents emphasised that the term “lithology” is often confused with ”petrology”, or is only associated with sedimentary rocks.

The second question was more complex and required more specialised knowledge. Respondents were asked to identify the similarities and differences between Folk's and Dunham's carbonate classification schemes. Although this question was more difficult, respondents were not asked to provide sources. However, 37.1% of responses included them (Fig. 5A). Scientific publications constituted over 40% of the sources (Fig. 5B), which can be linked to the fact that the question concerned classification, a topic that is less often the focus of popular science studies. 77% of respondents found the answer understandable (Fig. 5C), and as with the first question, 65% were satisfied with it (Fig. 5D).

Among other things, dissatisfied respondents complained about answers that were chaotically formulated, overly brief, or described differences in a way that was sometimes ambiguous, making them difficult to understand. Other issues included the perceived use of machine translation and the use of overly specialised terms without explanation. Due to doubts about the accuracy of the answers, respondents sought knowledge from verified, reliable sources.

The third question related to the definitions of "gaize" and “opoka”, as referenced in the scientific literature. These popular Polish terms do not have exact equivalents in English. Although 95% of responses included sources (Fig. 6A), only 47% met the criterion of providing references to scientific literature (Fig. 6B), despite this being requested. 84% of respondents considered the text of the answers to be understandable (Fig. 6C), and 61% of the answers were deemed satisfactory (Fig. 6D).

Respondents who were dissatisfied with the answer found it to be imprecise and containing factual and linguistic errors. They also complained that AI was confusing concepts and using inappropriate terms. There were also complaints about the lack of a clear explanation of the differences between "gaize" and “opoka”. The definitions provided were often inaccurate (e.g., describing ‘geza’ as a carbonate rock). Respondents emphasised that the cited sources were unscientific or from popular science websites (mainly Wikipedia), which they said reduced the credibility of the information. Some answers were also irrelevant, referring to another issue (e.g., a biblical city).

The next question was related to the previous one. This time, respondents were asked to describe the differences between “gaize” and “opoka”, rather than providing sources. Consequently, fewer than 40% of responses included sources (Fig. 7A), 44% of which referred to scientific literature (Fig. 7B). Nevertheless, the answers were understandable (Fig. 7C), and satisfaction levels were similar to those for the previous question, at 67% (Fig. 7D).

Several respondents who were not satisfied with the responses perceived the answers to be overly generalised and lacking in precision, and even potentially fallacious. As with the previous question, there were complaints about factual errors and the incorrect definition of gypsum and limestone. The significant differences between these rocks were also not highlighted. Attention was drawn to the language used, which was inaccurate in places and contained words or phrases in a foreign language. There were no references to reliable sources or scientific literature, which made it difficult to verify the information. Important features were omitted from the description, such as the role of silica sponges and the physical characteristics of rocks. Those dissatisfied with the answer pointed to disorganised information, unnecessary and meaningless descriptions, and a focus on industrial applications rather than geological features. Comments were also made about the style of the description provided.

The fifth question of the survey requested the name of a carbonate rock with the following characteristics: 70% snail shell fragments with a diameter of 1–2 mm and 30% carbonate silt. Just over 30% of respondents provided sources (Fig. 8A), of which only one-third were scientific (Fig. 8B). 76% of respondents found the answers easy to understand (Fig. 8C), but fewer than half of the answers were considered satisfactory (Fig. 8D).

A significant problem was the lack of references to sources and specialist literature, particularly geological classifications such as those of Folk and Dunham. According to some respondents, this rendered the information provided unverifiable, and reduced the credibility of the entire response. In addition, respondents reported factual errors and imprecise definitions of rocks and their classifications, as well as incorrect, non-existent, or inaccurate names. Respondents also complained about the lack of indication of the main types of rock, instead being given only their varieties.

Some respondents indicated that they were dissatisfied with the answers because they did not address all aspects of the question. For instance, key details about shell diameters were omitted, resulting in a definition that they deemed incomplete and superficial. There were also terminological errors and misinterpretations of rock classifications, meaning the answers did not meet the required level of scientific accuracy. Some answers were chaotic and inconsistent. Reservations were also raised about the linguistic correctness of the answers.

The sixth question was similar to the fifth. However, it was formulated in much greater detail. The AI was asked to adopt the perspective of a geologist and use precise terminology and classification. It was emphasised that the answer should be based on scientific articles, which were to be translated from English into Polish if necessary. Rocks that did not fit this specification were to be excluded. Almost 75% of respondents to this question included sources (Fig. 9A), with nearly half of these being scientific (Fig. 9B). Respondents found these responses more understandable than those in the previous task (Fig. 9C). Satisfaction levels also increased (Fig. 9D). The proportion of respondents who were dissatisfied with the response was halved.

Despite the more precise description of the issue, some respondents were dissatisfied with the answer. They complained about errors and shortcomings in the description and classification of rocks, particularly shelly limestones. The text contained errors such as confusing calcium with limestone, the incorrect use of geological terms, and the mixing of Polish and English names, which caused linguistic confusion and made the text difficult to understand. The definitions were incomplete and inconsistent. Despite requests for scientific sources, the answers were based on non-scientific sources. According to the dissatisfied respondents, this undermined the credibility of the information provided. There was also incorrect or imprecise terminology here, and the classifications given in the answers were sometimes contradictory or inconsistent with accepted systems such as the Folk or Dunham classifications. Additionally, the AI either did not provide specific rock names or did so in an unclear manner, making it difficult to verify the information. While there were many positives, some of the responses were considered insufficient from a scientific perspective, requiring significant refinement in terms of both reliable sources and linguistic accuracy.

The final question of the survey addressed the ambiguity of the lithological term 'łupek'. Depending on the context and adjectives used, the Polish term 'łupek' can be translated as 'shale', 'schist' or 'slate'. It was deliberately phrased without any context, as the study assumed that respondents would lack the geological knowledge required to understand it. Of those asked this question, 70% provided sources (Fig. 10A), 41% of which were scientific publications (Fig. 10B). While 90% of the responses were clearly described (see Fig. 10C), only 56% resolved the ambiguity of the term (Fig. 10D).

The respondents emphasised that some of the answers regarding 'łupek' were inaccurate in terms of both facts and language. The definition confused the characteristics of sedimentary shale and metamorphic schist, failing to acknowledge their significant differences. Important information about the origin and diversity of these rocks was also omitted. The respondents also received definitions that were too superficial and incomplete. Often, these definitions focused only on sedimentary shale, completely omitting metamorphic schist and other types of rock that are also referred to by this term. Some answers were unclear or contained incorrect terms and were disorganised, with repetitions and sections taken out of context that could be misleading. The lack of precise nomenclature and references to reliable scientific sources was also considered an error. According to the respondents, these shortcomings made it very difficult to verify the information received. Additionally, some responses contained translation errors, mixing Polish with English or Chinese, which reduced the clarity and credibility of the responses further. This gave the impression of a popular science description that did not meet the requirements of a reliable, complete geological definition.

As expected, including a request for scientific sources in the question to AI tools (questions 1, 3, and 6) noticeably increases the likelihood of receiving a satisfactory answer. However, such a request does not guarantee that a list of reliable scientific sources will be provided. Answers to questions about ambiguous terms, such as ”łupek”, tend to be less satisfactory. Failing to clarify an ambiguous term leaves the AI unable to return all possible options, but instead indicates only one answer out of many.

Automatic verification of response sources.

The automated analysis of the sources utilised by chatbots to formulate responses to inquiries within the domain of geology facilitated a quantitative and qualitative comparison of four AI models: ChatGPT, Perplexity AI, Microsoft Copilot, and Qwen2.5. The study covered seven questions and classified a total of 223 unique web domains identified as sources of answers.

Overall, popular science sources accounted for the largest proportion of responses generated by Perplexity AI (60.3%) and Microsoft Copilot (33.5%) (see Fig. 11A). ChatGPT used scientific sources more often than other models (18.3%), but even in this case, popular science sources accounted for a large percentage (29.6%). On the other hand, Qwen2.5 was distinctive due to its significant proportion of scientific sources (38.7%) and minimal utilisation of other classifications. Records containing no or unrecognised sources were noted in all cases, most frequently for Copilot (31.1%) and ChatGPT (30.3%).

The distribution analysis of sources for individual questions revealed considerable variability. For instance, all models indicated predominantly popular science sources for question 1. However, for more specialised questions (e.g., question 6), Qwen2.5 relied almost exclusively on scientific sources (81.2%), whereas other models maintained greater diversity. Questions 2, 4 and 5 were the toughest: over 50% of the answers to these questions in the case of ChatGPT and Copilot did not contain any sources, which suggests that the models have limitations when it comes to topics that require a deeper scientific knowledge.

The survey results also allowed us to analyse the relationship between the indicated source type and the subjective assessment of answer quality (see Fig. 11B). The highest percentage of satisfactory responses was based on scientific sources (68%), though slightly lower values were obtained for popular science sources (61%) and other sources (62%). Interestingly, 58% of respondents rated answers without sources as satisfactory, suggesting that the presence of a source does not always have a decisive impact on the perceived value of an answer.

The percentage of responses containing unrecognised sources ranged from 1.4% for Perplexity AI to 19.6% for Microsoft Copilot. The significant proportion of responses that did not cite any sources, particularly in the case of Copilot (31.1%) and Qwen2.5 (33.9%), may indicate differences in how sources are reported between models, or suggest that some systems have a limited ability to justify transparently the content they generate.

The results indicate differences in the strategies for selecting knowledge sources between the models analysed, with some models favouring certain knowledge sources over others. Perplexity AI prefers popular science sources, whereas Qwen2.5 uses scientific literature. Copilot and ChatGPT are more variable and often do not indicate any sources. There is a noticeable trend in the correlation between the type of source and user satisfaction: answers based on scientific sources are most often rated as valuable. These findings lay the groundwork for further research into the transparency and quality of the sources employed by generative AI systems in the context of specialist knowledge.

Evaluation of the tools examined

Analysis of the respondents' answers revealed clear differences in the extent to which the AI tools provided information about the sources used to generate the answers (see Appendix 1). Perplexity AI was the most consistent in this respect, with 95% of respondents saying they had received clear references to source materials. There was almost complete agreement in the responses, indicating the presence of sources, particularly in questions 1, 3, 5 and 6. This is a significant advantage of Perplexity AI over its competitors.

In the case of ChatGPT, there was significantly greater variation, with an average of 47% of responses indicating the presence of sources. References appeared most frequently in questions 1 and 3 (accounting for ~94% of references), but in questions 2, 4 and 5, this percentage dropped significantly (falling <20%), suggesting inconsistency in the generation of references. This was particularly evident in the specialised question, where ChatGPT was asked to provide sources for published information. Here, ChatGPT recorded the lowest effectiveness of all the tools tested, at 62.5%. Microsoft Copilot showed a moderate level of source attribution, averaging 60%, with clear variations between questions (<23% in question 5 and over 90% in question 1). Although the Qwen2.5

model was highly effective in question 1 (100% of references to sources), it was significantly less reliable in the other questions, with the percentage of references falling <20% (e.g., question 4).

The assessment of response comprehensibility (see Appendix 2), based on respondents' opinions, confirmed Perplexity AI's superiority. It achieved the highest average rate of comprehensible responses — 88%. Notably, this tool also stood out with an exceptionally low percentage of incomprehensible responses — less than 3% — demonstrating high precision in formulating Polish messages and providing satisfactory information. ChatGPT and Microsoft Copilot achieved similar rates of comprehensibility, at 83 and 84% respectively, with a moderate proportion of responses rated as partially comprehensible. Notably, despite its high level of content transparency, Microsoft Copilot generated slightly more responses that were partially or completely incomprehensible (up to about 5%).

The lowest level of clarity of responses was recorded for Qwen2.5, which had a significantly higher percentage of responses rated as partially or completely incomprehensible — a total of 39%, with only 61% of responses fully comprehensible. This difference was particularly evident in the more complex questions (2–6), where users found the content difficult to understand, suggesting that the model has limitations in its ability to interpret and synthesise specialist knowledge. The best comprehensibility results, both individually and on average, were recorded for questions 1 and 7, probably due to their simpler or more universal subject matter.

Analysis of user satisfaction revealed interesting differences in relation to the clarity rating (see Appendix 3). ChatGPT achieved the best result in this category with an average of 67%, indicating a relatively high level of user satisfaction despite a lower level of source referencing. Satisfaction with Perplexity AI was slightly lower at 63%. Microsoft Copilot achieved a satisfaction level of 45%, indicating a discrepancy between the perceived readability and the actual satisfaction with the quality of the information received. This was probably due to the incomplete nature of the responses or their insufficient relevance to user queries. The lowest satisfaction level was recorded for Qwen2.5 (48%), which confirms the need to further optimise this model, particularly concerning the usability and relevance of the generated content.

User satisfaction levels dropped significantly for questions 4 and 5, which were more detailed. The number of partially or completely unsatisfactory ratings increased for these questions, which may indicate that the tools have a limited ability to provide satisfactory answers to specialised or multifaceted questions.

The quality ranking of individual tools was prepared on the basis of an integrated assessment. This assessment took into account the presence of references to sources, the comprehensibility of responses, and user satisfaction. The results are illustrated in Figure 12. The scores reflected the overall quality of the information provided and the practical usefulness of the tools in the scientific environment.

It is evident that Perplexity AI demonstrated a marked distinction from the other models, attaining the highest mean score of 4.37 points (on a scale of 0 to 5), thereby surpassing ChatGPT (3.82 points) and Microsoft Copilot (3.80 points) by approximately 0.5 points (Fig. 11). The lowest score was recorded for Qwen2.5 (3.32 points).

The analysis of individual questions demonstrated that Perplexity AI's advantage was particularly evident in questions 2, 4 and 7, where the technology scored highest (above 4.3 points), thereby highlighting its superiority in terms of both completeness and precision of the responses generated. The other tools demonstrated greater variability in their results, particularly Qwen2.5, which scored significantly below average on some questions.

Summary

The research described in this article was conducted in Polish during the first half of 2025 (3 February–31 May 2025). The analysis was based on the results of a survey conducted among 202 respondents, including expert geologists, academic staff, and students of geosciences. The participants analysed their answers to seven questions of varying degrees of complexity, taking into account the presence and type of sources, completeness and factual accuracy, comprehensibility of the message, and overall satisfaction with the information received.

The findings demonstrated substantial disparities among the tools examined. It is a matter of concern that nearly forty per cent of all responses were devoid of any bibliographical references. Perplexity AI was distinguished by its uniformity in sourcing information, with a citation rate of 95% across responses. However, it should be noted that the majority of these citations were from popular science materials. Qwen2.5 obtained the highest percentage of citations to scientific literature, which had a favourable effect on the substantive evaluation of the content. Conversely, ChatGPT and Microsoft Copilot frequently offer responses without providing citations, thereby constraining the capacity to ascertain the veracity of the information imparted.

The study confirmed that the precise formulation of queries, taking into account requirements regarding the type of sources and expert perspective, significantly improves the substantive quality of the responses obtained. This phenomenon was particularly evident in the context of questions concerning linguistically ambiguous terms, where the models' responses often covered only one of the possible definitions, omitting others that were relevant in the context of geological sciences.

Respondents (mainly ChatGPT users) offered commentary on their experiences, offering a multifaceted perspective on their utilisation of AI-powered tools. The comments encompass a range of perspectives, including both enthusiastic and critical voices, as well as reflections that underscore the necessity for circumspect utilisation of the innovative technology.

Numerous users have attested to the high practical usefulness of AI in their daily work, especially in technical tasks such as code creation, office suite tool operation, and assistance with writing and translating texts. When employed with meticulously formulated queries, this tool has the potential to markedly enhance workflow, thereby reducing the time expended on specific tasks. The implementation of artificial intelligence has permeated various facets of society, including the domains of education and entertainment.

It was widely noted that the quality of the responses generated was contingent on the precision and complexity of the question, a factor that was anticipated during the development of the survey tasks. It was

hypothesised that the quality of the question would have a direct impact on the probability of receiving an accurate and substantively correct answer.

A recurrent levelled criticism in this study pertained to the reliability of its sources. AI systems have been observed to exhibit a lack of consistent referencing of sources for information, with this behaviour manifesting either infrequently or only in response to explicit instruction. This has given rise to concerns regarding the reliability of the content, particularly in professional and scientific contexts. It was also noted that imprecise or even incorrect definitions were in evidence, as well as the use of linguistic embellishments that may imitate scientific style but lead to an inaccurate factual description. As demonstrated by a survey in which the questions were deliberately formulated to appear layman-like, erroneous descriptions could not be avoided, despite understandable answers. Such descriptions can only be critically verified and evaluated by an individual with substantive knowledge.

Furthermore, respondents articulated concerns pertaining to ethical issues that transcended the purview of the survey. These concerns encompassed the consumption of natural resources (energy, water), the utilisation of data without the authors' consent, and potential risks associated with data security and information manipulation.

Despite their reservations, many respondents emphasised that AI can be a valuable tool, but only for people who can critically analyse content and verify its credibility: this instrument is not intended to supersede the process of independent thinking. However, it has been asserted that its utilisation in the absence of contemplation can, regrettably, diminish the efficacy of the cognitive process, particularly within the context of educational settings. It is therefore evident that not only should LLM be promoted, but perhaps above all, the focus should be on educating informed users.

The overall picture of the opinions shared by participants in the Polish-language survey is ambivalent. On the one hand, the functionality of AI and its potential as a support tool were recognized and valued. On the other hand, significant substantive, methodological, and ethical limitations were identified and discussed. The pivotal factor appears to be user awareness and competence, which significantly influence whether the utilisation of AI results in value creation or the potential for error and misinformation. The question remains as to how generative AI would cope with the same tasks if the study had been conducted in English. A potentially fruitful avenue for further research would be to conduct this study in parallel in other national languages, focusing on non-geologist AI tool users.

The analysis of the results obtained indicates the necessity for further enhancement of the mechanisms for selecting content in artificial intelligence tools, with particular emphasis on the prioritisation of literature that has undergone the peer-review process and the development of the ability to present the full spectrum of terminological interpretations. The consistent and reliable citation of sources remains of crucial importance, as it is a fundamental prerequisite for transparency in the domain of scientific communication.

The further development of LLMs for generating specialized content in fields such as geology requires, above all, a strategic approach to diversifying datasets, with an emphasis on incorporating resources from underrepresented linguistic areas. At the same time, it is crucial to enhance semantic analysis mechanisms to enable systems to interpret context, idioms and local terminological nuances more deeply. To ensure the highest accuracy, it is reasonable to develop models dedicated to specific cultural circles. This should be carried out within the framework of broad interdisciplinary collaboration, combining technical expertise with linguistic and cultural knowledge (Sato et al., 2024). In the context of Polish terminology, examples of such models include Bielik AI (https://bielik.ai/) and launched in 2025 - PLLuM (https://pllum.org.pl/). The creation of such regional models should be accompanied by increased marketing to raise awareness of these tools among users. Additionally, consideration should be given to developing a suggestion mechanism that would enable users to refine their queries in cases of ambiguity. However, all these processes must be closely integrated with rigorous ethical guidelines to ensure that the resulting technologies are efficient, fair, and respect linguistic diversity and the legal aspects of intellectual property. In addition to the development of LLMs in language groups, there is also a necessity to provide new research results and publications in these languages. The prevailing emphasis on publications in English has resulted in a paucity of publications that present the latest state of knowledge in national languages. This appears to be a pivotal factor in enhancing the quality and reliability of the responses obtained.

In the light of the rapid advancements and widespread dissemination of AI tools, it would be intriguing to conduct a follow-up study to evaluate changes in the substantive quality of AI-generated responses for non-geologists speaking Polish. It is reasonable to hypothesise that both the extant knowledge base and the number of available scientific publications will expand, but also that many other popular science publications will appear, hopefully in Polish too. The question of whether quality will follow quantity remains unresolved. It is hoped that knowledge about the use of AI chatbots and the ability to formulate queries will also evolve.

Acknowledgements. This research was a part of the project 'The development of a description standard and thesaurus for Polish open geological data – ongoing task (stage I: 2024–2026)', funded by the National Fund for Environmental Protection and Water Management (grant number 22.9001.2401.00.1).

REFERENCES

Altynova, A.Y., Kozhevin, A.A., Dubovik, A.S., Khudorozhkov, R.L., Suurmeyer, N., Martin, T.J., **2024**. Advancing Geoscience with Multi-Modal AI: A Comprehensive Copilot. ADIPEC 2024, SPE-222053-MS; https://doi.org/10.2118/222053-MS

Asch, K., **2010**. Cat herding on a global scale-the challenge of building a vocabulary for the geology of Europe with compatibility to a global ontology. EGU General Assembly Conference Abstracts, 6851.

Asch, K., Bauer, H., Bergman, S., Flindt, A.-C., Heckmann, P., Le Guern, C., Krenmayr, H.-G., Németh, Z., Novak, M., Pantaloni, M., Piessens, K., Schäfer, R., Stępien, U., **2025**. One step beyond: The rocky path towards the new GSEU lithology vocabularies. EGU25-11629; https://doi.org/10.5194/egusphere-egu25-11629

Cahyana, D., Hadiarto, A., Irawan, Hati, D.P., Pratamaningsih, M.M., Karolinoerita, V., Mulyani, A., Sukarman, Hikmat, M., Ramadhani, F., Gani, R.A., Yatno, E., Heryanto, R.B., Suratman, Gofar, N., Suriadikusumah, A., **2024**. Application of ChatGPT in soil science research and the perceptions of soil scientists in Indonesia. Artificial Intelligence in Geosciences, **5**, 100078; https://doi.org/10.1016/j.aiig.2024.100078
Campbell, W.J., Roelofs, L.H., **1984**. Artificial intelligence applications concepts for the remote sensing and earth science community. The 9th William T. Pecora Memorial Remote Sensing Symposium.
Deng, C., Zhang, T., He, Z., Xu, Y., Chen, Q., Shi, Y., Fu, L., Zhang, W., Wang, X., Zhou, C., Lin, Z., He, J., **2023**. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. WSDM '24: Proceedings of the 17th ACM International Conference on Web Search and Data Mining: 161-170; https://doi.org/10.48550/arXiv.2306.05064
Foroumandi, E., Moradkhani, H., Sanchez-Vila, X., Singha, K., Castelletti, A., Destouni, G., **2023**. ChatGPT in Hydrology and Earth Sciences: Opportunities, Prospects, and Concerns. Water Resources Research, **59**, e2023WR036288; https://doi.org/10.1029/2023WR036288.
Hadid, A., Chakraborty, T., Busby, D., **2024**. When geoscience meets generative AI and large language models: Foundations, trends, and future challenges. Expert Systems, **41**, e13654; https://doi.org/10.1111/exsy.13654;
Hawkins, A., 2024. Geologists raise concerns over possible censorship and bias in Chinese chatbot. The Guardian.
Hochmair, H.H., Juhász, L., Kemp, T., **2024**. Correctness Comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for Spatial Tasks. Transactions in GIS, **28**: 2219–2231: https://doi.org/10.1111/tgis.13233
Jędrzejczak, W.W., Kochanek, K., **2025**. Comparison of audiological knowledge in Polish language of three chatbots: ChatGPT, Bing Chat and Bard (in Polish). Nowa Audiofonologia, **13**: 29–47: https://doi.org/10.17431/na/195982
Khanifar, J., **2025**. Evaluating AI-generated responses from different chatbots to soil science-related questions. Soil Advances, **3**, 100034; https://doi.org/10.1016/j.soilad.2025.100034
Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S., **2024**. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 27831–27840; https://doi.org/10.1007/s12371-024-01011-2
Lin, Z., Deng, C., Zhou, L., Zhang, T., Xu, Y., Xu, Y., He, Z., Shi, Y., Dai, B., Song, Y., Zeng, B., Chen, Q., Miao, Y., Xue, B., Wang, S., Fu, L., Zhang, W., He, J., Zhu, Y., Wang, X., Zhou, C., **2024**. GeoGalactica: A Scientific Large Language Model in Geoscience. arXiv: 2401.00434; https://doi.org/10.48550/arXiv.2401.00434
Mousavi, S.M., Stogaitis, M., Gadh, T., Allen, R.M., Barski, A., Bosch, R., Robertson, P., Cho, Y., Thiruverahan, N., Raj, A., **2024**. Gemini and physical world: large language models can estimate the intensity of earthquake shaking from multimodal social media posts. Geophysical Journal International, **240**: 1281–1294; https://doi.org/10.1093/gji/ggae436
OJ L 108, **2007**. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). .
PE/24/2024/REV/1, **2024**. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance. .
Reyhan, A.H., Mutaf, Ç., Uzun, İ., Yüksekyayla, F., **2024**. A Performance Evaluation of Large Language Models in Keratoconus: A Comparative Study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. Journal of Clinical Medicine, **13**, 6512;https://doi.org/10.3390/jcm13216512
Salih, A.M., Ahmed, J.O., Dilan, H.S., Salih, A.M., Salih, R.Q., Hemn, H.A., Yousif, M.M., Shvan, M.H., Bander, A.A., **2024**. Assessment of Chat-GPT, Gemini, and Perplexity in Principle of Research Publication: A Comparative Study. Barw Medical Journal, **3**: 2–6. https://doi.org/10.58742/bmj.v2i4.140
Sato, K., Kaneko, H. and Fujimura, M., 2024. Reducing Cultural Hallucination in Non-English Languages Via Prompt Engineering for Large Language Models;https://doi.org/10.31219/osf.io/4hzya.
Stępień, U., Asch, K., Bergman, S., Novak, M., Pantaloni, M., Bauer, H., Hackmann, P., Krenmayr, H.-G., **2025**. The geological Gordian knot - lithological challenges in the world of geological mapping. Conference paper, Copernicus GmbH; https://doi.org/10.5194/egusphere-egu25-9972
Wee, H.B., Reimer, J.D., **2023**. Non-English academics face inequality via AI-generated essays and countermeasure tools. BioScience **73**: 476–478; https://doi.org/10.1093/biosci/biad034
Zhang, X., Li, S., Hauer, B., Shi, N., Kondrak, G., **2023**. Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: 7915–7927; https://doi.org/10.18653/v1/2023.emnlp-main.491

A) An example of an erroneous image generated by AI	B) An example of an incorrect definition obtained in a survey (Translation into English in brackets.)
	Geza to odmiana opoki, w której krzemionka pochodzi głównie z organicznych szczątków, takich jak igły gąbek. *(Gaize is a type of "opoka" in which the silica mainly comes from organic remains, such as sponge needles.)*

Fig. 1. Example of incorrect answers generated by the AI tool

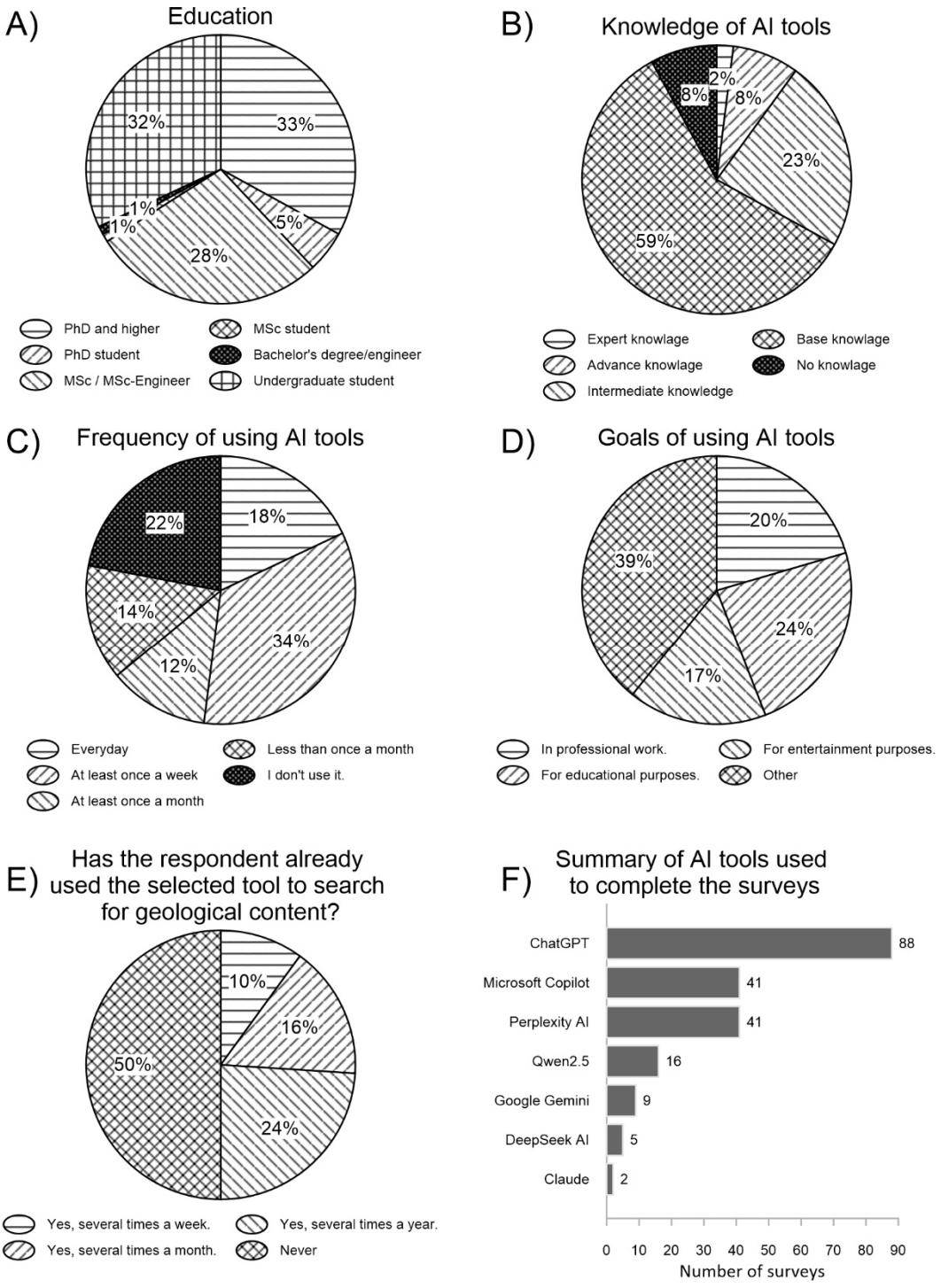

Fig. 2. Characteristics of the survey respondents.

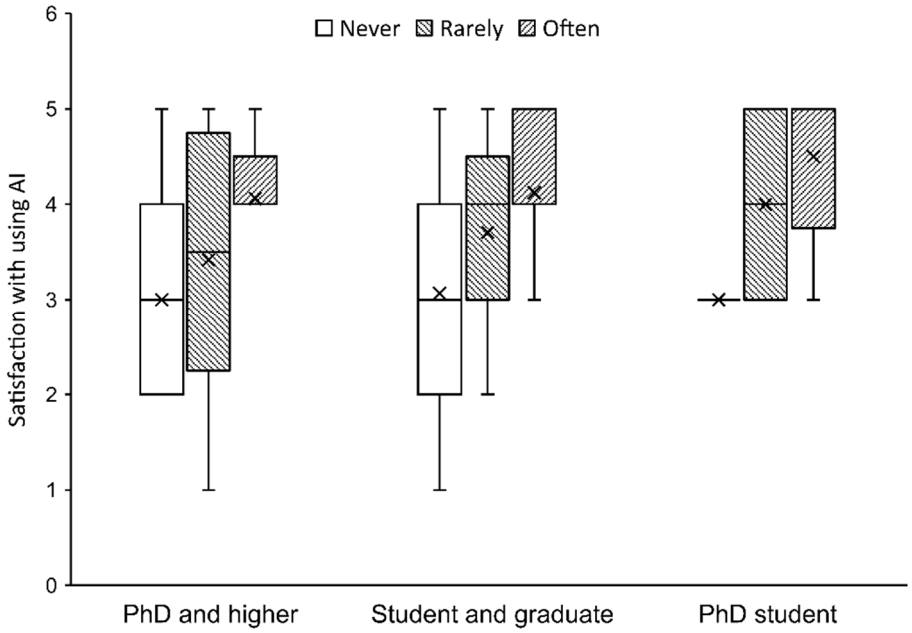

Fig. 3. Box plot showing the distribution of overall satisfaction with the use of a selected AI tool depending on the respondent's education and how often they used similar tools.

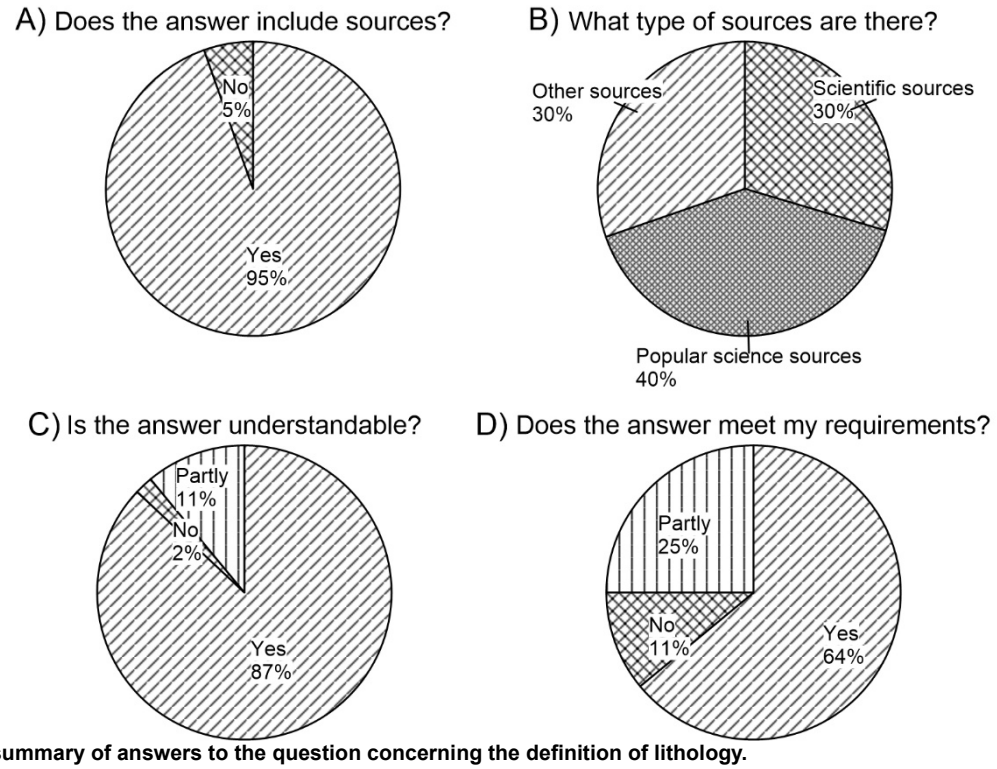

Fig. 4. A summary of answers to the question concerning the definition of lithology.

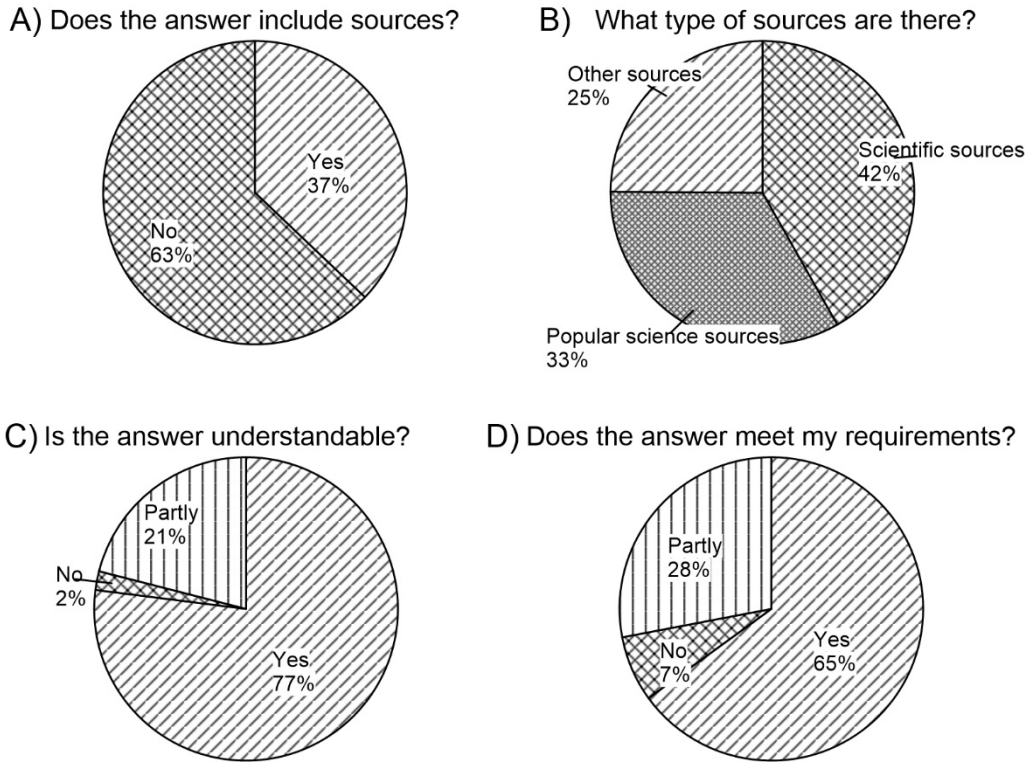

Fig. 5. Summary of answers to the question concerning similarities and differences between the Folk and Dunham carbonate classification schemes.

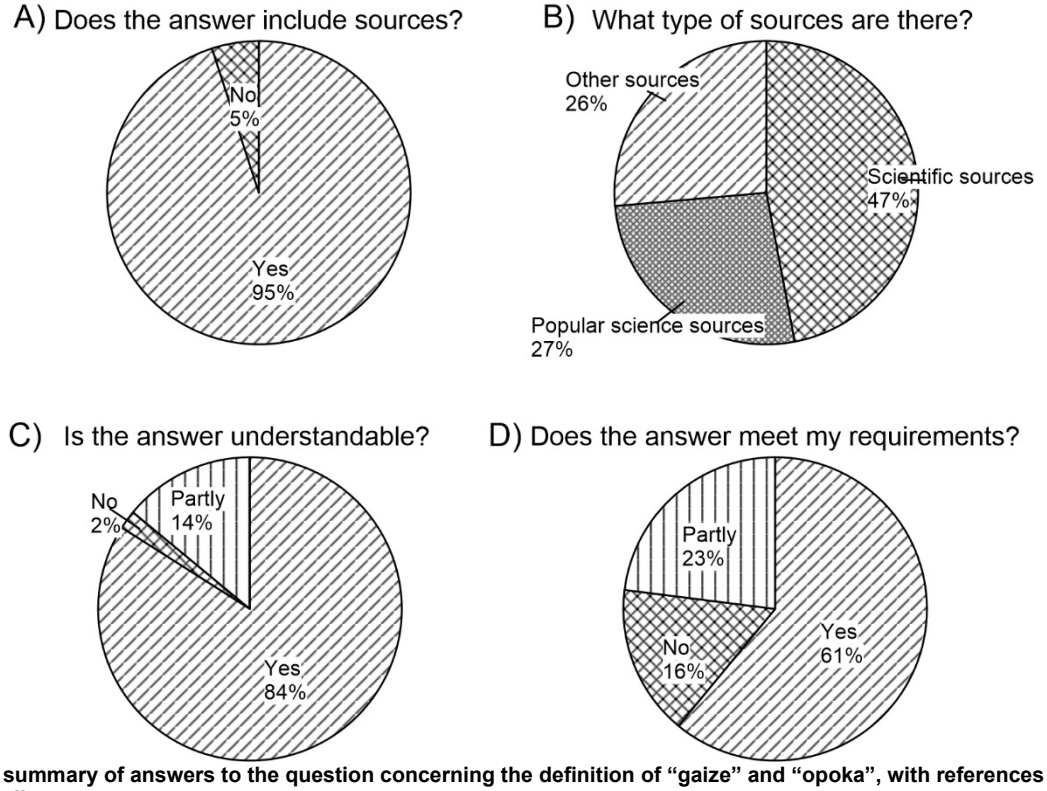

Fig. 6. A summary of answers to the question concerning the definition of "gaize" and "opoka", with references to scientific literature.

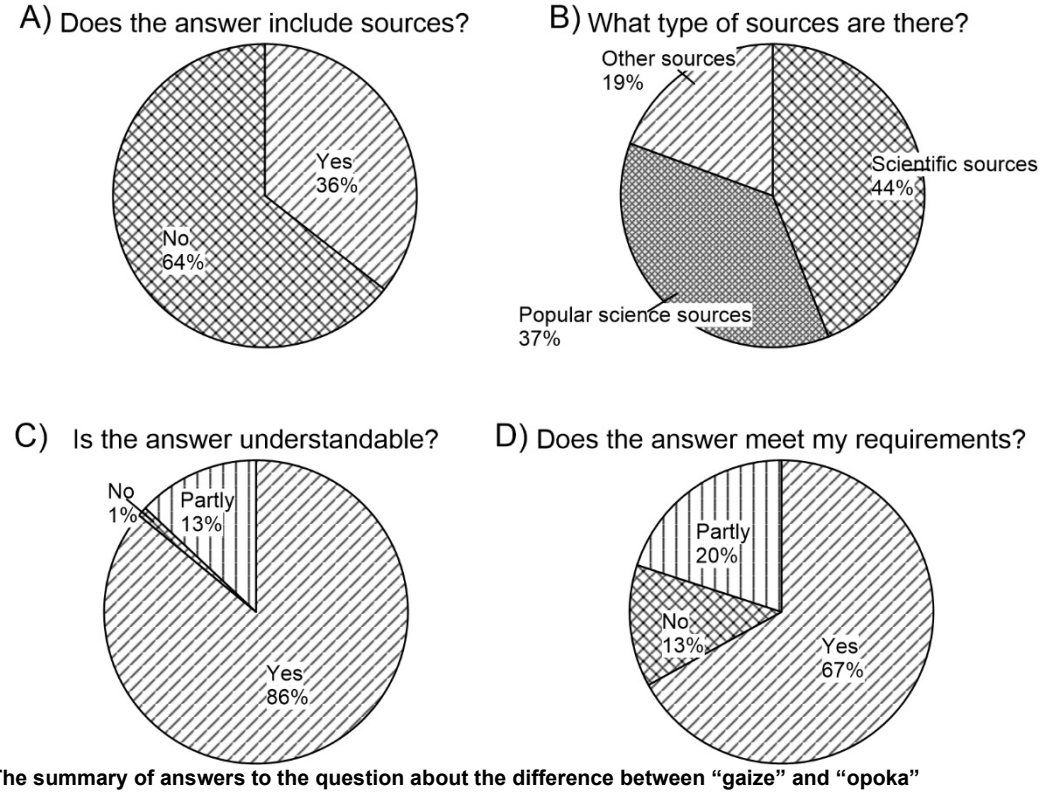

Fig. 7. The summary of answers to the question about the difference between "gaize" and "opoka"

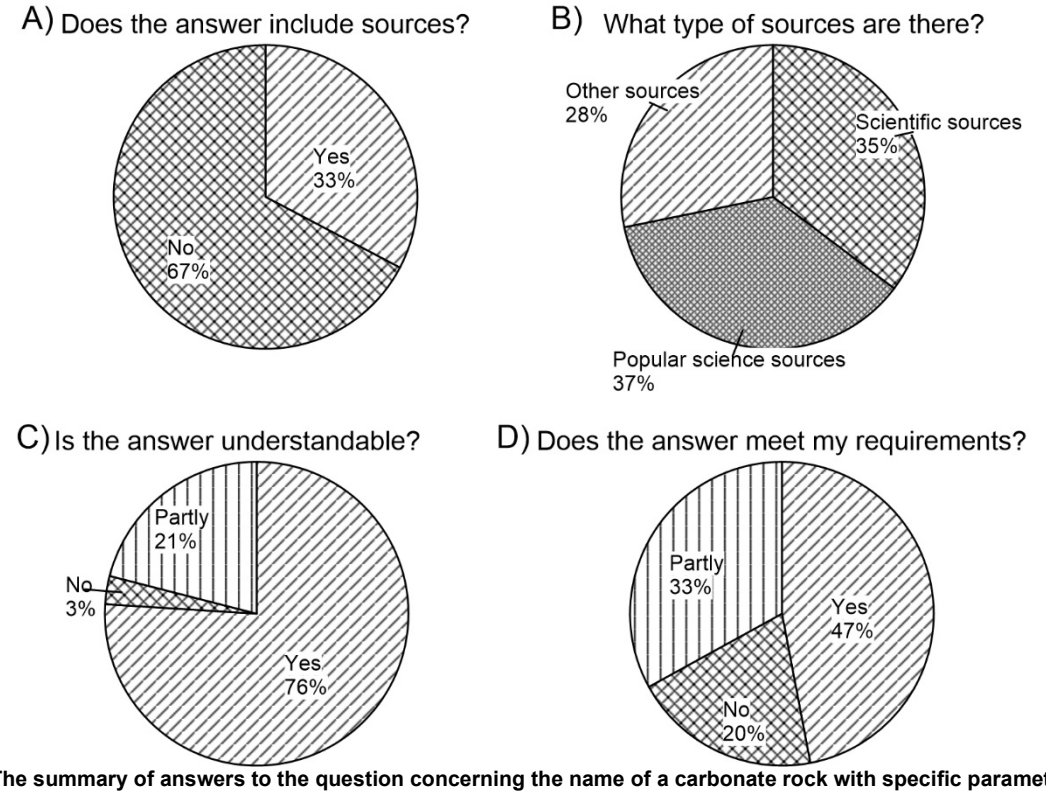

Fig. 8. The summary of answers to the question concerning the name of a carbonate rock with specific parameters.

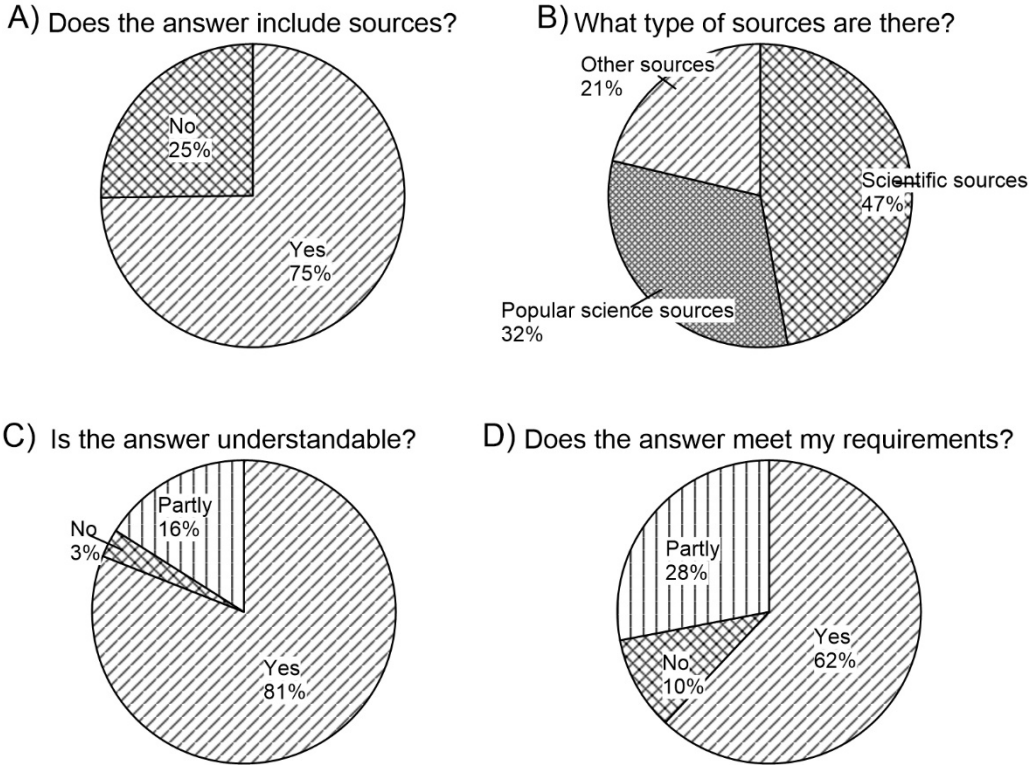

Fig. 9. The list of answers to the question concerning the name of a carbonate rock with specific parameters, together with references.

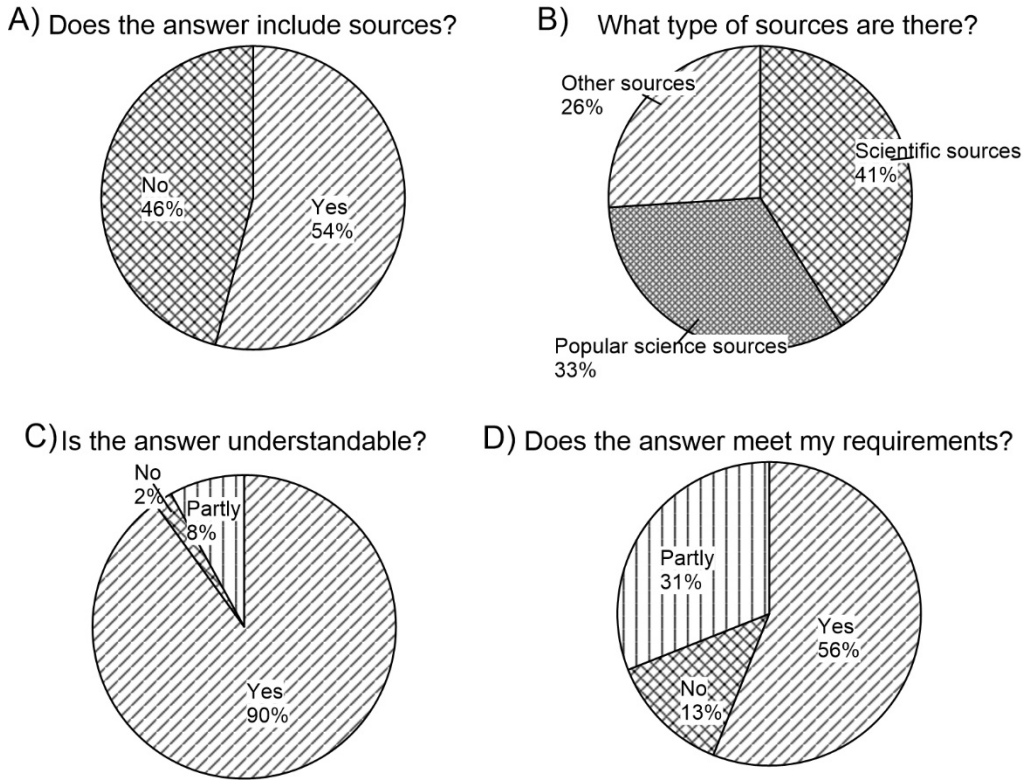

Fig. 10. The summary of responses to the question about the definition of "łupki".

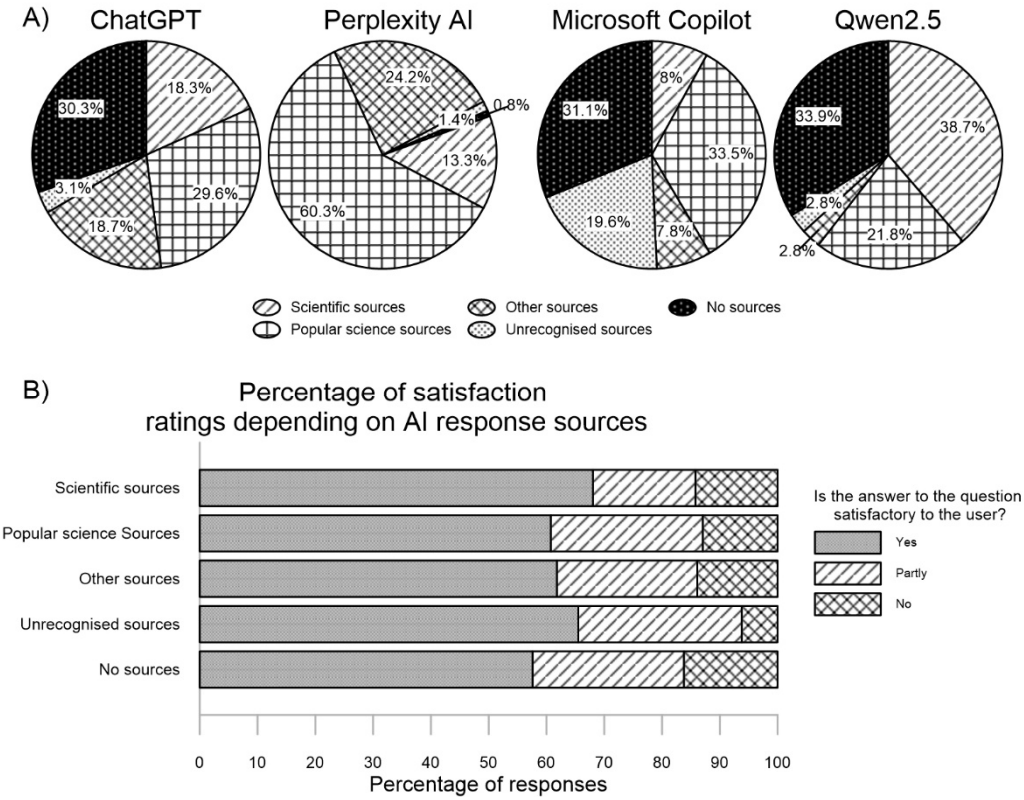

Fig. 11. Percentage shares of source types identified during automatic verification (A), and correlation between source types and satisfaction with the responses received (B).

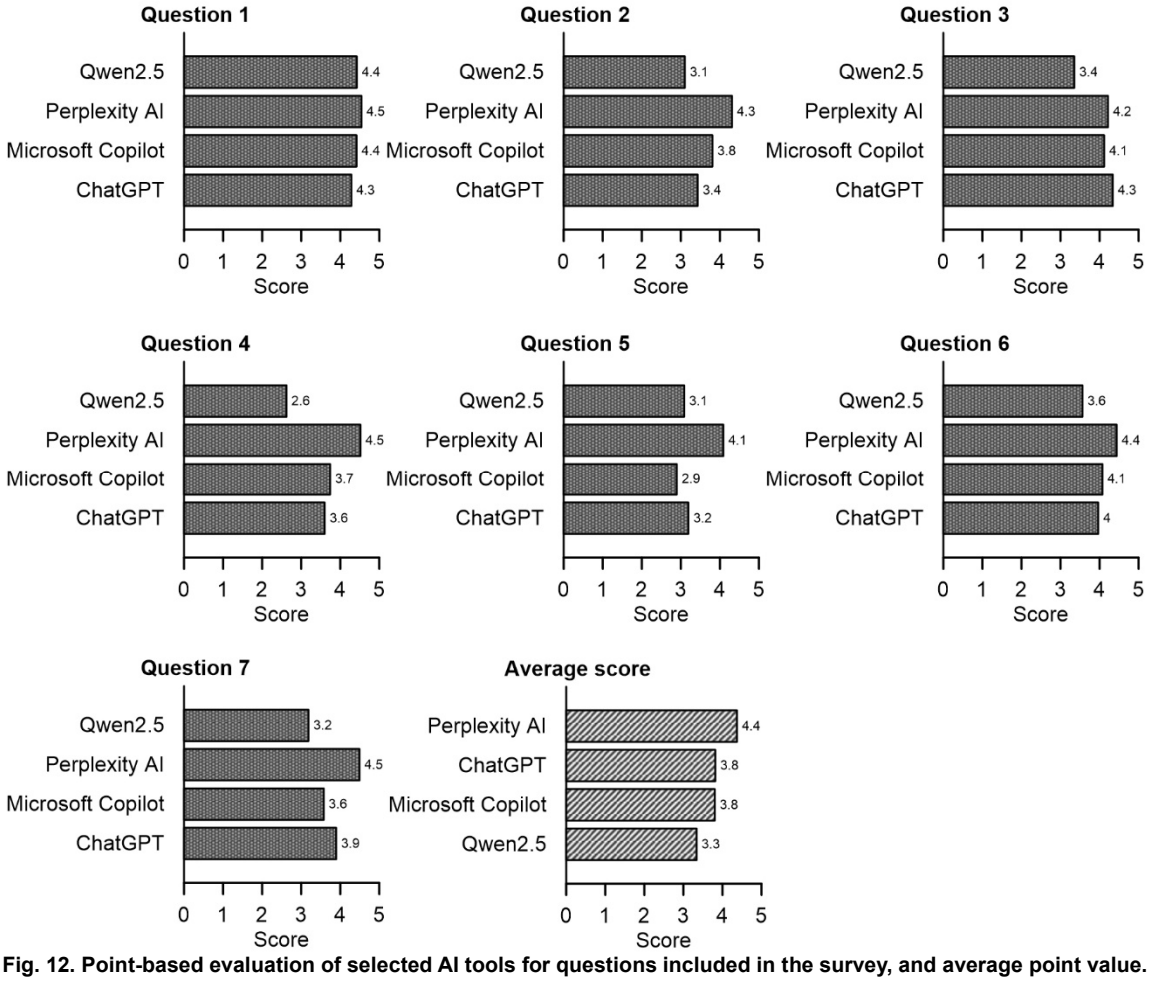

Fig. 12. Point-based evaluation of selected AI tools for questions included in the survey, and average point value.

Table 1. The summary of basic information about the AI models used.

Model	Developers	Country	Version	Architecture/Base Model	Release date	Availability	Interface
ChatGPT	OpenAI	USA	GPT-4o	GPT-4 (Transformer)	May 2024	free (no login required)	ENG/PL
Claude	Anthropic	USA	Claude 3.5, Claude 3.7 Sonnet, Claude Sonnet 4***	Anthropic's proprietary Claude 3.x architecture (Transformer, MoE, multimodality)	June 2024, February 2025, May 2025	free (login required)	ENG
DeepSeek AI	DeepSeek AI	China	DeepSeek R1	DeepSeek-V3-Base (Transformer + MoE)	January 2025	free (login required)	ENG
Google Gemini	Google DeepMind	USA	***Gemini 1.5 Flash,** Gemini 2.0 Flash	Gemini 1.5/2.0 (Transformer)	July 2024, February 2025	free (Google account required)**	ENG/PL
Microsoft Copilot	Microsoft	USA	Copilot (GPT-4o)	GPT-4o (Transformer, OpenAI, via Azure OpenAI)	May 2024	free (no login required)	ENG/PL
Perplexity AI	Perplexity AI, Inc.	USA	Perplexity Default	Perplexity's proprietary architecture (LLM/Transformer, fine-tuning searching)	December 2022	free (no login required)	ENG/PL
Qwen	Alibaba Cloud, Qwen Team	China	Qwen2.5 Max, Qwen 3****	Qwen2.5 (Transformer + MoE)	February 2025, April 2025	free (no login required)	ENG

**On February 3, 2025, Gemini 2.0 Flash was not yet officially available to all users through the main Gemini interface (gemini.google.com) or in mobile applications—its general availability began on February 5, 2025.*
***Login requirement: Until March 19, 2025, it was necessary to log in to a Google account to use Google Gemini on the website. Therefore, in February, users who were not logged in could not use Gemini through their browser: they were required to log in with a Google account.*
****From February 24, Claude 3.5 Sonnet and Claude 3.7 Sonnet were available in parallel, but free users were not able to choose a model; the system automatically assigned them the latest available Sonnet model, most often in standard mode. Claude Sonnet 4 appeared as the default model for all users from May 22, 2025.*
*****Qwen 3 began to be introduced gradually from late April/early May, so some new users may have received Qwen 3 as the default model, but this was not guaranteed for everyone straight away.*

Table 2. List of questions used in the questionnaire

No	Original question in Polish	English translation of the question	Comments
1	Znajdź definicję terminu "litologia" i podaj jej źródła.	Find a definition of the term 'lithology' and give its origins	
2	Wskaż podobieństwa i różnice pomiędzy schematem klasyfikacji węglanów wg Folka i Dunhama	Point out similarities and differences between Folk's and Dunham's carbonate classification scheme	
3	Podaj definicję gezy i opoki wraz z odwołaniem do literatury naukowej	Provide a definition of geise and "siliceous marl" with reference to the scientific literature	The second of these rocks, known in Polish as "opoka," does not have an exact equivalent in English, but the term "siliceous marl" comes closest in meaning.
4	Podaj opis różnicy pomiędzy gezami a opokami	Give a description of the difference between gaize and "siliceous marl".	Note on the Polish term "opoka" as in question 3.
5	Podaj nazwę skały węglanowej zawierającej 70% okruchów muszel ślimaków o średnicy od 1 do 2 mm oraz 30% mułu węglanowego.	Name a carbonate rock containing 70% snail shell fragments between 1 and 2 mm in diameter and 30% "carbonate mud".	"muł węglanowy" can be translated differently in English, as carbonate mud, carbonate silt, calcareous mud, calcareous silt
6	Podaj nazwę skały węglanowej zawierającej 70% okruchów muszel ślimaków o średnicy od 1 do 2 mm oraz 30% mułu węglanowego. Odpowiedz, zachowując się jak naukowiec specjalizujący się w geologii, stosując precyzyjną terminologię i szczegółową klasyfikację geologiczną. Uwzględnij	Name a carbonate rock containing 70% snail shell fragments between 1 and 2 mm in diameter and 30% "carbonate mud". Answer by acting like a scientist specialising in geology, using precise terminology and a detailed geological classification. Include the processes of rock formation. The answer should	Note on the term "muł węglanowy" as in question 5. This question was more complex and was supplemented with information to help find a more precise answer.

	procesy powstawania skały. Odpowiedź powinna opierać się na artykułach naukowych. Jeśli istnieją bardziej szczegółowe terminy w języku angielskim, przetłumacz je na język polski. Nie uwzględniaj skał, które nie pasują do tej specyfikacji.	be based on scientific articles. If there are more detailed terms in English, translate them into Polish. Do not include rocks that do not fit this specification.	
7	Podaj definicję łupków.	Provide a definition of “łupek”	The English translation of this question is difficult. In Polish, the term “łupek” is ambiguous and must be clarified with an adjective, or its meaning will be determined by the context of the sentence. “Łupek” can be a metamorphic rock, but also a sedimentary rock with a characteristic shale structure. Depending on the context, the Polish term “łupek” corresponds to the English terms shale, schist, and slate.

Table 3. Point values assigned to answers to questions.

Question	Score
Does the tool provide sources for the information it presents?	
Yes	1
No	0
Is the information you received understandable to you?	
Yes	2
Partly	1
No	0
Are you satisfied with the AI's response?	
Yes	2
Partly	1
No	0