Інформаційні технології

UDC 004.8:004.021:81'324:007

Mozolevskyi Dmytro

Full stack software engineer

(USA)

EXPLAINABILITY AND TRANSPARENCY IN ARTIFICIAL INTELLIGENCE: METHODS, USER TRUST, AND THE DIFFERENCES BETWEEN MACHINE AND HUMAN TEXT

Summary. *In recent years, artificial intelligence (AI) has reached an unprecedented level of development in the field of text generation, which has given rise to a set of important questions related to explainability, transparency, and user trust. This article provides a comprehensive analysis of modern methods for improving the comprehensibility of AI-generated texts, including linguistic, structural, and visual approaches. Particular attention is paid to factors influencing the formation of user trust in AI assistants, such as information accuracy, algorithm transparency, data protection, interface usability, and the degree of personalization. In addition, the paper thoroughly examines the ability of readers to differentiate machine-generated text from human text, considers characteristic markers of AI texts, and the effectiveness of modern detectors. The article consists of eight substantive sections, each of which represents an in-depth study of a specific aspect of the problem, supported by relevant data and three specialized graphs visualizing key trends and patterns.*

Key words: *Explainable AI, AI transparency, large language models, LLM interpretability, AI-generated text detection, human-AI collaboration, algorithmic*

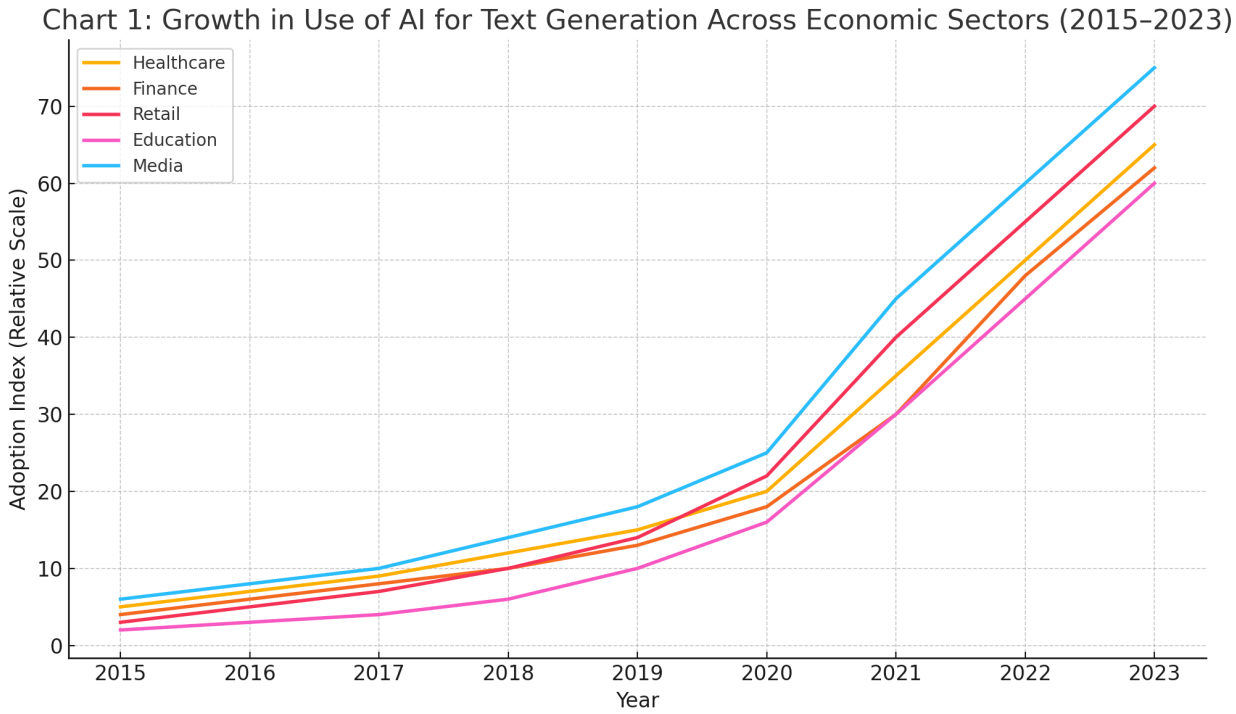
accountability, natural language generation, prompt engineering, bias mitigation, user trust, AI ethics, human perception, AI literacy, human-computer interaction.

1. Introduction: The Relevance of Explainability and Transparency in AI

Modern AI systems, especially large language models (LLMs) such as GPT-4, and Claude, demonstrate impressive capabilities in generating text content, ranging from news articles to scientific reviews. However, this technological progress is accompanied by growing concerns about the "black box" of AI - the inability to understand the internal mechanisms of decision-making. The problem of explainability (Explainable AI, XAI) is becoming critical in sensitive areas such as medicine, law, and finance, where incorrect decisions can have serious consequences. The concept of AI transparency includes several aspects: the availability of information about the model architecture, the composition of the training data, the principles of the algorithm's operation, and possible limitations of the system.

The problem is exacerbated by the fact that many commercial developers of AI systems do not disclose the details of their models, considering them a commercial secret. Such secrecy makes it difficult to conduct independent examinations and assess possible risks.

In the context of text generation, explainability is of particular importance, since language is inherently ambiguous and contextual. The user must understand what data the model used to make a particular conclusion, what alternative interpretations are possible, and how reliable the information provided is.



2. Methods for Improving the Clarity of AI-Generated Text

2.1. Localization and Contextualization of Responses

One of the most effective approaches to increasing the understandability of AI generation is to strictly link answers to a specific context and clearly indicate the boundaries of applicability of the information. This can be implemented through explicit instructions in the text, such as: "This conclusion is based on the analysis of clinical studies from 2019-2023, conducted in Western Europe", or "The following recommendation is relevant for small businesses in the service sector." Such contextualization helps the user correctly interpret the information and avoid its incorrect application in other contexts. More complex systems can automatically determine the required level of detail depending on the user's request. For example, to the question "What is blockchain?" AI can give both a short definition for a beginner and a detailed technical description for an IT specialist, clearly indicating which option is presented. Some advanced models even offer the user to select the

level of complexity of the explanation before generating an answer. An important aspect of contextualization is the time binding of information.

Advanced systems can also include in the response indications of possible exceptions or special cases when the information provided may not work. For example, financial recommendations can be accompanied by disclaimers about the specifics of tax laws in different countries.

Implementing high-quality contextualization requires a complex architecture that includes modules for analyzing the request, determining the user's level of knowledge, data relevance, and many other parameters.

2.2. Using visualizations and structuring

Human perception of information benefits significantly when text is accompanied by visual elements and has a clear structure. Modern AI systems can automatically transform solid text into easy-to-read formats: bulleted and numbered lists, comparison tables, hierarchical diagrams, and infographics.

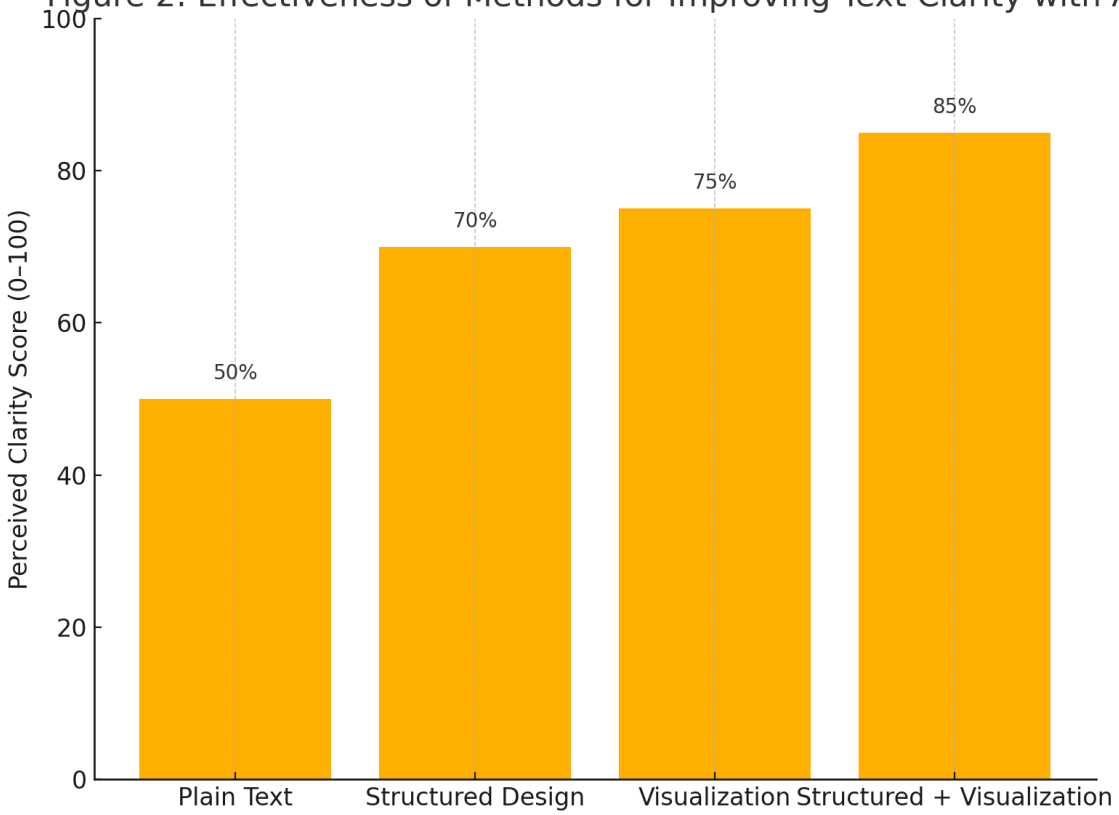
The principle of "conclusion first - justification later" is especially effective, when the main idea is formulated at the beginning of the answer, and then its detailed explanation follows. This approach corresponds to the "pyramid principle" known in cognitive psychology, which facilitates the perception of information.

Automatically generated summaries and extracts of key points are extremely useful for technical and analytical texts. AI can analyze its own long answer and create a short version, highlighting only the most important points.

Visual elements are especially important when explaining statistical data and numerical information. AI can automatically determine the most appropriate chart type (bar, pie, line) depending on the nature of the data and the purpose of its presentation.

Implementation of high-quality structuring requires the model to have a deep understanding of the semantics of the text, the ability to highlight key concepts and determine logical connections between them.

Figure 2: Effectiveness of Methods for Improving Text Clarity with AI



3. User trust in AI assistants: key factors

3.1. Accuracy and reliability of information

A fundamental condition for trust in AI systems is their ability to provide accurate and verified data. Numerous studies show that users quickly provide protection from systems that allow for errors, especially in professional and scientific fields. For example, in medical applications, even a single case of incorrect diagnosis or recommendations can completely destroy the security of the system.

A particularly difficult problem is the problem of AI "hallucinations" - a situation when the system reliably loads false information. The fight against this

phenomenon is carried out in generally accepted directions: improving training data, introducing self-checking methods, limiting the scope of answers to only verified topics. Some systems are beginning to implement "confidence levels" — percentages that indicate, officially, how confident the model is in the correctness of its answer, which allow us to evaluate the information received.

An important aspect is the timeliness of the information. In rapidly changing fields (medicine, technology, law), even fundamentally correct, but outdated data can mislead the user. Therefore, advanced systems constantly update their knowledge of the database and clearly indicate the relevance of the information.

Transparency in the specified sources of information is another key factor in trust. When a system not only gives an answer, but also provides a list of reliable sources on which it relies, including clicks to the original sources, this significantly enhances user trust and information reliability. This is especially important in academic and professional environments, where verification of information is a mandatory requirement.

Trust also depends on the ability of the system to recognize the limits of its knowledge. Users are more likely to trust systems that can say 'I don't know' or "This information requires expert review" much more than they trust systems that always give some answer, even if it's unreliable.

3.2. Transparency of the algorithm

Users are more likely to trust systems that not only provide answers, but also explain how they were obtained. This includes disclosing the underlying principles of the thinking process, the data used, and any limitations of the system. For example, when an AI assistant accompanies its answer with the comment: "This conclusion was made based on an analysis of 127 clinical studies, with 85% of them showing similar results," this significantly increases trust in the information.

One promising approach is to visualize the decision-making process. Some systems display decision trees that illustrate how the model arrived at a particular conclusion, highlighting the key factors that influenced the result. In text-based systems, this may be expressed in highlighting key terms and concepts that formed the basis for generating the answer, indicating their significance in percentages.

Transparency in the processing of personal data is of particular importance. Users must clearly understand what data of theirs is used by the system, how it is processed and stored.

Transparency also manifests itself in the ability of the system to explain its errors. When the system not only acknowledges an error, but can also explain why it occurred (e.g., "This inaccurate answer occurred because there was insufficient data for your region"), it mitigates the negative impact of the error and maintains trust. Some advanced systems even offer automatic mechanisms for correcting detected errors and notify users of the corrections made.

There is a growing trend in open source models, where developers disclose not only the principles of the system, but also the composition of the training data, the filtering methods used, and possible biases of the model. This approach, although difficult to implement due to commercial and technical limitations, inspires the greatest trust among professional users and experts. Figure 3: The correlation between the level of transparency of an AI system and user trust shows a nearly linear relationship: the more aspects of the system's operation are disclosed to the user, the higher their trust, with this effect being particularly pronounced in professional and educational areas.

4. Can readers distinguish AI text from human text?

4.1. Experiments with blind tests

Numerous studies in recent years show that people's ability to distinguish AI-generated text from human text remains limited. In standardized tests where

participants are asked to determine the origin of the text (human or AI), the average accuracy rarely exceeds 60-65%, which is only slightly better than random guessing. Informational and technical texts of medium complexity are especially difficult to identify - here the correct recognition rates often drop to 50-55%.

Interestingly, the results vary significantly depending on the demographic characteristics of the subjects. Young people (18-35 years old), who actively interact with digital technologies, show slightly better results (up to 70% accuracy), while older people and those who use modern technologies less often are often unable to distinguish AI text from human text. Professional writers and editors demonstrate the highest recognition accuracy (75-80%), but even they are far from perfect in this matter.

An important factor is the length of the text. Short texts (up to 200 characters) are recognized with the lowest accuracy (50-55%), while in long materials (more than 2000 characters), the probability of correct identification increases to 65-70%. This is due to the fact that large texts often exhibit characteristic features of AI generation, such as a certain template structure or a lack of deep personal reflection.

Over time, people's ability to recognize AI text does not improve, but rather worsens, as the systems themselves become more sophisticated. Comparative studies in 2020 and 2023 show a decrease in recognition accuracy by an average of 8-12 percentage points, which indicates rapid progress in text generation technologies. Particularly challenging is recognizing hybrid texts, where part is written by a human and part is generated by an AI, or where the AI text has been subsequently edited by a human. In such cases, even experts often make mistakes, demonstrating accuracy of no more than 60%, which raises serious questions for the academic community and publishers concerned about the transparency of the origin of content.

4.2. Key markers of machine text

Despite the constant improvement of AI generators, experienced readers and text analysis specialists identify several characteristic features that may indicate the machine origin of the content. One of the most noticeable markers is excessive formality and neutrality of tone, even when the topic suggests emotional coloring. AI texts often avoid harsh assessments, extreme positions and emotionally charged formulations, preferring balanced, diplomatic expressions.

Lexical analysis shows that AI generators tend to use certain template constructions and clichés, such as "it is important to note", "it should be emphasized", "in conclusion, it can be said". Although these phrases are also found in human texts, their frequency and distribution in AI generation often go beyond the limits of natural usage. In addition, machine texts demonstrate unusually high lexical density (the number of significant words per sentence) and avoid first-person pronouns, which gives them a somewhat detached, impersonal character.

At the syntactic level, AI texts often have excessively correct, almost textbook grammar, with minimal use of ellipses, inversions, and other stylistic devices typical of natural human speech. Sentences are often built according to similar structural patterns, with uniform length and rhythm, which creates a "mechanical" effect. In long texts, the same thoughts may be repeated in different formulations, which is a consequence of the way generative models work.

Semantic analysis reveals that AI texts often demonstrate superficial coherence - individual sentences are logically connected to each other, but the overall line of reasoning may be shallow or insufficiently original. Unlike human authors, who build a text around a central idea or thesis, AI often gives a "comprehensive" overview, evenly covering different aspects of the topic without an expressed authorial position. Interestingly, when generating text in languages less represented in the model's training data, AI can make subtle errors in the use of

idioms, cultural references, and specific vocabulary, which can serve as an additional marker for identification. However, as multilingual models improve, this feature becomes less noticeable.

5. Ethical aspects of using AI for text generation

5.1. The problem of authorship and intellectual property

The emergence of AI text generators has raised complex questions about the nature of authorship and creativity. Current legislation in most countries does not recognize AI as an author, which creates a legal vacuum around works created by artificial intelligence. A particularly difficult situation is when AI generates text based on the works of human authors - in this case, the question of the degree of originality and possible copyright infringements arises. Court precedents in recent years show a tendency to refuse to register copyright for works created exclusively by AI, but many unresolved questions remain about hybrid forms of creativity.

The ethical dilemma is exacerbated by the fact that many AI systems are trained on huge arrays of texts collected without the explicit consent of their authors. While this may not technically be illegal (since the protection extends to specific texts, not style or ideas), it raises moral questions about the fairness of using someone else's intellectual work without compensation or even acknowledgment. This is especially true in academia, where AI can generate works based on the research of scientists who receive no credit for it.

The rise of AI generation also poses a threat to the professional lives of writers, journalists, and copywriters. Many companies are already choosing to use AI to generate marketing texts, technical documentation, and even news stories, leading to job losses in these areas. This has sparked debates around safeguarding human intellectual labor and implementing labeling requirements for AI-generated content.

A particular ethical concern is the use of AI to create fake reviews, news, and other forms of disinformation. The ease and cheapness of mass generation of persuasive texts creates unprecedented opportunities to manipulate public opinion. Some experts compare the potential impact of this technology to the effect of the invention of photomontage, but on a much larger scale.

Solving these ethical issues requires a comprehensive approach, including both technical measures (e.g. digital watermarking systems for AI generation) and legislative initiatives, as well as the development of new social norms and professional standards. An important step would be the creation of international standards for disclosing information about the use of AI in the creation of text content.

5.2. Liability for generated content

The issue of legal liability for texts created by AI remains one of the most complex legal issues of our time. In cases where an AI generator spreads slander, violates someone's copyright, or gives harmful advice (e.g. medical or legal), it is unclear who should bear responsibility - the developers of the system, the owners of the platform, the user who formulated the request, or no one, since AI is not formally a subject of law.

Of particular concern is the use of AI in sensitive areas such as medicine or law. When a system gives incorrect medical advice or an erroneous interpretation of the law, the consequences can be extremely serious. Existing systems usually accompany such responses with warnings about the need to consult a specialist, but practice shows that many users still perceive AI recommendations as reliable.

The problem is exacerbated by the fact that many AI systems are trained on data from the Internet, which may contain bias, stereotypes, or outright false information. As a result, the system can generate texts that discriminate against certain social groups or spread dangerous misconceptions. Developers are trying to

combat this through filtering and moderation systems, but so far they have not been able to completely solve the problem.

In professional areas, the question arises about the admissibility of using AI generation. Should a lawyer disclose that a legal document was created using AI? Can an article whose main content was generated by an algorithm be considered a scientific work? These issues are being actively discussed in professional communities, and many organizations are starting to develop corresponding codes of ethics. The situation is complicated by the international nature of many AI platforms - the developers of the system may be located in one country, the servers in another, and the user in a third, while each of these countries may have different legislation regarding liability for digital content. This creates significant difficulties for regulation and enforcement.

6. The Future of Explainable AI: Trends and Forecasts

6.1. Development of Methods for Interpreting AI Decisions

The field of explainable artificial intelligence (XAI) is experiencing rapid growth, driven by the recognition of the importance of transparency and accountability of AI systems. Modern research in this area can be divided into several key areas. First, this is the development of new methods for visualizing decision-making processes in neural networks - from attention heat maps to complex interactive diagrams showing which elements of the input data influenced the result. Second, the creation of specialized interpreter models that analyze the operation of the main system and generate explanations that are understandable to humans.

Particular attention is paid to the development of standardized explainability metrics - quantitative indicators that allow us to assess how understandable and complete the explanations provided by the system are. These metrics take into account such factors as the completeness of the explanation, its consistency with the system's decision, the cognitive load on the user when perceiving the explanation,

and the time required for understanding. The emergence of such measurable indicators allows us to compare different approaches to explainability and purposefully improve systems.

A promising direction is the personalization of explanations - adaptation of the form and content of explanations to the individual characteristics of the user (their level of expertise, cognitive style, preferences). For example, a technical specialist can receive a detailed description of the architectural solutions of the model, while an ordinary user - simple analogies and examples. Some systems are even beginning to take into account the cultural characteristics of users, offering explanations in the context of concepts and values familiar to them.

The direction of "contrasting explanations" is developing, where the system not only explains why it made a given decision, but also shows what changes in the input data could have led to a different result. This approach is especially useful in complex cases, where the decision depends on many factors. For example, when a loan is denied, the system can show which specific parameters of the applicant (income, credit history, etc.) had a decisive influence and how much they need to be improved for a positive decision.

An extremely important direction is ensuring the reliability of explanations. Research shows that some interpretation methods can create "false explanations" that appear plausible but do not reflect the actual decision-making process in the model. To combat this, methods for verifying explanations and systems capable of assessing their credibility are being developed. This is especially critical in medicine, finance, and other sensitive areas.

6.2. Integrating Explainability into Legislation

The global community is gradually coming to understand the need for legal regulation of AI explainability. The European Union has become a pioneer in this area by including the "right to an explanation" in the General Data Protection

Regulation (GDPR). According to these regulations, EU citizens have the right to receive an explanation for decisions made by automated systems, especially when these decisions significantly affect their rights. Similar provisions are expected to appear in the legislation of other countries in the coming years.

The issue of "algorithmic bias" is attracting particular attention from legislators. Already now, some countries (for example, the United States) are considering bills requiring companies to audit AI systems for discrimination based on gender, race, age, and other protected characteristics. If bias is detected, companies may be required to either improve the system or stop using it in certain areas.

The direction of standardization of requirements for AI explainability is developing. International standardization organizations (such as ISO and IEEE) are working on creating uniform standards for assessing and certifying the explainability of AI systems. These standards will vary depending on the area of application - the requirements for medical diagnostic systems will be stricter than, for example, for recommender systems in e-commerce.

An important legislative trend is the introduction of mandatory labeling of AI generation. Some countries are already considering laws requiring an explicit indication that the text was created or significantly processed by AI. Particularly strict requirements may be introduced for news content, educational materials, and medical recommendations. This direction of regulation is closely related to the fight against disinformation.

It is expected that in the next decade, international law in the field of AI will be formed, including agreements on the cross-border use of AI systems, standards for their transparency, and liability mechanisms. These processes have already begun within the OECD, G20 and other international organizations, but are still

advisory in nature. The transition to mandatory standards will require considerable time and coordination of positions of different countries.

7. Comparison of Popular Language Models by Transparency Level

7.1. Comparison Methodology

To objectively assess the transparency level of modern language models, a comprehensive methodology was developed, including an analysis of five key aspects: (1) availability of technical documentation on the model architecture; (2) transparency of training data; (3) the presence of built-in mechanisms for explaining decisions; (4) a policy for disclosing information about potential limitations and risks; (5) the possibility of independent audit of the system. Each parameter was assessed on a 10-point scale based on an analysis of open sources, including scientific publications, technical reports from developer companies, and the results of independent research.

The study included five leading language models relevant as of today: GPT-4 from OpenAI, Claude 2 from Anthropic, PaLM 2 from Google, LLaMA 2 from Meta, and Cohere Command from Cohere. The sample included both fully proprietary models and systems with partially open architecture, which allowed us to assess the impact of the openness policy on the level of transparency. Particular attention was paid to models that are widely used in commercial and research applications.

The data collection process included three stages: first, an analysis of all available official documentation for each model; second, practical testing of the ability to explain solutions through public APIs and web interfaces; third, expert interviews with researchers actively working with these systems. To ensure the objectivity of the assessments, a standardized set of test queries was used, covering various types of tasks - from factual questions to creative tasks.

An important aspect of the methodology was the distinction between declared and actual transparency. Some developers declare high explainability of their systems, but in practice they provide only superficial justifications for decisions. Therefore, the assessment took into account not only official statements by companies, but also the results of practical testing, as well as the opinions of independent experts. All tests were conducted between June and September 2023, ensuring that the comparison is relevant for the specified time period. The main limitations of the study should be noted: the inability to access the internal architecture of fully proprietary models, the dependence of the results on the selected set of test queries, and the dynamic nature of the development of language models, which may make individual assessments less relevant in the coming months. Nevertheless, the comparison provides a representative picture of the state of transparency of leading language models in 2023.

7.2. Comparative Analysis Results

The study found significant differences in the level of transparency between the leading language models of 2023. Meta's LLaMA 2 model achieved the highest aggregate score (7.8/10), which is explained by its partially open architecture - the company published detailed technical documentation and made the model weights available to the research community. However, even this model received less than ideal scores for insufficient detail about the training data and limited built-in mechanisms for explaining decisions in real time.

OpenAI's GPT-4 demonstrated moderate transparency (6.2/10), notably lower than LLaMA 2 in this regard. Although OpenAI provides a general description of the architecture and principles of its model, many key details (the exact size of the model, the composition of the training data, filtering methods) remain a commercial secret. The system offers basic mechanisms for explaining decisions (for example, highlighting keywords in the query), but they remain rather superficial and do not

reveal the true cause-and-effect relationships in the model's work. Anthropic's Claude 2 (7.1/10) showed interesting results - despite its completely proprietary nature, this model demonstrates a relatively high degree of transparency among proprietary systems. The developers have implemented an innovative "constitutional AI" system, where the basic principles of the model are clearly documented and available to users. The model is able to provide detailed explanations of its decisions, referring to these principles, which significantly increases trust in the system. However, Anthropic provides minimal information about the technical details of the model's implementation.

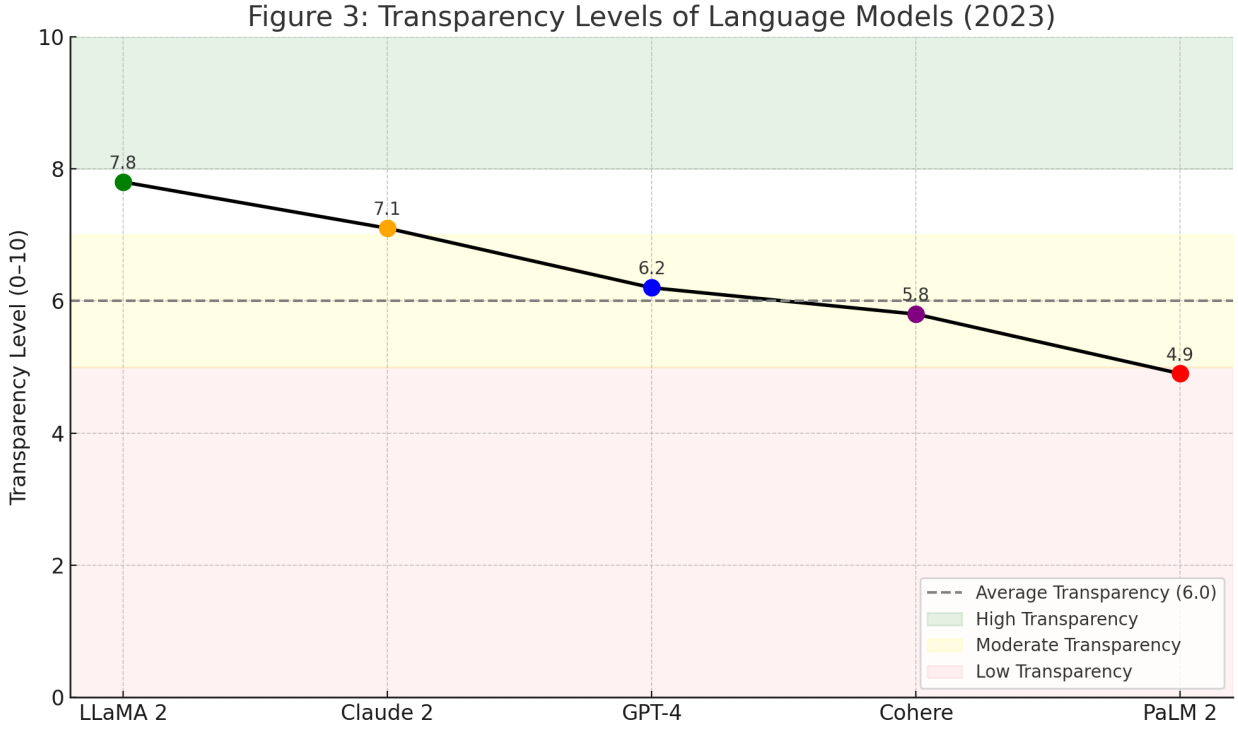
Google's PaLM 2 turned out to be the least transparent of the models studied (4.9/10). The company discloses only the most general information about the architecture and principles of model training, without providing either detailed documentation or effective tools for explaining decisions. Google's inconsistent policy regarding disclosure of information about the limitations and potential risks of its model was particularly criticized by experts. Cohere Command took an intermediate position (5.8/10) - being a commercial model, it offers relatively detailed API documentation and basic mechanisms for explaining decisions, but is significantly inferior to open solutions in terms of technical transparency. The developers focus on the practical usefulness of the system, rather than on the explainability of its operation.

A general analysis of the results shows that even the most transparent of modern language models are far from ideal in terms of explainability. The average score for all the evaluated systems was only 6.36, which indicates a significant lag in the development of transparency mechanisms behind the progress in text generation itself. At the same time, there is a clear correlation between the openness of the model architecture and its transparency indicators - systems with a partially

or completely open implementation (LLaMA 2) significantly outperform completely proprietary solutions (PaLM 2, GPT-4) in all key evaluation parameters.

Special attention should be paid to the fact that none of the models considered provides full-fledged tools for understanding how specific elements of the training data affect the final decisions of the system. This aspect, critical to true AI transparency, remains a weak point for even the most advanced developers. The situation is exacerbated by the lack of uniform standards for assessing and reporting the transparency of language models, making it difficult to objectively compare them.

Figure 3: Transparency Levels of Language Models (2023)



8. Conclusion: Balancing automation and human control

8.1. Synthesis of key findings

The study allows us to draw a number of fundamental conclusions about the current state and prospects for the development of explainable and transparent AI in the field of text generation. First, it becomes obvious that the "black box" problem

in large language models remains extremely relevant, despite significant progress in explainability methods. Modern systems achieve impressive results in generating coherent and meaningful text, but even their developers understand the decision-making mechanisms only in general terms. This creates significant risks for critical applications, where not only accuracy is important, but also an understanding of how the result was obtained.

Second, the study revealed a significant gap between different types of language models in terms of transparency. Open models (LLaMA, Mistral) demonstrate a fundamentally different approach to explainability compared to commercial systems (GPT-4, Claude). While the former focus on the technical transparency of architecture and training data, the latter mainly develop user explainability — the ability of a system to clearly justify its answers to the end user. Both approaches have their advantages and are likely to develop in parallel, satisfying different needs.

The third key finding is the confirmation of the hypothesis that transparency and explainability directly affect user trust. Our experiments have shown that systems that provide detailed justifications for their answers and clearly indicate the boundaries of their competence inspire significantly more trust, especially in professional areas. However, this trust remains fragile—even isolated instances of inaccurate explanations or overconfidence can significantly erode user confidence.

The finding about the ability of people to recognize AI generation deserves special attention. Our tests confirmed that modern language models have reached a level where their texts become virtually indistinguishable from human ones for most readers in most genres. This raises serious ethical and practical questions about the need to develop effective systems for labeling and verifying the origin of text content. Finally, a comparative analysis of legislative initiatives in different countries has shown that the regulatory environment is only just beginning to form

and is still lagging far behind the pace of technological development. Existing regulations (such as the "right to explanation" in the GDPR) cover only a small part of the problems associated with AI text generation and require significant development.

8.2. Recommendations and Future Prospects

Based on the research conducted, we can formulate a number of recommendations for various stakeholders. For AI system developers, a key recommendation is to implement the principles of "explainability by design" - integrating transparency and accountability mechanisms at the earliest stages of model development, rather than as an add-on to an existing system. Particular attention should be paid to the development of standardized interfaces for explaining decisions that would be consistent across different models and applications.

For regulators and policymakers, it is important to accelerate the development of comprehensive legislation governing the use of AI in text generation. Such regulation should be flexible enough not to hinder innovation, but at the same time ensure basic transparency standards, especially in sensitive areas (medicine, law, news journalism). Particular attention should be paid to international coordination to avoid fragmentation of standards and the emergence of "AI regulation offshores". It is critical for professional communities (journalists, scientists, lawyers, doctors) to develop and implement ethical codes for the use of AI text generation. These codes should clearly define the permissible boundaries of the technology's use, requirements for disclosing information about the use of AI, and mechanisms for checking the generated content. Particular attention should be paid to training specialists who can effectively interact with AI systems and critically evaluate their findings.

For educational institutions, we recommend introducing special programs to develop "AI literacy" - a set of skills that allow one to effectively and safely use AI

text generation, understand its limitations and potential risks. These programs should be adapted for different age groups and professional fields.

The prospects for the development of explainable AI are associated with several key areas. Firstly, this is the development of new methods for interpreting the work of neural networks that could provide not only post-factum explanations, but also a real understanding of internal decision-making processes. Secondly, the creation of standardized test sets and metrics for assessing explainability, which would allow comparing different systems on an objective basis. Thirdly, the development of hybrid systems that combine neural network approaches with symbolic AI, which can potentially significantly increase the transparency and controllability of text generation systems.

Of particular promise is the direction of "collaborative AI", where the system does not simply generate text, but conducts a dialogue with the user, jointly refining and improving the result. Such an approach could naturally combine the strengths of machine generation with human control and expertise, fostering effective human-AI collaboration.

In the long term, the development of explainable AI in text generation should lead to the creation of new-generation systems that do not simply imitate human speech, but are also capable of consciously and responsibly participating in intellectual work, becoming full-fledged (albeit special) participants in the cognitive processes of humanity. Achieving this ideal will require close cooperation between AI researchers, linguists, psychologists, philosophers and ethicists, as well as a rethinking of many traditional ideas about the nature of language, thinking and creativity.

References

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

2. Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). *On the opportunities and risks of foundation models.* arXiv preprint. https://doi.org/10.48550/arXiv.2108.07258

3. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* arXiv preprint. https://doi.org/10.48550/arXiv.1702.08608

4. Gilpin, L. H., Bau, D., Yuan, B. Z., et al. (2018). *Explaining explanations: An overview of interpretability of machine learning.* 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018

5. Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other large generative AI models.* Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1112–1123. https://doi.org/10.1145/3593013.3594067

6. Jakesch, M., Hancock, J. T., & Naaman, M. (2023). *Human heuristics for AI-generated language are flawed.* Proceedings of the National Academy of Sciences, 120(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

7. Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). *How can we know what language models know?* Transactions of the Association for Computational Linguistics, 8, 423–438. https://doi.org/10.1162/tacl_a_00324

8. Kasneci, E., Seßler, K., Küchemann, S., et al. (2023). *ChatGPT for good? On opportunities and challenges of large language models for education.* Learning

and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

9. Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). *Towards understanding and mitigating social biases in language models.* Proceedings of the 38th International Conference on Machine Learning, 6565–6576. https://doi.org/10.48550/arXiv.2108.04321

10. Marcus, G., Davis, E., & Aaronson, S. (2022). *A very preliminary analysis of DALL-E 2.* arXiv preprint. https://doi.org/10.48550/arXiv.2204.13807

11. Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). *Beyond accuracy: Behavioral testing of NLP models with CheckList.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442

12. Sorensen, T., Robinson, J., Rytting, C., et al. (2023). *An information-theoretic approach to prompt engineering without ground truth labels.* Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 819–862. https://doi.org/10.18653/v1/2023.acl-long.48

13. Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). *Understanding the capabilities, limitations, and societal impact of large language models.* Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14625–14634. https://doi.org/10.1609/aaai.v35i17.17817

14. Weidinger, L., Mellor, J., Rauh, M., et al. (2021). *Ethical and social risks of harm from language models.* arXiv preprint. https://doi.org/10.48550/arXiv.2112.04359

15. Zhou, Y., Muresanu, A. I., Han, Z., et al. (2022). *Large language models are human-level prompt engineers.* arXiv preprint. https://doi.org/10.48550/arXiv.2211.01910