

A Deep User Interface for Exploring LLaMa

Divya Perumal & Swaroop Panda

Northumbria University

Abstract. The growing popularity and widespread adoption of large language models (LLMs) necessitates the development of tools that enhance the effectiveness of user interactions with these models. Understanding the structures and functions of these models poses a significant challenge for users. Visual analytics-driven tools enables users to explore and compare, facilitating better decision-making. This paper presents a visual analytics-driven tool equipped with interactive controls for key hyperparameters, including top-p, frequency and presence penalty, enabling users to explore, examine and compare the outputs of LLMs. In a user study, we assessed the tool’s effectiveness, which received favorable feedback for its visual design, with particular commendation for the interface layout and ease of navigation. Additionally, the feedback provided valuable insights for enhancing the effectiveness of Human-LLM interaction tools.

Keywords: Large Language Models · User Interface · Hyperparameters · Explainable AI · Visual Analytics

1 Introduction

The emergence of Large Language Models (LLMs), such as OpenAI’s GPT, Google’s BERT, and Meta’s Llama, has fundamentally transformed the field of natural language processing. These models have introduced significant advancements in areas such as text generation [24], summarization [44], and other related applications. In everyday tasks, these LLMs have also found widespread application, significantly enhancing user experiences across various domains. For instance, they are embedded in conversational voice assistants allowing users to interact with technology through natural, conversational language [38,29].

Despite the impressive capabilities of LLMs, their complex nature often presents challenges in terms of interpretability and usability [41]. As these models continue to grow in size and complexity, the need for effective tools to facilitate the communication and visualization of their inputs and outputs becomes increasingly urgent. For example, LLMs are being increasingly utilized in the medical field for diagnosing diseases [6], recommending treatments [40], and generating medical reports [36], as well as in data science, where they contribute to data analysis[28], trend prediction[18], and decision support [39]. However, the opaque, *black-box* nature of LLMs [5,26] presents a considerable barrier to validating the accuracy and reliability of their results. There remains a substantial

gap in understanding [27] how these models function and how they can be modified to produce different outcomes. The impact of these applications is largely contingent on the precision and appropriateness of the models, underscoring the critical need for transparency [23] and accountability [10] in AI systems, particularly as they become increasingly integrated into daily life.

Explainable AI and visual analytics [1,12] methods have been adapted to mitigate this black box nature of LLMs. These approaches facilitate a deeper understanding of model behavior by providing insights [42] into decision-making processes and enabling users to visualize the relationships between input parameters and output predictions, thereby enhancing interpretability [2] and trustworthiness [31]. By employing techniques such as interactive visualizations[14,4] or explorations[17], these methods empower users to gain a clearer understanding of how LLMs generate responses. This increased transparency is essential for fostering user confidence[21] and ensuring responsible AI [3] deployment, particularly in sensitive applications such as healthcare, finance, and legal decision-making.

In this paper, we present a visual analytics-driven tool to facilitate an understanding of working of LLMs by examining their underlying hyperparameters. The contributions of the paper are,

1. We design and develop a deep user interface to facilitate the hyperparameter driven exploration of a Large Language Model
2. We evaluate this user interface through user feedback and incorporate their suggestions to enhance its design.
3. We provide preliminarily actionable insights for researchers to design and develop user interfaces for LLMs.

2 Background

2.1 Visual Analytics based User Interfaces

Visual Analytics (VA) combines visualization, human interaction, and data analysis to support analytical reasoning through interactive visual interfaces [9]. Visual analytics tools such as model performance dashboards help in identifying issues like overfitting, bias, or data leakage by visually representing model performance metrics enhance the AI interpretability [43]. VA has also been used to understand the inner workings of neural networks by visualizing activations and weights help the interpretation of the model decisions [13]. VA supports these cognitive processes by providing interactive tools that clear the way for deeper insights and better decision-making [11].

2.2 Interpretability and Exploration for LLMs

Interpretability in machine learning refers to the capacity to comprehend, trust, and manage a model’s outputs and decision-making processes. The inherent black-box nature of LLMs, attributed to their deep neural network architectures comprising millions or even billions of parameters, presents significant challenges

in this regard. Various interpretability methodologies have been developed, including post-hoc approaches, which generate explanations after a model has produced a prediction. Many of these methods do not directly interpret the model itself but rather attempt to elucidate its decision-making process. Several techniques, such as SHapley Additive exPlanations (SHAP) [20] and Local Interpretable Model-agnostic Explanations (LIME) [35], assign importance scores to input features. In the context of LLMs, these methods are commonly employed to analyze the significance of specific words, phrases, or tokens. The resulting visualizations aid in understanding how different linguistic elements influence predictions, thereby offering insights into complex model behavior. These techniques may not always capture the true reasoning of the model, potentially leading to misleading interpretations. Moreover, the application of these methods to LLMs, which involve extensive parameter sets and input tokens, is computationally expensive and complex. Many LLMs, such as BERT, utilize self-attention mechanisms to assign varying levels of importance to different words within a sentence [16]. Researchers have explored the interpretation of attention weights as proxies for understanding model decisions [8,22]. However, the extent to which these weights provide meaningful insights into a model’s reasoning remains an area of ongoing investigation.

2.3 User Interfaces for LLMs

One of the key challenges in utilizing LLMs is the selection of optimal hyper-parameters, which significantly influence model performance [25]. Visual analytics facilitates hyper-parameter tuning by providing interactive visualizations that help both experts and non-experts discern parameter relationships, enhancing model optimization. For example, Google Vizier employs parallel coordinates plots to analyze hyper-parameters, elucidating the relationships between inputs and outputs [15]. Sacha et al. [34] proposed an interactive machine learning framework that incorporates human feedback via visualization, improving parameter refinement. Similarly, Kahng et al. [19] introduced a tool for comparing and evaluating LLM outputs during fine-tuning, addressing visualization challenges but still limited in scope and standardization. Chen et al. [7] explored visual analytics in ChatGPT through StudentGPT, a tool that analyzes student interactions to derive cognitive insights. Despite the availability of such tools, there remains a gap in integrating visual analytics for hyper-parameter analysis where human intuition is crucial [32].

Although visual analytics has been explored in various machine learning contexts, its application to hyper-parameter tuning in LLMs remains relatively under-explored, with limited dedicated tools. Hyperparameter tuning is crucial for optimizing model performance, yet it often remains an empirical process with limited interpretability. Visual analytics could aid in exploring the hyper-parameter space and understanding parameter interactions, though its impact on tuning efficiency requires further validation.

3 Design & Development of the User Interface

The visual analytics tool was designed and developed to enable users to adjust model hyperparameters visually, making it easier to see how changes affect the LLM’s outputs. This involved creating intuitive visualizations for each hyperparameter to facilitate user interaction. For this tool, we choose the open source large language model Llama [37].

3.1 Hyperparameters

We selected two hyperparameters, top p and frequency and presence penalty based on their availability via the LLaMA API.

Top-P: The top-p hyperparameter [33], also known as nucleus sampling, is a technique used in NLP to control text generation. It plays the role of selecting the next word in a sequence impacting the creativity of the generated text. In LLMs, content generation is the prediction of the next word in a sequence based on the words that generated previously. The model generates a list of all possible tokens and ranks them by their predicted probability. Top-p sampling is a method used to choose the next word based on these probabilities. The model accumulates the probabilities until the sum reaches a threshold value, p , which is set by the top-p parameter. At this point, the model forms a candidate pool of tokens, and the next token is randomly selected from this pool. This randomness introduces variation in the generated text, making it more creative. The top-p vocabulary $V^{(p)}$ is the smallest subset of the total vocabulary V where the cumulative probability mass meets or exceeds the threshold value p (Eq. 1). The selection of the next word depends on ensuring that the cumulative probability is at least p , which adds randomness and diversity to the generated text.

$$\sum_{x \in V^{(p)}} P(x | x_{1:i-1}) \geq p. \quad (1)$$

Frequency and Presence penalty: Frequency and presence penalty [30] are two hyperparameters used in language models to control the repetition and diversity of the generated text. Both are designed to influence how often certain words or phrases appear in the generated text. Frequency penalty reduces the repetition of the same token within the generated text. The presence penalty encourages the model to introduce a new token into the generated text. The presence penalty works by reducing the probability of selecting a token that has already appeared in the text, regardless of how many times it has been used. To adjust token probabilities for diversity and repetition control, we define the new modified probability $P'(t)$ for token t as:

$$P'(t) = \frac{P(t)}{(1 + \alpha \cdot f(t))(1 + \beta \cdot \mathbf{1}_{\{f(t) > 0\}})} \quad (2)$$

where $P(t)$ is the original probability, $f(t)$ is the frequency of t in prior text, α is the frequency penalty, and β is the presence penalty and $\mathbf{1}_{\{f(t) > 0\}}$ is an

indicator function that equals 1 if token t has appeared at least once in the text and 0 otherwise. The frequency penalty reduces $P(t)$ proportional to $f(t)$, while the presence penalty reduces $P(t)$ if t has already appeared, enhancing diversity in the generated text.

3.2 Development & Visual Design

The tool was developed using Python for backend logic and integration with the Llama API from *Meta LLAMA-7B*, Flask for building the web interface, d3js for interactive data visualizations in web browsers and MongoDB for storing user interactions and survey responses.

To visualize the top-p hyperparameter we use a knob (clock-like) interface that allows users to adjust the parameter values (which controls the nucleus sampling in the language model). This visual representation aims to provide an intuitive way for users to adjust and understand the impact of the top-p hyperparameter on text generation, allowing for more or less randomness in the output based on the setting.

To visualize the frequency and presence penalty we use a co-ordinate plane graph with x and y axis which represent presence and frequency penalty respectively. This scatter plot-type of visualization aids users in comprehending the relationship between these two hyperparameters as reflected in the model’s output.

In Fig. 1 (A) is the prompt section where the user enters their prompts to interact with the LLM. (B) represents the hyperparameters section, which is designed to be visually engaging. (C) is the graphical representation to adjust the presence and frequency penalty. (D) denotes the previous point used, which enables the user to compare outputs given the hyperparameter value.

Fig. 2 (E) denotes the output of the LLaMA. The content given in the image is for sample. (F) is the rating scale for the users to score the generated output; for the user study.

Fig. 3 (G) represents the graph with points of frequency and presence previously used for the given prompt (from Fig. 1 D). The shade of the points represents the score given to the output by the user. Fig. 3 (H) indicates the table with all the prompts, parameter values used and rating for each output.

4 Evaluation of the User Interface

4.1 User Study

With approval from the relevant institutional ethics committee, participants were selected from among students enrolled in postgraduate programs, ensuring a decent understanding of LLM usage. Participants had academic backgrounds in computer and data science. A total of 10 participants took part in the experiment. Participants were of the age group *mean:24.3 yrs, stddev:0.82*.

Participants received a list of ten predefined prompts and were instructed to select three. They subsequently inputted these chosen prompts into the visual

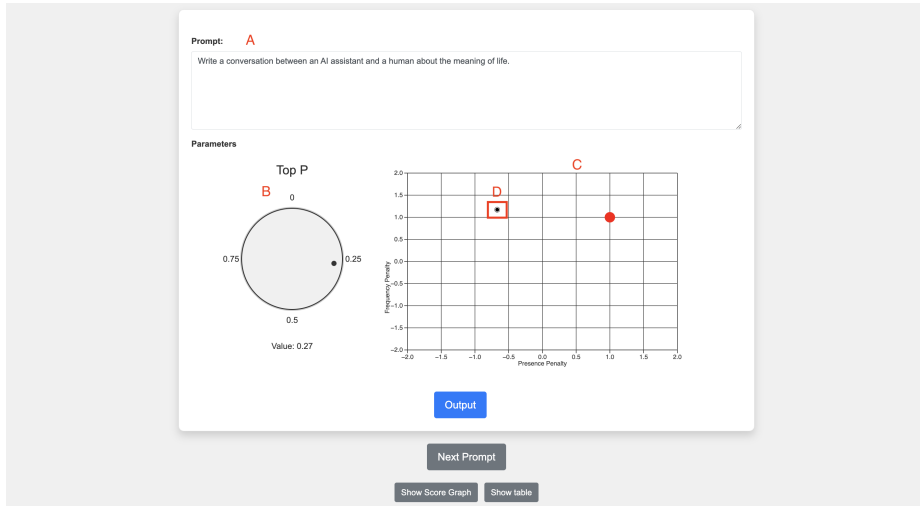


Fig. 1. Visualization of Hyperparameters in the UI

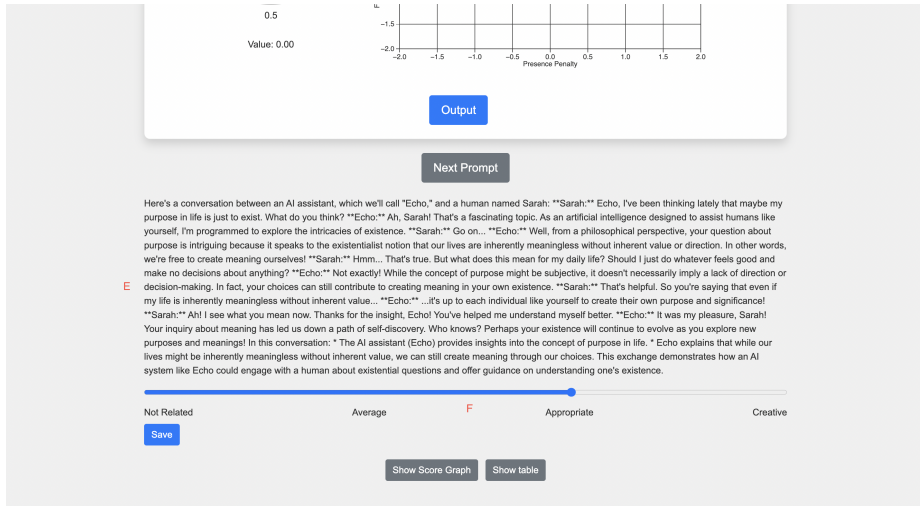


Fig. 2. Visual Representation of LLM Outputs with the Scale for User Rating

analytics tool. Utilizing the tool’s visualized interfaces, participants adjusted hyperparameters— top-p, frequency penalty, and presence penalty—to examine their impacts on the model’s outputs. After setting these hyperparameters, participants submitted the prompts to the LLaMA to generate corresponding text outputs. Each output was subsequently evaluated on a standardized rating scale. These tasks and exploration were designed to assess how effectively participants could use the tool to change hyperparameters and compare, explore,

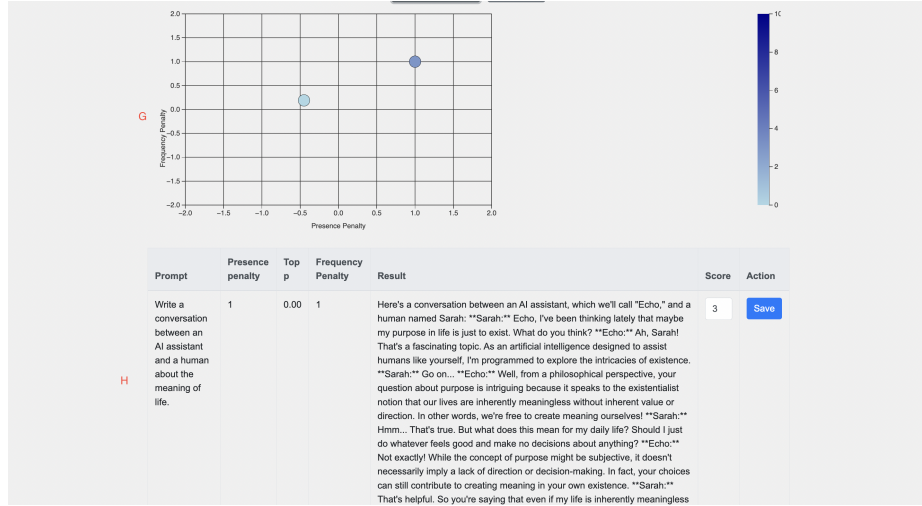


Fig. 3. Visual Representation of Score Graph and Table

evaluate the generated outputs. Followed by the experiment participants were asked to complete survey with combination of likert scale and qualitative questions to gain feedback on the tool’s functionalities, interface and overall user experience.

4.2 Results & Analysis

Based on the user study results, participants provided overall positive feedback, particularly regarding the tool’s interface and its visual appeal (Table 1). Users found the interface visually engaging, with the layout receiving especially high ratings, indicating that the design is well-organized and easy to navigate. The interface of the tool is found to be visually appealing to the users, with a mean rating of 4.22 and a relatively low standard deviation of 0.83, indicating the most respondents appreciated the design. The layout of the interface was rated even higher, with a mean of 4.56 and a standard deviation of 0.53, reflecting that the interface was well organized and easy to navigate.

Qualitative feedback yielded valuable insights; users expressed a desire for more clarification regarding the functionality of the hyperparameters. Participants highlighted features that enhanced their experience with the system. They appreciated the ability to see varying outputs for prompts. One participant stated that *"It was a good experience to see varying outputs for the prompts. The outputs differed on the parameters which provided different kinds of outputs which could then be rated as creative, average, or satisfying."* The feature of adjusting hyperparameters using visualizations stood out as particularly convenient, with a participant noting, *"the feature of adjusting parameters using visualizations was the most convenient one, along with the option of accessing history, which*

helped to compare the previous parameters." Access to history was also valued for its role in facilitating comparisons between different parameter settings. The user interface was praised for its simplicity, making the selection of parameters straightforward. One participant suggested "selecting the parameters was easy as it was simple as the user interface was simple. For the prompts provided, it was helpful to play around with the parameters to check various outputs.". Another participant stated, "most helpful parameter my opinion would be changing the answers as user changes the parameter."

Following the user study, we refined the tool based on user feedback to enhance clarity and usability. One key improvement was the addition of written descriptions for each hyperparameter, providing users with a clearer understanding of their influence on the model's behavior. This change aims to make the tool more accessible, especially for those with limited prior knowledge, enabling more informed adjustments and interactions.

<i>Question</i>	<i>Mean</i>	<i>StdDev</i>
How visually appealing do you find the interface of the visual analytics tool?	4.22	0.83
How well-organized is the layout of the tool's interface?	4.55	0.52
How readable and appropriate is the typography used in the tool?	4.66	0.5
How easy was it for you to navigate and use the visual analytics tool?	4.33	0.86
How effective are the tool's features in helping you achieve your goals?	4.33	0.70
How effective is the answer changing according to the given parameters?	4.22	0.66
How clear and understandable are the visualizations provided by the tool?	4.33	0.5
How would you rate the interactivity of the visualizations?	4.3	0.67

Table 1. Results of the User Study on a 5-point Likert Scale

5 Discussion

While the findings of the user study offer valuable perspectives, it is important to interpret them with caution due to the limited sample size. While the tool's navigation was generally rated as easy with a mean score of 4.33, the standard deviation of 0.86 suggests that a minority of users encountered difficulties. The tool's functional effectiveness, particularly in how well the features helped users achieve their goals, the responses were favorable. This indicates that the tool was generally effective though there were some variations in user experience,

suggesting that certain features might benefit from further enhancement to ensure consistent effectiveness.

5.1 Actionable Insights

Preliminary insights suggest some considerations for the design of user interfaces for LLMs. These following actionable insights provide guidance on optimizing interaction paradigms, improving usability, and enhancing the overall user experience.

1. Providing users with access to adjustable hyperparameters within the interface appears to facilitate exploration, particularly when these hyperparameters have a substantial impact on the generated outputs. This feature may enhance user engagement and comprehension of model behavior.
2. While enabling hyperparameter-based exploration, it is beneficial to incorporate a history of generated responses. Such a feature allows users to compare outputs systematically, fostering a clearer understanding of how hyperparameter adjustments influence the model's responses over time.

5.2 Future Work

Future research includes exploring more hyperparameters of LLMs and developing visualizations that enhance user understanding, thereby facilitating more targeted LLM outputs. Furthermore, a more comprehensive user study involving diverse participants could reveal patterns in hyperparameter adjustments and user ratings, thereby providing deeper insights into user behavior and preferences.

6 Conclusion

The aim of our study was to explore how visual analytics could be integrated into the process of tuning hyperparameters in LLMs to improve user experience and make these models more interpretable. From our initial findings, we discovered that a more intuitive user interface—one that lets users adjust model settings, view and compare results—can transform complex models like LLaMa from a "black box" into something more transparent and user-friendly. The positive feedback we received on the design of this tool shows great potential for its wider use, making these advanced models more accessible and easier to understand.

References

1. Alicioglu, G., Sun, B.: A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics* **102**, 502–520 (2022)
2. Andrienko, N., Andrienko, G., Adilova, L., Wrobel, S.: Visual analytics for human-centered machine learning. *IEEE Computer Graphics and Applications* **42**(1), 123–133 (2022)
3. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* **58**, 82–115 (2020)
4. Batch, A., Elmqvist, N.: The interactive visualization gap in initial exploratory data analysis. *IEEE transactions on visualization and computer graphics* **24**(1), 278–287 (2017)
5. Bhattacharjee, A., Moraffah, R., Garland, J., Liu, H.: Llms as counterfactual explanation modules: Can chatgpt explain black-box text classifiers? *arXiv preprint arXiv:2309.13340* (2023)
6. Chen, X., Mao, X., Guo, Q., Wang, L., Zhang, S., Chen, T.: Rarebench: Can llms serve as rare diseases specialists? In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pp. 4850–4861 (2024)
7. Chen, Z., Wang, J., Xia, M., Shigyo, K., Liu, D., Zhang, R., Qu, H.: Stugptviz: A visual analytics approach to understand student-chatgpt interactions. *IEEE Transactions on Visualization and Computer Graphics* (2024)
8. Coscia, A., Holmes, L., Morris, W., Choi, J.S., Crossley, S., Endert, A.: iscore: Visual analytics for interpreting how language models automatically score summaries. In: *Proceedings of the 29th International Conference on Intelligent User Interfaces*. pp. 787–802 (2024)
9. Cui, W.: Visual analytics: A comprehensive overview. *IEEE access* **7**, 81555–81573 (2019)
10. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O’Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., et al.: Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017)
11. Endert, A., Ribarsky, W., Turkay, C., Wong, B.W., Nabney, I., Blanco, I.D., Rossi, F.: The state of the art in integrating machine learning into visual analytics. In: *Computer Graphics Forum*. vol. 36, pp. 458–486. Wiley Online Library (2017)
12. Epp, C.D., Bull, S.: Uncertainty representation in visualizations of learning analytics for learners: Current approaches and opportunities. *IEEE Transactions on Learning Technologies* **8**(3), 242–260 (2015)
13. Garcia, R., Telea, A.C., da Silva, B.C., Tørresen, J., Comba, J.L.D.: A task-and-technique centered survey on visual analytics for deep learning model engineering. *Computers & Graphics* **77**, 30–49 (2018)
14. Godfrey, P., Gryz, J., Lasek, P.: Interactive visualization of large data sets. *IEEE transactions on knowledge and data engineering* **28**(8), 2142–2157 (2016)
15. Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., Sculley, D.: Google vizier: A service for black-box optimization. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1487–1495 (2017)
16. Insuasti, J., Roa, F., Zapata-Jaramillo, C.M.: Computers’ interpretations of knowledge representation using pre-conceptual schemas: An approach based on the bert and llama 2-chat models. *Big Data and Cognitive Computing* **7**(4), 182 (2023)

17. Jankun-Kelly, T., Ma, K.L., Gertz, M.: A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics* **13**(2), 357–369 (2007)
18. Jin, M., Tang, H., Zhang, C., Yu, Q., Liu, C., Zhu, S., Zhang, Y., Du, M.: Time series forecasting with llms: Understanding and enhancing model capabilities. *arXiv preprint arXiv:2402.10835* (2024)
19. Kahng, M., Tenney, I., Pushkarna, M., Liu, M.X., Wexler, J., Reif, E., Kallarackal, K., Chang, M., Terry, M., Dixon, L.: Llm comparator: Visual analytics for side-by-side evaluation of large language models. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. pp. 1–7 (2024)
20. Khediri, A., Slimi, H., Yahiaoui, A., Derdour, M., Bendjenna, H., Ghenai, C.E.: Enhancing machine learning model interpretability in intrusion detection systems through shap explanations and llm-generated descriptions. In: *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*. pp. 1–6. IEEE (2024)
21. Kizilcec, R.F.: How much information? effects of transparency on trust in an algorithmic interface. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. pp. 2390–2395 (2016)
22. La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., Angelini, M.: State of the art of visual analytics for explainable deep learning. In: *Computer Graphics Forum*. vol. 42, pp. 319–355. Wiley Online Library (2023)
23. Larsson, S., Heintz, F.: Transparency in artificial intelligence. *Internet policy review* **9**(2) (2020)
24. Li, J., Tang, T., Zhao, W.X., Nie, J.Y., Wen, J.R.: Pre-trained language models for text generation: A survey. *ACM Computing Surveys* **56**(9), 1–39 (2024)
25. Li, T., Convertino, G., Wang, W., Most, H., Zajonc, T., Tsai, Y.H.: Hypertuner: Visual analytics for hyperparameter tuning by professionals. In: *2018 IEEE Workshop on Machine Learning from User Interaction for Visualization and Analytics (MLUI)*. pp. 1–11. IEEE (2018)
26. Liu, H., Yin, Q., Wang, W.Y.: Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196* (2018)
27. Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., Zhang, Y., Wang, J., Gao, X., Zhong, T., et al.: Understanding llms: A comprehensive overview from training to inference. *arXiv preprint arXiv:2401.02038* (2024)
28. Ma, P., Ding, R., Wang, S., Han, S., Zhang, D.: Insightpilot: An llm-empowered automated data exploration system. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 346–352 (2023)
29. Mahmood, A., Wang, J., Yao, B., Wang, D., Huang, C.M.: Llm-powered conversational voice assistants: Interaction patterns, opportunities, challenges, and design guidelines. *arXiv preprint arXiv:2309.13879* (2023)
30. Martínez, G., Hernández, J.A., Conde, J., Reviriego, P., Merino-Gómez, E.: Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology* (2024)
31. Ooge, J., Verbert, K.: Explaining artificial intelligence with tailored interactive visualisations. In: *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces*. pp. 120–123 (2022)
32. Park, H., Nam, Y., Kim, J.H., Choo, J.: Hypertendril: Visual analytics for user-driven hyperparameter optimization of deep neural networks. *IEEE Transactions on Visualization and Computer Graphics* **27**(2), 1407–1416 (2020)

33. Requeima, J., Bronskill, J., Choi, D., Turner, R.E., Duvenaud, D.: Llm processes: Numerical predictive distributions conditioned on natural language. arXiv preprint arXiv:2405.12856 (2024)
34. Sacha, D., Sedlmair, M., Zhang, L., Lee, J.A., Peltonen, J., Weiskopf, D., North, S.C., Keim, D.A.: What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* **268**, 164–175 (2017)
35. Spinner, T., Schlegel, U., Schäfer, H., El-Assady, M.: explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics* **26**(1), 1064–1074 (2019)
36. Tan, Y., Zhang, Z., Li, M., Pan, F., Duan, H., Huang, Z., Deng, H., Yu, Z., Yang, C., Shen, G., et al.: Medchatzh: A tuning llm for traditional chinese medicine consultations. *Computers in Biology and Medicine* **172**, 108290 (2024)
37. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
38. Wang, B., Li, G., Li, Y.: Enabling conversational interaction with mobile ui using large language models. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. pp. 1–17 (2023)
39. Wang, D., Weisz, J.D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., Gray, A.: Human-ai collaboration in data science: Exploring data scientists’ perceptions of automated ai. *Proceedings of the ACM on human-computer interaction* **3**(CSCW), 1–24 (2019)
40. Wilhelm, T.I., Roos, J., Kaczmarczyk, R.: Large language models for therapy recommendations across 3 clinical specialties: comparative study. *Journal of medical Internet research* **25**, e49324 (2023)
41. Xu, Z., Wall, E.: Exploring the capability of llms in performing low-level visual analytic tasks on svg data visualizations. arXiv preprint arXiv:2404.19097 (2024)
42. Yi, J.S., Kang, Y.a., Stasko, J.T., Jacko, J.A.: Understanding and characterizing insights: how do people gain insights using information visualization? In: *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaluation methods for Information Visualization*. pp. 1–6 (2008)
43. Yuan, J., Chen, C., Yang, W., Liu, M., Xia, J., Liu, S.: A survey of visual analytics techniques for machine learning. *Computational Visual Media* **7**, 3–36 (2021)
44. Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B.: Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics* **12**, 39–57 (2024)