


SHORT REPORT

Artificial Intelligence

Evaluating the performance of large language models in haematopoietic stem cell transplantation decision-making

Ivan Civettini^{1,2}  | Arianna Zappaterra^{1,2,3} | Bianca Maria Granelli^{1,2} | Giovanni Rindone^{1,2} | Andrea Aroldi² | Stefano Bonfanti^{1,2} | Federica Colombo^{1,2} | Marilena Fedele² | Giovanni Grillo³ | Matteo Parma² | Paola Perfetti² | Elisabetta Terruzzi² | Carlo Gambacorti-Passerini^{1,2} | Daniele Ramazzotti¹ | Fabrizio Cavalca²

¹Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

²Department of Haematology and Bone Marrow Transplantation Unit, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy

³Department of Haematology and Bone Marrow Transplantation Unit, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy

Correspondence

Ivan Civettini, Haematology Division, University of Milano-Bicocca, Fondazione IRCCS San Gerardo dei Tintori Hospital Monza, via Cadore 48, Monza 20900, Italy. Email: i.civettini@campus.unimib.it; ivan.civettini@gmail.com

Summary

In a first-of-its-kind study, we assessed the capabilities of large language models (LLMs) in making complex decisions in haematopoietic stem cell transplantation. The evaluation was conducted not only for Generative Pre-trained Transformer 4 (GPT-4) but also conducted on other artificial intelligence models: PaLM 2 and Llama-2. Using detailed haematological histories that include both clinical, molecular and donor data, we conducted a triple-blind survey to compare LLMs to haematology residents. We found that residents significantly outperformed LLMs ($p=0.02$), particularly in transplant eligibility assessment ($p=0.01$). Our triple-blind methodology aimed to mitigate potential biases in evaluating LLMs and revealed both their promise and limitations in deciphering complex haematological clinical scenarios.

KEY WORDS

artificial intelligence, GPT, HSC transplantation, interrater agreement, transplant

INTRODUCTION

Large language models (LLMs) are a type of artificial intelligence (AI) that employs deep learning techniques and big datasets to understand, summarize and predict new content. By analysing data and identifying patterns and connections, they can predict the most likely words or phrases in specific contexts. Previous studies have indicated that Generative Pre-trained Transformer (GPT) developed by OpenAI performs well in answering single-choice clinical questions. However, its performance seems to be less satisfactory when dealing with multiple-choice questions and more intricate clinical cases.^{1,2} A marked enhancement is noted with Flan-PaLM, registering a commendable 67.6% precision on MedQA.³⁻⁵ A new model named Med-PaLM

was recently published in Nature,⁶ but, despite its promising results, it still lags behind the discernment of seasoned clinicians.

A critical gap in the current literature is the exploration of LLMs within the ambit of haematopoietic stem cell transplantation (HSCT) decision-making—a multifaceted process deeply rooted in clinical acumen. Essential to this decision-making is the evaluation of candidates and risk determinants of HSCT. Key factors include disease-specific considerations and patient-specific elements (age, comorbidities and infectious diseases/colonization).⁷ Over the years, several predictive models became an essential tool to aid clinicians in gauging transplantation risks (EBMT risk score,⁸ DRI,⁹ HCT-comorbidity index⁹ and PAM score¹⁰). Additionally, donor-associated factors, encompassing stem

Daniele Ramazzotti and Fabrizio Cavalca contributed equally as senior authors to this work.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *British Journal of Haematology* published by British Society for Haematology and John Wiley & Sons Ltd.

cell sources, play a pivotal role in influencing disease control and transplant-related mortality (TRM).⁷

In this evolving landscape of growing interest in LLMs within medical research, it is pertinent to note that the lion's share of clinical research has been fixated on GPT's capabilities, often side-lining robust competitors like Llama-2 (Meta) and PaLM 2 (Google). Considering these observations, our research aimed to rigorously evaluate the potential of LLMs within the intricate domain of HSCT.

METHODS

Patient histories

Six haematological patient histories were created. An experienced haematologist subsequently reviewed and validated these records. The data spanned demographics, prior medical histories, haematological disease characteristics (including genetic data and minimal residual disease [MRD]), treatment outcomes, complications from prior therapies and pertinent donor details (related/unrelated, HLA and CMV). From a haematological standpoint, the patient histories comprised four cases of acute myeloid leukaemia, one of acute lymphoblastic leukaemia and one of myelofibrosis (detailed medical histories in Data S1).

Expert and resident review

Six bone marrow transplant specialists from two leading JACIE-accredited hospitals, along with 11 haematology residents from the University of Milano-Bicocca, were presented with the clinical histories. The following questions were posed:

1. Would you recommend a transplant for this patient?
2. Which donor, from the provided list, would you select?
3. What would be your preferred conditioning regimen?
4. Could you estimate the transplant-related mortality (TRM)?

The potential answers included:

1. Yes, No, I don't know
2. Specified donor choice or I don't know
3. Several conditioning regimens, including 'thiotepa-fludarabine-busulfan myeloablative (TBF-MAC)', 'thiotepa-fludarabine-busulfan reduced intensity conditioning (TBF-RIC)', 'treosulfan-fludarabine (Treo-FLU)', 'cyclophosphamide-total body irradiation (C-TBI)', 'busulfan-fludarabine (BU-FLU)', or 'I don't know'
4. Risk classifications: 'Low (14% at 2 years)', 'Intermediate (21% at 2 years)', 'High (41% at 2 years)' or 'I don't know'

Notably, those opting for 'I don't know' for the first question (pertaining to HSCT eligibility) could not proceed with subsequent answers for that specific patient.

LLM analysis

GPT-4, PaLM 2, Llama-2 13b and 70b were the chosen LLMs for analysis.^{11–15} Except for GPT-4, which retained default temperature setting, the LLMs underwent calibration with variable temperature settings. The degree of unpredictability in a language model's output is determined by its temperature setting. Higher temperature settings increase the likelihood of less likely tokens while decreasing the likelihood of more likely ones, making outputs less predictable and more creative. Lower temperatures, on the other hand, produce outcomes that are more conservative and reliable. We favoured lower settings to foster more deterministic outputs. Initially, LLMs were tasked with direct answers to the key questions. Later, we expanded the token limit to explore the reasoning behind their selections (detailed LLM settings and methods are provided in [supplements](#)).

Data collection and analysis

A triple-blind survey was conducted using Typeform (<https://www.typeform.com/>), where both senior haematologists and residents submitted anonymized responses via unique tokens. Critically, each group (senior haematologists, residents and LLM testers) remained uninformed about the others' responses. The consensus answer (CoA) was defined as the predominant response from the experts. To evaluate the agreement between residents or LLMs and the CoA of experts, we computed Fleiss kappa (K) and overall agreement (OA). Responses were treated as nominal dichotomous variables, categorized as either matching or differing from the CoA of experts. OA denoted the percentage of agreement between residents or LLMs and the consensus answer provided by the experts. This indicated the proportion in which residents or LLMs selected the CoA among all possible answers. Fleiss kappa, in contrast to OA, not only assessed the percentage of agreement but also considered the probability of chance alone in selecting the answers.^{16,17} Given the ununiformed use of cut-off values for K ,¹⁸ and considering the complex nature of the transplant scenario, we compared the groups by evaluating OA and K values as continuous variables. The comparison between the groups was executed using T -tests or Mann-Whitney tests, with GraphPad version 10.0.1. For detailed Methods, see [Figure 1](#).

RESULTS

Expert consensus

The expert consensus showed a perfect agreement in evaluating patient transplant eligibility ($K=1.0$, OA 100.0%). They also demonstrated substantial consensus when choosing donors ($K=0.62$, OA 66.7%) and deciding on conditioning

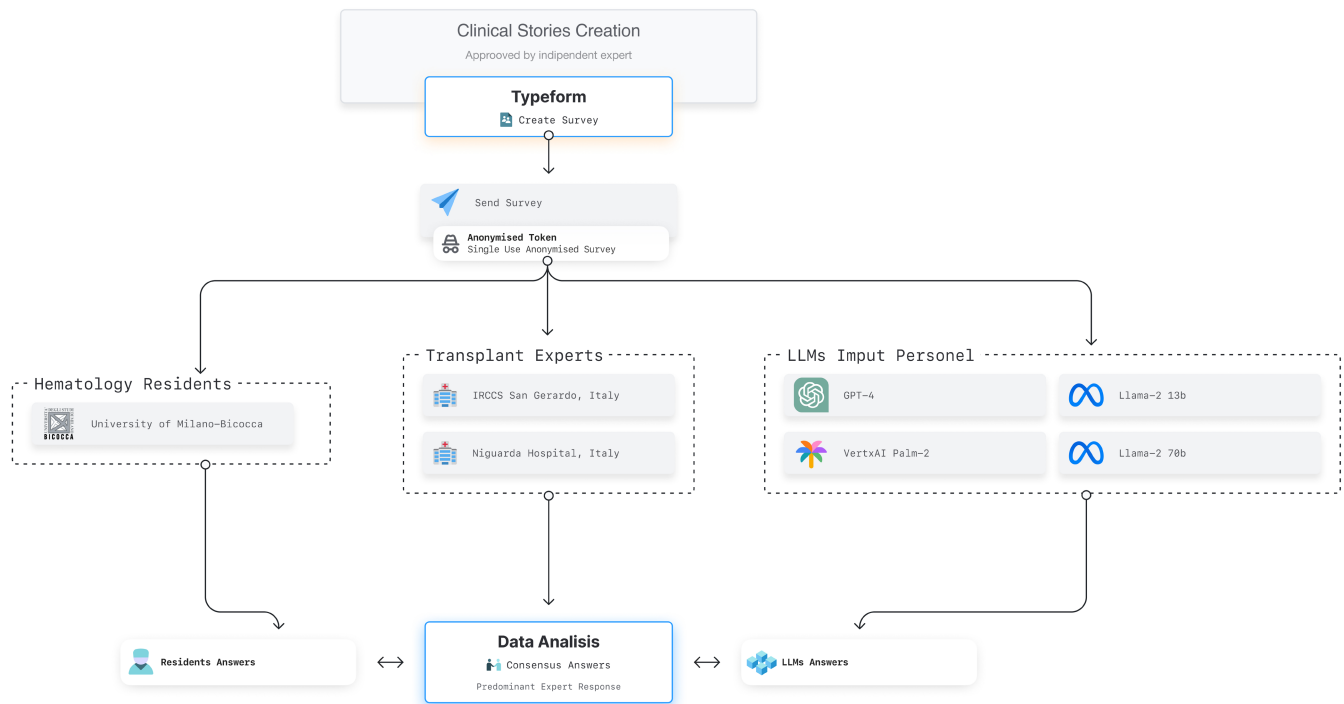


FIGURE 1 Flow chart outlining the multistep methodology employed in the study to compare the performance of LLMs with that of haematology residents. Underlined rectangle indicates key phases like clinical stories creation and validation, survey creation and data analysis. Dotted rectangle indicates the macro-areas where the survey was sent, which includes haematology residents, transplant experts and large language model (LLM) input personnel. Data analysis was conducted by comparing the answers of LLMs and residents with the consensus answers of experts, defined as the most frequent answers provided by experts. This figure was created using standard UI Kit.

regimens ($K=0.62$, OA 68.3%). However, the agreement waned somewhat in the TRM estimation, achieving only fair consensus ($K=0.22$, OA 41.7%).

LLMs versus resident responses

Of the 150 questions evaluated in the LLMs statistical analysis, LLMs responded with 'I don't know' in only 0.03% (five questions) of the instances. On the other hand, out of the 198 questions posed to residents, they opted for 'I don't know' in 12.1% (24 questions).

The median OA and K values between residents and the CoA of experts were 76.5% (range 52.9%–88.2%) and 0.61 (range 0.4–0.8) respectively. The median OA and K values between LLMs answers and experts were 58.8% (range 47%–71%) and 0.45 (range 0.3–0.61) respectively. The mean OA and K values of residents were significantly higher compared to LLMs ($p=0.02$). Specifically, residents showed higher median OA and K values in patient eligibility assessment (median OA 100% vs. 83% and K 1 vs. 0.78; $p=0.01$). However, there were no significant differences in median K or OA between residents and LLMs for donor choice (0.56 vs. 0.56; OA 67% vs. 67%; $p=0.3$), conditioning regimen (0.67 vs. 0.33; OA 75% vs. 50%; $p=0.6$) and TRM evaluation (0.33 vs. 0; OA 50% vs. 25%; $p=0.1$) (Table S1). The median K values of GPT-4, PaLM 2, Llama2-13b and Llama2-70b were 0.49 (OA 61.7%),

0.53 (OA 64.7%), 0.33 (OA 50%) and 0.53 respectively (OA 64.7%) (Figure 2).

LLM limitations, a deeper insight

Upon a closer look at the data, LLMs exhibited notable shortcomings in two specific questions:

1. Patient 3 HSCT indication. A case of favourable risk MRD negative AML following one induction and three consolidation chemotherapy rounds—all LLMs invariably endorsed HSCT. This starkly contrasted with the residents' unanimous decision against HSCT for this patient.
2. Patient 4 conditioning regimen—lymphoblastic leukaemia. Experts predominantly opted for C-TBI. In 89% of the cases, LLMs suggested the TBF-MAC and never C-TBI. Residents achieved a mean OA of 36.3% for this question.

CONCLUSION

Our study provides a comprehensive assessment of the role and capabilities of LLMs in the intricate domain of HSCT decision-making. Although LLMs showed promising results with a median OA of 59%, residents demonstrated

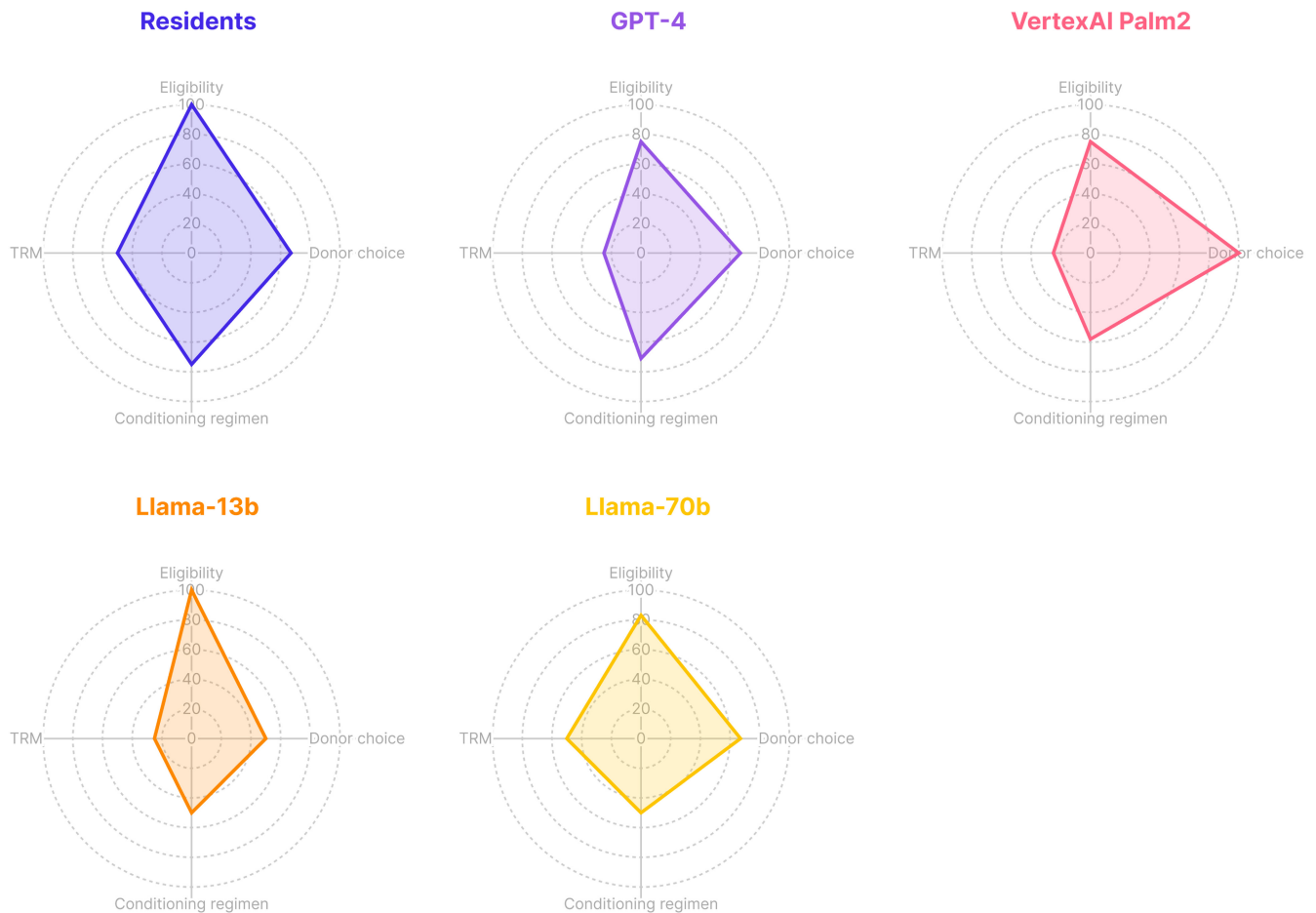


FIGURE 2 Comparative analysis of median overall agreement (OA) with radar charts. Displayed as a percentage ranged from 0 to 100, across various subclasses including transplant eligibility, donor choice, conditioning regimen choice and TRM (transplant-related mortality). Figures were created with <https://flourish.studio>.

superior performance. It is particularly worth noting that the residents involved in our study ranged from the first year of residency to the last. This means that many of them, especially the younger ones, might not have had any direct transplant experience yet. The main differences between residents and LLMs were most evident in the HSCT indication for the third patient. A notable discrepancy also arose with the second patient, an elderly unfit patient, where certain LLMs (GPT and Llama) showed uncertainty of responses in HSCT indication ('I don't know') and Vertex AI PaLM 2 set at a temperature of 0.2 proposed the HSCT, contrary to the unanimous expert and resident opinion (no HSCT eligibility). It is worth noting that simply by lowering the temperature at 0 and by raising tokens to better understand the motivation beyond the HSCT choice, Vertex AI PaLM 2 gave a complete and correct response. In this case, PaLM did not consider the second patient for HSCT, given the patient age (78 years old), comorbidities and *Klebsiella pneumoniae* carbapenemase (KPC)-producing bacteria colonization, and at least available donor (HLA match 8/10). PaLM therefore recommended continuing with hypomethylating therapy and supportive care for this patient.

We then asked LLMs about the specific documents or web sources that contributed the decision within the training dataset. The PaLM model's response indicates that the decisions were influenced by UpToDate information on stem cell transplantation in acute myeloid leukaemia as well as articles published by the European Society for Blood and Marrow Transplantation (EBMT) between 2017 and 2021. It is interesting to note that PaLM consistently cited these reputable sources when asked about acute myeloid leukaemia as well as acute lymphoblastic leukaemia. This may explain why LLMs consistently suggested a myeloid conditioning regimen and never C-TBI for the fourth patient with acute lymphoblastic leukaemia. Given that LLMs are not specifically designed for this purpose, interpreting this response requires careful consideration. Conversely, the same query posed to Llama and GPT models resulted in a response indicating their inability to recall the precise documents or web-pages present in the training dataset.

LLMs displayed good performances also in donor choice but showed shortcomings in conditioning regimens and TRM evaluation.

Previous studies have evaluated the performance of LLMs such as GPT by asking experts to use a rating scale for

assessment.¹⁹ While this method does not account for the concordance among experts, it also introduces the possibility of bias, especially if the experts are aware of the responses provided by the LLMs.

To mitigate the risk of bias, experts did not use a rating scale when evaluating LLMs' responses. However, it is crucial to acknowledge that the consensus answer, although the most frequent, does not automatically imply that other responses provided by the experts were incorrect. Therefore, the lower consensus among experts in TRM evaluation, likely due to the challenge of precisely calculating TRM in a survey-based evaluation, should also prompt a cautious approach when evaluating residents' and LLMs' answers in this context.

Finally, we must underline that, despite the effort to enrich the clinical cases with clinical and molecular data to closely simulate real-world scenarios, and subsequent validation and approval by an independent haematologist, these cases might not fully capture the complexity and varied nature of actual cases. The choice to avoid employing real-world scenarios primarily emerged during the early stages of the study in early 2023, driven by concerns initially raised by the Italian Data Protection Authority regarding the use of LLMs.²⁰

In conclusion, our research underscores the indispensable value of human expertise in HSCT decisions. While LLMs did not outperform haematology residents in this complex context, their results are promising. LLMs, with potential further refinements using specialized haematological datasets, could become a supplementary tool that can assist clinicians in intricate HSCT assessments in the future.

AUTHOR CONTRIBUTIONS

Ivan Civettini: conceptualization, writing—original draft, performed experiments and statistical analysis. **Arianna Zappaterra, Bianca Maria Granelli, Giovanni Rindone, Stefano Bonfanti and Federica Colombo:** data collection, writing and editing. **Andrea Aroldi, Marilena Fedele, Giovanni Grillo, Matteo Parma, Paola Perfetti and Elisabetta Terruzzi:** proofreading, transplant setting supervision. **Carlo Gambacorti-Passerini:** proofreading, scientific and haematological supervision. **Daniele Ramazzotti:** conceptualization, LLM supervision, writing and editing. **Fabrizio Cavalca:** conceptualization, clinical supervision, writing and editing.

FUNDING INFORMATION

The authors received no financial support for the research, authorship and publication of this article.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author, I.C., upon request.

ETHICS STATEMENT

Ethics Committee approval was not necessary.

PATIENT CONSENT STATEMENT

Patient consent was not necessary.

PERMISSION TO REPRODUCE MATERIAL FROM OTHER SOURCES

We confirm that all materials used in this article, including the Standard UI Kit used in [Figure 1](#), were properly licensed and authorized for use. The Standard UI Kit used in this article is licensed under the Email address i.civettini@campus.unimib.it.

CLINICAL TRIAL REGISTRATION NUMBER

This study does not involve a clinical trial, and therefore, it is not registered in a clinical trial database.

ORCID

Ivan Civettini  <https://orcid.org/0000-0002-7707-584X>

REFERENCES

1. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Med Educ.* 2023;9:e46599.
2. Hoch CC, Wollenberg B, Lüers J-C, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol.* 2023;280(9):4271–8.
3. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arxiv.org* Cornell University 2020.
4. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. *arXiv.* 2019.
5. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. *arXiv.* 2020.
6. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172–80.
7. Carreras E, Dufour C, Mohty M, Kröger N, editors. *The EBMT handbook: hematopoietic stem cell transplantation and cellular therapies.* 7th ed. Cham (CH): Springer; 2019.
8. Terwey TH, Hemmati PG, Martus P, Dietz E, Vuong LG, Massenkeil G, et al. A modified EBMT risk score and the hematopoietic cell transplantation-specific comorbidity index for pre-transplant risk assessment in adult acute lymphoblastic leukemia. *Haematologica.* 2010;95(5):810–8.
9. Sorror ML. Comorbidities and hematopoietic cell transplantation outcomes. *Hematology.* 2010;2010(1):237–47.
10. Parimon T, Au DH, Martin PJ, Chien JW. A risk score for mortality after allogeneic hematopoietic cell transplantation. *Ann Intern Med.* 2006;144(6):407–14.
11. Available from: <https://openai.com/>. Accessed 13 Oct 2023.
12. Available from: <https://cloud.google.com/vertex-ai/docs/generative-ai/start/quickstarts/api-quickstart>. Accessed 13 Oct 2023.
13. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. *arXiv.* 2307.09288. <https://doi.org/10.48550/arXiv.2307.09288>

14. Mahan D, Carlow R, Castricato L, Cooper N, Laforte C. Stable Beluga models. Available from: <https://huggingface.co/stabilityai/StableBeluga2>. Accessed 13 Oct 2023.
15. Available from: <https://huggingface.co/TheBloke/Llama-2-13B-chat-GGML>. Accessed 13 Oct 2023.
16. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
17. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5):378–82. <https://doi.org/10.1037/h0031619>
18. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;85(3):257–68.
19. Haemmerli J, Sveikata L, Nouri A, May A, Egervari K, Freyschlag C, et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? *BMJ Health Care Inform.* 2023;30(1):e100775.
20. Available from: <https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847>. Accessed 13 Oct 2023.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Civettini I, Zappaterra A, Granelli BM, Rindone G, Aroldi A, Bonfanti S, et al. Evaluating the performance of large language models in haematopoietic stem cell transplantation decision-making. *Br J Haematol.* 2023;00:1–6. <https://doi.org/10.1111/bjh.19200>