Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

BankGPT: the use of large language models in official communications[1]

Claudia Biancotti, Carolina Camassa, Marco Fruzzetti,
Luigi Palumbo and Myriam Portaluri,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

BankGPT: the use of large language models in official communications

Claudia Biancotti, Carolina Camassa, Marco Fruzzetti, Luigi Palumbo, Myriam Portaluri

Abstract

This study investigates the summarization performance of two prominent large language models (LLMs) – *GPT-3.5 turbo* and *LLaMA 2* – by evaluating their ability to generate summaries of the Bank of Italy's *Economic Bulletin*, a standardized institutional publication. The analysis compares model-generated summaries with the official human-produced summary, using evaluations from domain experts within the Bank's Directorate General for Economics, Statistics and Research. Employing the Bradley-Terry model for pairwise comparisons, the study assesses summaries across three criteria: fluency, relevance, and consistency. Results indicate no overall preference between human and machine-generated summaries; however, experts with greater familiarity with the *Bulletin* and higher educational attainment showed a marked preference for the official summaries. In addition, preferences correlated with demographic factors, with women and individuals over 35 more likely to prefer human-authored summaries. These findings highlight both the strengths and limitations of LLMs for tasks requiring specialized domain knowledge and audience alignment.

Keywords: Large Language Model (LLM), central banks, Bradley–Terry model, logistic regression.

JEL classification: O39, Z19, A19, C88, H89

1. Introduction

Large language models (LLMs) are artificial intelligence systems trained to understand the syntax and semantics of human language, making them appear well-suited for summarization tasks. These models can process large volumes of text and generate grammatically correct summaries at high speed.

This study assesses the summarization capabilities of two widely used LLMs - *GPT-3.5 turbo* and *LLama 2* - by analyzing their ability to summarize central bank publications, with a particular focus on the Bank of Italy's *Economic Bulletin*, an institutional report with a format and editorial structure that are repeated in every edition. The *Economic Bulletin* is meticulously drafted by senior economists and its English language version undergoes final revisions by professional English editors and translators to ensure clarity and accessibility for the general public.

To evaluate the summaries, we conducted a survey with domain experts from the Directorate General for Economics, Statistics and Research, comparing three types of summaries: the official summary from the *Economic Bulletin* and those generated by *GPT-3.5 turbo* and *Llama 2*. The evaluation was based on pairwise comparisons using three key criteria: fluency, consistency, and relevance. Following the definitions in Fabbri et al. (2021), fluency measures both grammatical correctness and stylistic coherence with the original text. Relevance assesses the summary's efficacy in distilling essential information from the source text. Finally, consistency examines factual accuracy, ensuring the absence of imprecisions or fabricated content.

We used the Bradley-Terry model for paired comparisons (Bradley and Terry 1952) to analyze the results, which – although limited by the small, non-random sample – are consistent with established patterns (T. Zhang et al. 2024). Overall, evaluators showed no clear preference for human and machine-generated summaries. However, evaluators with a deeper knowledge of the *Economic Bulletin* and higher education levels tended to favor human-generated summaries, indicating that LLMs may not yet fully capture specialized knowledge.

Our findings also suggest a preference for human summaries among women and individuals over 35. This pattern may reflect how preference training and fine-tuning processes shape LLM outputs in a way that aligns with different demographic groups' content preferences (Kirk et al. 2024; Kotek, Dockum, and Sun 2023; Pervez and Titus 2024).

2. Literature Review

Large language models (LLMs) are taking the world by storm. In 2018 OpenAI developed the initial Generative Pre-trained Transformers (GPT) model, a state-of-the-art language model, capable of understanding and replicating patterns in human language with good accuracy. Since September 2022, when OpenAI released "Chat Generative Pre-Trained Transformer", commonly known as "ChatGPT", the significant potential of these AI tools was clear to the public. Subsequently, more proprietary and open LLMs have been released over time with ever-increasing capabilities (OpenAI et al. 2023; Gemini Team et al. 2023; Touvron et al. 2023).

Large language models (LLMs) represent state-of-the-art artificial intelligence systems trained on extensive textual corpora to generate contextually relevant, human-like text. These models demonstrate remarkable capabilities in producing coherent and grammatically sophisticated content with high computational efficiency. Their applications span diverse tasks including text correction, writing assistance, and content summarization. In October 2023, an OECD survey of various institutions (Mitton 2023) revealed that while 23% of respondents reported using Generative AI (GenAI) technologies in their work, 76% lacked clear policies and guidelines for its use in communication tasks.

LLMs have found applications in different domains, including finance. Fatouros et al. (2023) investigated the capabilities of large language models in financial sentiment analysis, in the context of the foreign exchange market, and demonstrated their significant effectiveness in understanding market behavior. Further studies showed the potential of LLMs in generating investment signals given a set of information (Fatouros et al. 2024).

In the realm of central banking, Hansen and Kazinnik (2023) assessed whether ChatGPT can decipher FOMC announcements. Smales (2023) tested whether ChatGPT can classify hawkish/dovish monetary policy decisions made by the Reserve Bank of Australia (RBA). Gambacorta et al. (2024) introduced central bank language models (CB-LMs) designed for two primary purposes: (i) predicting masked words within central bank-specific terminology and (ii) classifying the monetary policy stance conveyed in Federal Open Market Committee (FOMC) statements. The authors advocated advancing Natural Language Process analysis in monetary economics and central banking, emphasizing the importance of knowledge sharing among central banks to enhance the application of language models in the processes underlying monetary policy decision-making. However, they also highlighted critical concerns regarding confidentiality, transparency, replicability, and cost-efficiency associated with integrating LLMs into central banking workflows.

The use of LLMs in text summarization is a widely studied topic, we refer to Zhang, Yu, and Zhang (2024) for a comprehensive review. Recent studies found that performance can be comparable to expert humans on generic news, although with consistent stylistic differences (T. Zhang et al. 2024). However, the evaluation of LLM-Generated summaries is a task that remains methodologically challenging, and current approaches remain incomplete (Liu et al. 2023; Schaik and Pugh 2024).

The evaluation of AI-generated content quality and the detection of AI-authored content are active areas of research. Recent studies on poetry found that non-expert readers were not able to identify AI-generated poems, and those were rated more favorably than human-authored ones (Porter and Machery 2024). Another study on artworks found that participants were unable to consistently distinguish between human and AI-created images, and they generally preferred AI-generated artworks over human-made ones (Grassini and Koivisto 2024). Additionally, the same study showed that participants displayed a negative bias against artworks that were perceived as AI-generated, independently on their true source.

3. Experiment design

The objective of this study is to assess whether a professional audience demonstrates a clear preference for human-generated summaries over those produced by LLMs in the context of central bank communications. As a case study, we analyze the quarterly *Economic Bulletin* of the Bank of Italy, a highly standardized institutional publication. This *Bulletin* offers a comprehensive overview of economic developments in Italy, contextualized within the broader international and euro-area context. It focuses on key aspects of the economy, including the real economy, inflation, banking, and the financial markets. Each section has an introductory summary, edited by the editorial committee and professional reviewers at the Library Division of Bank of Italy.

The *Economic Bulletin* is originally written in Italian and undergoes a multi-step translation process before publication. First, an institutional automated tool generates an initial translation. This version is then reviewed and refined by a professional translator before undergoing a final validation by editors to ensure accuracy and adherence to the publication's standards.

For our analysis, we randomly selected five out of the 13 sections from the April 2023 issue of the *Economic Bulletin*[1]. The selected sections included: (1.2) *The World Economy – The Euro Area*, (2.2) *The Italian Economy – Firms*, (2.6) *The Italian Economy – Price Developments in Italy*, (2.7) *The Italian Economy – Banks*, and (2.8) *The Italian Economy – Financial Markets*. We extracted the original text from each section and generated summaries using GPT-3.5 Turbo and LLaMA 2 70B. To ensure comparability with the original text, we designed the prompts to produce summaries with a word count approximately matching that of the human-authored introductory summaries at the beginning of each section.

In order to evaluate the preferences of a professional audience, we conducted a survey (full text available in Appendix B) among all employees in the Directorate General for Economics, Statistics, and Research at the Bank of Italy, comprising approximately 600 professionals. A number of divisions in the directorate are involved, in different capacities, in writing and editing of the *Economic Bulletin*. The survey was designed and distributed via an internal platform, which also recorded completion time for each survey response. Participants received a series of emails containing a link to the survey, facilitating access and encouraging participation.

The survey gathered demographic information about respondents, including gender, age, level of English proficiency and educational credentials including degree level and field. Additionally, it collected details on professional background, such as familiarity with LLMs, years of work experience, whether respondents had ever contributed to the authorship of the Bank of Italy's *Economic Bulletin*, and how often they read the publication.

For each selected section, we conducted a randomized pairwise comparison of two summaries chosen from three possible sources: the official human-generated introductory summary, GPT-3.5 Turbo, or LLaMA 2. To minimize potential judgment biases, the survey characterized all summaries as generated by LLMs. Participants were asked to evaluate their preferences for each pair of summaries based on three

[1] https://www.bancaditalia.it/pubblicazioni/bollettino-economico/2023-2/index.html

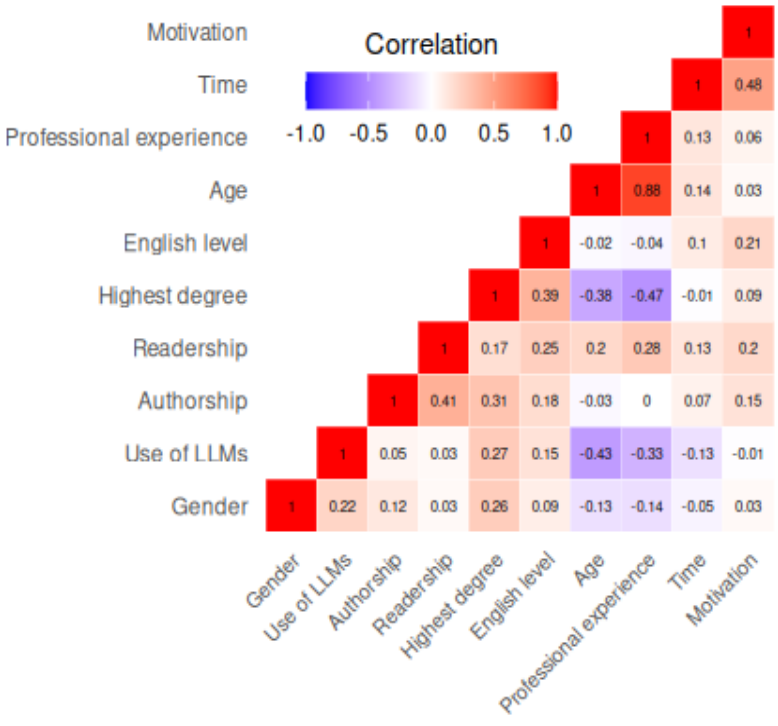
metrics - Fluency, Consistency, and Relevance - as defined by Fabbri et al. (2021). Preferences were recorded using a five-point Likert scale: Strong Preference for Text A, Preference for Text A, Neutral, Preference for Text B, and Strong Preference for Text B. Respondents were also given the opportunity to provide written justifications for their evaluations.

4. Data description

The survey was conducted in February 2024, yielding 175 responses, representing approximately 30% of the target population. Of these respondents, 133 completed evaluations for all randomly proposed text pairs, while 42 participants provided assessments for at least one text pair. Nine responses contained incomplete demographic information, which was subsequently imputed using the Random Forest algorithm provided by the mice package in R (Buuren and Groothuis-Oudshoorn 2011). Analyses performed both with and without imputed data yielded consistent results. Table 4 in Appendix A reports the summary figures for the covariates.

We conducted a preliminary analysis of correlations across covariates, with variables transformed into ordinal levels. The results are visualized in Figure 1. As expected, age and professional experience showed strong positive correlation. We observed a negative correlation between LLM usage and age, and a positive correlation with educational attainment. Finally, there is a positive correlation between the time spent for answering the survey and having left at least a written motivation for the preferences indicated between different texts.

Covariates correlation plot — Figure 1



We computed the mean preference scores for each metric and section using a weighted scoring system. This system allocates +1 point to the preferred text and -1 point to the non-preferred text, with strong preferences weighted at±2 points and neutral responses at zero points. As shown in Figure 2, preferences varied across writing samples: human-generated text was significantly preferred in sections 1.2 (*The World Economy – The Euro Area*) and 2.7 (*The Italian Economy – Banks*), whereas LLM-generated text performed better in sections 2.6 (*The Italian Economy – Price Developments in Italy*) and 2.8 (*The Italian Economy – Financial Markets*). Section 2.2 of the *Bulletin* (*The Italian Economy – Firms*) yielded heterogeneous results across different metrics, with no clear preference pattern emerging.

Mean preference scores

Figure 2

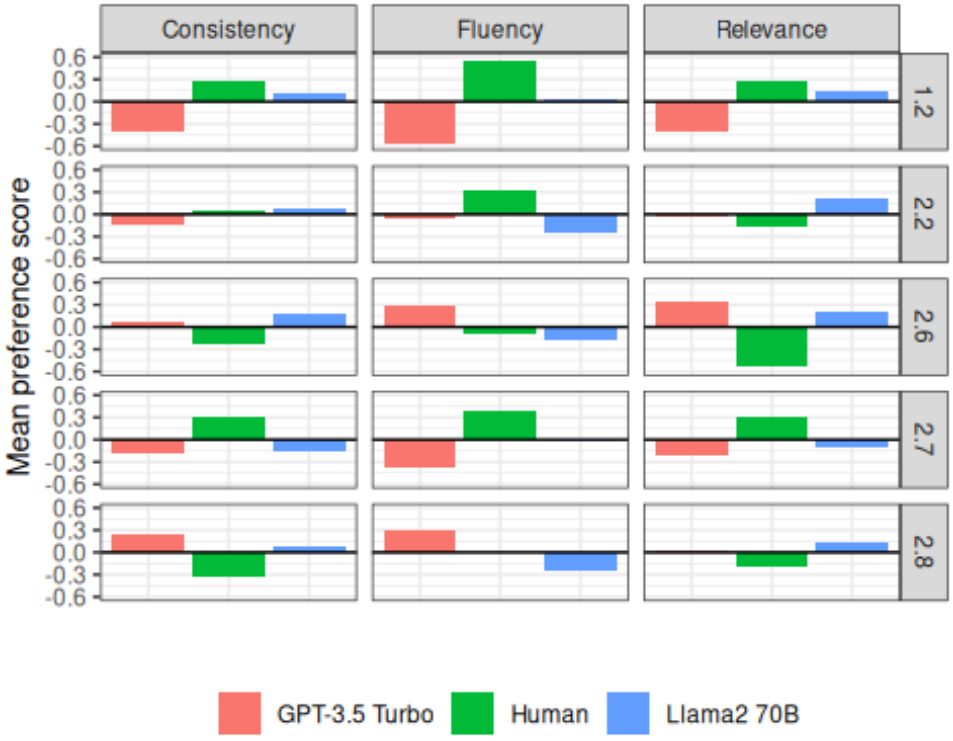
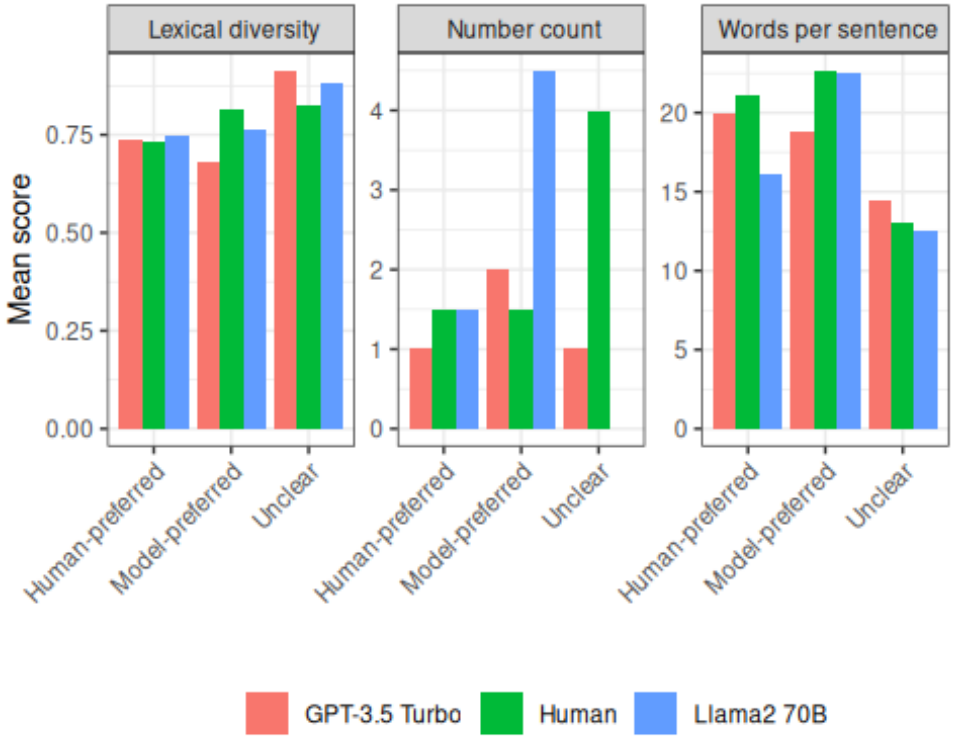


Figure 3 presents a set of descriptive metrics for each source and summary, including the average number of words per sentence, the frequency of numerical values in each summary, and lexical complexity, which is defined as to the ratio of different unique word stems to the total number of words. Based on the preferences outlined above, we classify sections 1.2 and 2.7 as "Human preferred," section 2.2 as "Unclear," and the remaining sections as "Model preferred." The emerging pattern shows that the favored summaries in human-preferred sections have longer sentences, and lower number counts, while model-preferred sections show higher number counts and shorter sentences. This pattern aligns with the different nature of these sections - model-preferred sections cover data-heavy topics like prices and financial markets, while human-preferred sections handle more complex, context-rich topics like macroeconomic analysis and banking sector developments.

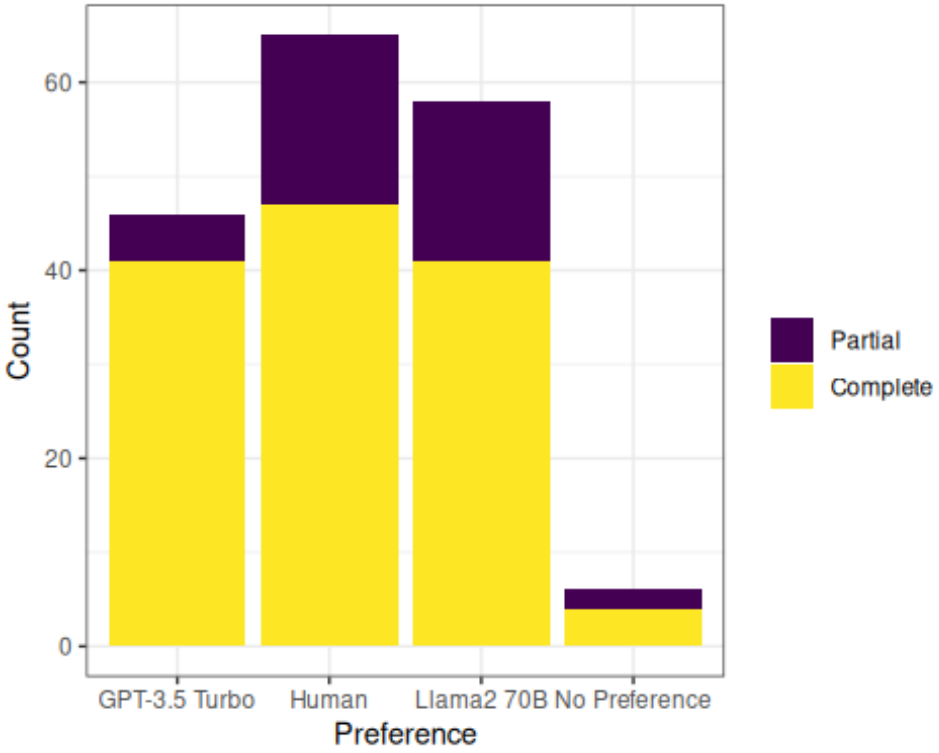
Text metrics

Figure 3



We chose to use The Bradley-Terry model (Bradley and Terry 1952) to convert the pairwise comparisons of the summaries into a comprehensive ranking of preferences for each respondent (see Appendix C for the formal specification and estimation procedure). This model is a popular choice to process the pairwise preference data, and it is commonly used to derive LLM rankings from conversational data (Chiang et al 2024). The results, shown in Figure 4, indicate that human-generated summaries are the most preferred overall, particularly among respondents who completed the survey. GPT-3.5 Turbo and LLaMA2 70B are competitive, with similar aggregate scores, while "No Preference" ranks lowest, indicating a strong tendency for respondents to choose a clear favorite.

Individual preferences for summary sources, derived from Bradley-Terry scores Figure 4



5. Main results

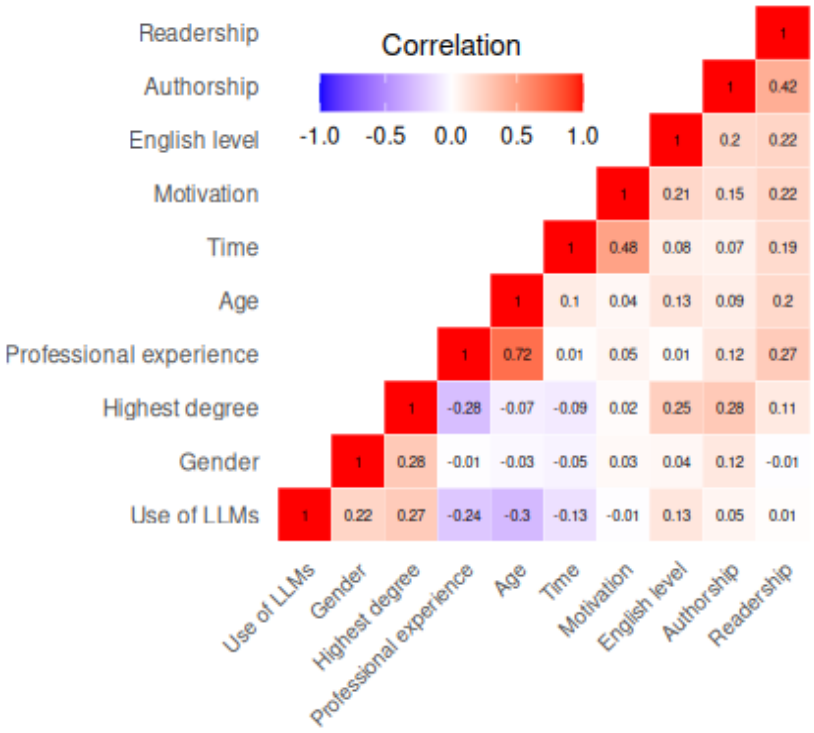
We employed a logistic regression model to explore whether specific respondent characteristics influence their preference for human-generated summaries over those created by LLMs. In this analysis the dependent variable is the Bradley-Terry calculated preference for human-generated text.

In order to streamline the model and reduce dimensionality, we recoded the covariates into binary variables. Age was categorized into two groups: "Under35" and "Over35." Professional experience was dichotomized into "Junior" (10 or fewer years of professional experience) and "Senior." English proficiency levels of "Elementary" and "Intermediate" were combined into a single category, as were all educational levels below the doctoral level. Additionally, readership intensity of the *Economic Bulletin* was simplified into two groups: "Low" for respondents who indicated reading the publication "Sometimes" or less frequently, and "High" for those reporting the two highest levels of readership intensity. "Motivation" indicates whether the respondent provided at least a written motivation for the preferences assigned.

This approach ensures a parsimonious model while retaining the ability to identify key patterns and associations in the data. Figure 5 presents the correlation matrix for the dichotomized variables.

Dichotomized covariates correlation plot

Figure 5



The logistic regression analysis using the full set of covariates, with preference for human-generated text as the dependent variable, is reported as (1) in Table 1. Since most variables were not statistically significant, we implemented a stepwise backward selection procedure using Akaike Information Criterion (AIC) (Akaike 1969, 1971) to identify the most parsimonious model. We present this reduced model as (2) in Table 1.

Logistic regression on Bradley-Terry preferences for human-generated text Table 1

	(1)	(2)
Intercept	-2.064 **	-2.175 ***
	(0.638)	(0.528)
Age: Over35	1.255 *	1.117 **
	(0.627)	(0.429)
Gender: Male	-1.008 *	-0.968 *
	(0.395)	(0.389)
Authorship: Yes	0.793 *	0.797 *
	(0.397)	(0.360)
Motivation: Yes	0.952 *	0.858 *
	(0.414)	(0.352)
Highest degree: PhD	1.119 *	1.052 **
	(0.447)	(0.396)
English level: Advanced	-0.660	
	(0.479)	
Professional Experience: Senior	0.016	
	(0.557)	
Use of LLMs: Yes	0.247	
	(0.428)	
Readership: High	0.164	
	(0.406)	
Time	-0.000	
	(0.000)	
N. obs.	175	175
AIC	218.931	211.088

*** p < 0.001; ** p < 0.01; * p < 0.05.

While the model demonstrates modest predictive performance, with an F1 score of 0.42, it achieves statistically significant improvement over the no-information rate baseline.

The first takeaway is that age and gender have the largest coefficients in absolute terms: economists older than 35 years tend to prefer human text while men tend on average to prefer LLMs. The second takeaway is that having already used LLMs does not substantially influence the choice. Our findings also showed that holding a doctoral degree is positively correlated with preference for human text. Lastly, not surprisingly, authors of past *Economic Bulletin* editions favor the human text on average.

Wong and Kim (2023) demonstrated that individuals are more likely to perceive ChatGPT as male rather than female. Additionally, research on gender differences in reading preferences (Summers 2013) found that male respondents were slightly more inclined to favor books authored by men. For our study, it is relevant to mention that the translation team for the *Economic Bulletin* is mostly composed by women, and 4 out of the 5 human-generated summaries under analysis were translated by a woman.

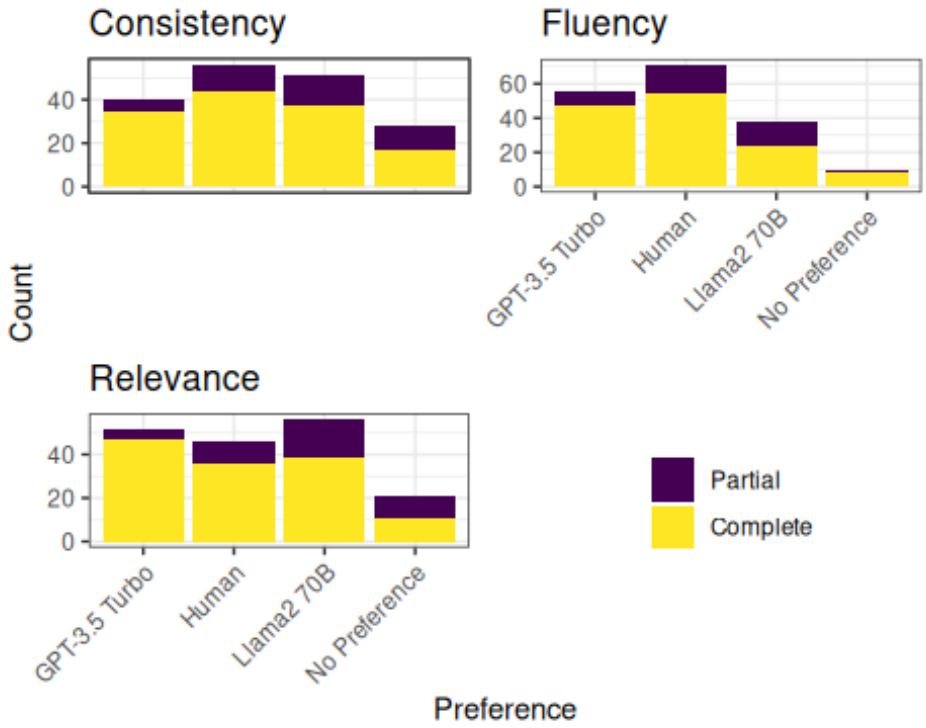
These findings suggest that the writing style of LLMs, shaped by their training and fine-tuning processes, may influence their appeal across different demographic groups.

Furthermore, having previously authored the *Economic Bulletin* and the provision of a written justification for preferences may reflect a heightened level of involvement and an emotional attachment to the publication and its distinctive writing style. Consequently, it is not surprising that these factors are associated with a stronger preference for human-generated text.

6. Robustness analysis

In this section we show the results of several robustness exercises that we performed in order to check the strength of our findings. The first exercise is the estimation of the reduced model presented as (2) in Table 1 on the three separate evaluation metrics: consistency, fluency, and relevance. Figure 6 presents the Bradley-Terry preferences for each individual metric.

Bradley-Terry preference scores for individual metrics Figure 6



Human-generated summaries seem to have an edge over LLMs in terms of consistency and fluency, while they score less favorably for relevance. Table 2 presents the results for logistic regression of preferences for human-generated text on each individual metric.

Logistic regression on individual metrics Bradley-Terry preferences for human-generated text

Table 2

	Consistency	Fluency	Relevance
Intercept	-1.772 ***	-1.647 ***	-2.707 ***
	(0.496)	(0.489)	(0.616)
Age: Over35	0.568	0.960 *	1.424 **
	(0.407)	(0.406)	(0.523)
Gender: Male	-0.604	-1.003 **	-0.665
	(0.373)	(0.375)	(0.397)
Authorship: Yes	0.588	0.588	0.215
	(0.355)	(0.356)	(0.384)
Motivation: Yes	0.579	1.032 **	0.854 *
	(0.344)	(0.342)	(0.376)
Highest degree: PhD	0.711	0.682	0.664
	(0.383)	(0.377)	(0.410)
N. obs.	175	175	175
AIC	216.710	220.133	194.082

*** p < 0.001; ** p < 0.01; * p < 0.05.

We note that the coefficients for age and having written a motivation for the preference retain significance for two metrics, and gender retains significance for fluency.

The second robustness check we perform is variable selection using LASSO (Tibshirani 1996). We select the optimal shrinkage coefficient with cross-validation using the R package glmnet (Friedman, Tibshirani, and Hastie 2010). The results are shown in Table 3.

Lasso logistic regression on Bradley-Terry preferences for human-generated text

Table 3

	(1)
Intercept	-1.688
Age: Over35	0.766
Gender: Male	-0.554
Authorship: Yes	0.603
Motivation: Yes	0.583
Highest degree: PhD	0.667
Professional experience: Senior	0.000
Use of LLMs: Yes	0.000
Readership: High	0.053
English level: Advanced	-0.034
Time	0.000

Variable selection via LASSO leads to results very similar to the stepwise regression, retaining the same sign and relative magnitude. Two additional predictor - readership and English level - are retained by LASSO, but only with marginal coefficients.

The third robustness check involves excluding past authors of the *Economic Bulletin* from both the descriptive statistics and the logistic regression analysis. Figure 7 presents the distribution of preferences among respondents who have not contributed to previous editions of the *Economic Bulletin*. As expected, summaries generated by humans are the least preferred among this group.

Bradley-Terry preference scores excluding past authors of the *Economic Bulletin* Figure 7

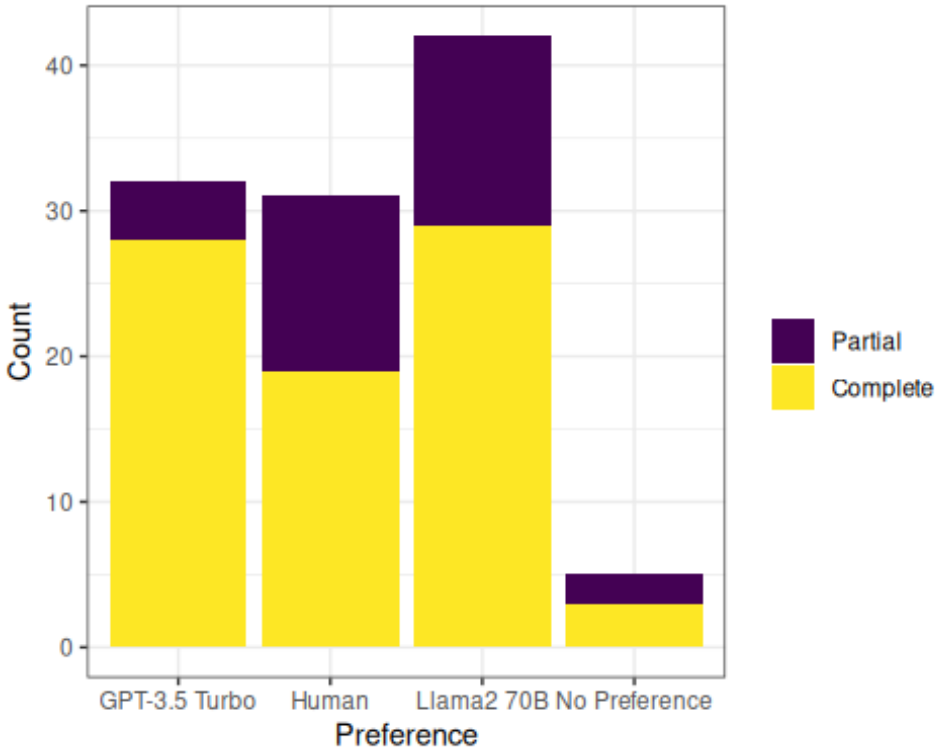


Table 4 reports the results of the logistic regression and stepwise selection procedure, excluding past authors of the *Economic Bulletin*, following the same methodology as in Table 1. The findings remain consistent with the analysis of the full sample, with the sole exception that the "Motivation" parameter is not statistically significant in the reduced sample.

Logistic regression on Bradley-Terry preferences for human-generated text excluding past authors of the *Economic Bulletin* Table 4

	(1a)	(2a)
Intercept	-2.858 **	-2.965 ***
	(0.910)	(0.750)
Age: Over35	2.580 **	2.207 **
	(0.912)	(0.696)
Gender: Male	-1.050 *	-0.967
	(0.533)	(0.513)
Motivation: Yes	0.759	
	(0.605)	
Highest degree: PhD	1.787 **	1.721 **
	(0.639)	(0.525)
English level: Advanced	-0.913	
	(0.630)	
Professional Experience: Senior	-0.404	
	(0.759)	
Use of LLMs: Yes	0.021	
	(0.578)	
Readership: High	0.733	
	(0.582)	
Time	-0.000	
	(0.000)	
N. obs.	110	110
AIC	121.835	114.818

*** p < 0.001; ** p < 0.01; * p < 0.05.

7. Conclusions

Our study confirms that large language models (LLMs) are capable of producing text of comparable quality to that produced by expert humans. However, the findings also highlight the presence of audience segmentation in preferences, influenced by demographic factors such as age, gender, and education level. Specifically, older respondents and those with doctoral degrees are more likely to prefer human-generated text, suggesting that experience and advanced education may shape expectations or familiarity with certain writing styles. Conversely, male respondents show a tendency to prefer LLM-generated content, which could be influenced by gendered perceptions of writing styles, as supported by previous research.

Additionally, the study highlights the role of emotional engagement and connection with the source material, as evidenced by the positive association between past authorship of the *Economic Bulletin* or providing a written justification for preferences and the inclination towards human-generated text. These findings

suggest that factors such as personal involvement and familiarity with a publication's distinctive style play a crucial role in shaping audience preferences.

While the models used to analyze these patterns show moderate predictive accuracy, they offer statistically significant insights over a no-information baseline. The consistency of the results across robustness checks supports the validity of the observed relationships between demographic characteristics and preferences.

These findings have important implications for the use of LLM-generated content in professional communications, suggesting that organizations should carefully consider their audience composition when determining optimal content generation strategies. Future research might productively explore whether these preference patterns persist across different document types and institutional contexts, and investigate the specific textual features that drive these demographically-based preferences.

References

Akaike, Hirotugu. 1969. "Fitting autoregressive models for prediction." *Annals of the Institute of Statistical Mathematics* 21 (1): 243–47.

Akaike, Hirotugu. 1971. "Autoregressive Model Fitting for Control." *Annals of the Institute of Statistical Mathematics* 23 (1): 163–80.

Bradley, Ralph Allan, and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39 (3/4): 324. https://doi.org/10.2307/2334029.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in r" 45: 1–67. https://doi.org/10.18637/jss.v045.i03.

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2025. "Chatbot arena: an open platform for evaluating LLMs by human preference." In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Vol. 235. JMLR.org, Article 331, 8359–8388. https://dl.acm.org/doi/10.5555/3692070.3692401.

Fabbri, Alexander R., Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. "SummEval: Re-Evaluating Summarization Evaluation." *Transactions of the Association for Computational Linguistics* 9: 391–409. https://doi.org/10.1162/tacl_a_00373.

Fatouros, Georgios, Kostas Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2024. "Can Large Language Models Beat Wall Street? Evaluating GPT-4's Impact on Financial Decision-Making with MarketSenseAI." *Neural Computing and Applications*, December. https://doi.org/10.1007/s00521-024-10613-4.

Fatouros, Georgios, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. 2023. "Transforming Sentiment Analysis in the Financial Domain with ChatGPT." *Machine Learning with Applications* 14 (December): 100508. https://doi.org/10.1016/j.mlwa.2023.100508.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent" 33. https://doi.org/10.18637/jss.v033.i01.

Gambacorta, Leonardo, Byeungchun Kwon, Taejin Park, Pietro Patelli, and Sonya Zhu. 2024. "CB-LMs: language models for central banking." BIS Working Papers 1215. Bank for International Settlements. https://ideas.repec.org/p/bis/biswps/1215.html.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2023. "Gemini: A Family of Highly Capable Multimodal Models." https://doi.org/10.48550/ARXIV.2312.11805.

Grassini, Simone, and Mika Koivisto. 2024. "Understanding How Personality Traits, Experiences, and Attitudes Shape Negative Bias Toward AI-Generated Artworks." *Scientific Reports* 14 (1). https://doi.org/10.1038/s41598-024-54294-4.

Hansen, Anne Lundgaard, and Sophia Kazinnik. 2023. "Can ChatGPT Decipher Fedspeak?" *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4399406.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. 2024. *The PRISM Alignment Dataset*. https://doi.org/10.57967/hf/2113.

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. "Gender Bias and Stereotypes in Large Language Models." *Proceedings of The ACM Collective Intelligence Conference*, November, 12–24. https://doi.org/10.1145/3582269.3615599.

Liu, Yixin, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, et al. 2023. "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2023.acl-long.228.

Mitton, Terri. 2023. "Results from the Survey on the Use of Generative AI for Communication." In *UNECE Expert Meeting on Dissemination and Communication of Statistics 2023*. Geneva. https://unece.org/sites/default/files/2023-10/DissComm2023_S3_OECD_Mitton_PPT.pdf.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. "GPT-4 Technical Report." https://doi.org/10.48550/ARXIV.2303.08774.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback". *Advances in neural information processing systems*, *35*, 27730-27744. https://dblp.org/rec/conf/nips/Ouyang0JAWMZASR22.html.

Pervez, Naseela, and Alexander J. Titus. 2024. "Inclusivity in Large Language Models: Personality Traits and Gender Bias in Scientific Abstracts." https://doi.org/10.48550/ARXIV.2406.19497.

Porter, Brian, and Edouard Machery. 2024. "AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably." *Scientific Reports* 14 (1). https://doi.org/10.1038/s41598-024-76900-1.

Schaik, Tempest A. van, and Brittany Pugh. 2024. "A Field Guide to Automatic Evaluation of LLM-Generated Summaries." *Proceedings of the 47th International ACM*

SIGIR Conference on Research and Development in Information Retrieval, July, 2832–36. https://doi.org/10.1145/3626772.3661346.

Smales, Lee A. 2023. "Classification of RBA Monetary Policy Announcements Using ChatGPT." *Finance Research Letters* 58 (December): 104514. https://doi.org/10.1016/j.frl.2023.104514.

Summers, Kate. 2013. "Adult Reading Habits and Preferences in Relation to Gender Differences." *Reference & User Services Quarterly* 52 (3): 243–49. https://doi.org/10.5860/rusq.52.3.3319.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." https://doi.org/10.48550/ARXIV.2302.13971.

Turner, Heather, and David Firth. 2012. "Bradley-Terry Models in R: The BradleyTerry2 Package". *Journal of Statistical Software* 48 (9):1-21. https://doi.org/10.18637/jss.v048.i09.

Wong, Jared, and Jin Kim. 2023. "ChatGPT Is More Likely to Be Perceived as Male Than Female." https://doi.org/10.48550/ARXIV.2305.12564.

Zhang, Haopeng, Philip S. Yu, and Jiawei Zhang. 2024. "A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models." https://doi.org/10.48550/ARXIV.2406.11289.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. "Benchmarking Large Language Models for News Summarization." *Transactions of the Association for Computational Linguistics* 12: 39–57. https://doi.org/10.1162/tacl_a_00632.

Appendix A

Summary of demographic characteristics for survey's participants.

Distribution of demographic details — Table 4

Gender	*Female*	*Male*					
	61	114					
Age	*26-35*	*36-45*	*45-55*	*>55*			
	48	63	43	21			
Professional experience	*≤5*	*6-10*	*11-15*	*16-20*	*21-25*	*26-30*	*>30*
	31	37	27	31	24	7	18
Highest degree	*High School*	*Bachelor*	*Master*	*PhD*			
	5	21	50	99			
Highest degree field	*Economic and Finance*	*Engineering*	*Law and Humanities*	*Mathematics and Statistics*	*Other*		
	118	10	12	23	12		
English level	*Elementary*	*Intermediate*	*Advanced*				
	5	32	138				
Authorship of the Economic Bulletin	*No*	*Yes*					
	110	63					
Readership of the Economic Bulletin	*Never*	*Rarely*	*Sometimes*	*Often*	*Regularly*		
	8	34	49	41	43		
Use of LLMs	*No*	*Yes*					
	51	124					
Directorate	*Economic Outlook and Monetary Policy*	*Financial Stability*	*International Relations and Economics*	*Statistical Analysis*	*Statistical Data Collection and Processing*	*Structural Economic Analysis*	*Other*
	45	21	20	23	23	35	8

Appendix B

Survey administered to the participants.

Large Language Models

This survey aims to **evaluate the usage of Large Language Models** (LLMs) by personnel in the Economics and Statistics Department and **assess the quality of LLMs output** in the context of official Central Bank communications.

Your answers will help support the development of Bank of Italy's strategy and guidelines for LLMs usage. Please read each question carefully and answer using your best judgement.

The expected time to complete the questionnaire is **20 minutes**.

Selected chapters from the Bank of Italy Economic Bulletin (April 2023) were processed to obtain a concise summary using various LLMs. Your task is to express your preference between five pairs of text based on three metrics:

- **Fluency**: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.
- **Relevance**:A summary is fully relevant if it extracts the most important points from the source.
- **Consistency**: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Your responses will be processed anonymously and only used for this research.

There are 33 questions in this survey.

Demografics

Have you ever used ChatGPT or other Large Language models (LLMs)? *

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Yes
- ◯ No

How often do you use ChatGPT or other LLM? *

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '1 [LLMUSEYN]' (Have you ever used ChatGPT or other Large Language models (LLMs)?)

ⓘ Choose one of the following answers
Please choose **only one** of the following:

- ◯ I have only used it once or twice
- ◯ Once a month
- ◯ Once per week
- ◯ Several times a week
- ◯ Every day

How do you use ChatGPT or other LLMs?

*

Only answer this question if the following conditions are met:
Answer was 'Yes' at question '1 [LLMUSEYN]' (Have you ever used ChatGPT or other Large Language models (LLMs)?)

ⓘ Check any that apply
Please choose **all** that apply:

- ☐ Institutional purposes
- ☐ Drafting text
- ☐ Summarizing text
- ☐ Writing software code
- ☐ Translating documents
- ☐ Personal purposes (research, etc)
- ☐ Other:

What applications of ChatGPT or other LLMs may better support your work?

Check any that apply
Please choose **all** that apply:

- ☐ Drafting text
- ☐ Summarizing text
- ☐ Writing software code
- ☐ Translating documents
- ☐ Other:

Assume you have access to fully private LLM services suitable for confidential data.

Gender:

Choose one of the following answers
Please choose **only one** of the following:

- ◯ Female
- ◯ Male
- ◯ Prefer not to say

Age

Choose one of the following answers
Please choose **only one** of the following:

- ◯ =< 25
- ◯ 26 - 35
- ◯ 36 - 45
- ◯ 46 - 55
- ◯ > 55

Familiarity with English:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Beginner
- ◯ Elementary
- ◯ Intermediate
- ◯ Advanced

Highest degree obtained:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ High School
- ◯ Bachelor
- ◯ Master
- ◯ PhD

Degree topic:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Economics and Finance
- ◯ Mathematics and Statistics
- ◯ Engineering
- ◯ Law and Humanities
- ◯ Other

Years of professional experience:

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ =< 5
- ◯ 6 - 10
- ◯ 11 - 15
- ◯ 16 - 20
- ◯ 21 - 25
- ◯ 26 - 30
- ◯ > 30

Did you author or edit any edition of Bank of Italy's Economic Bulletin in the last 5 years? *

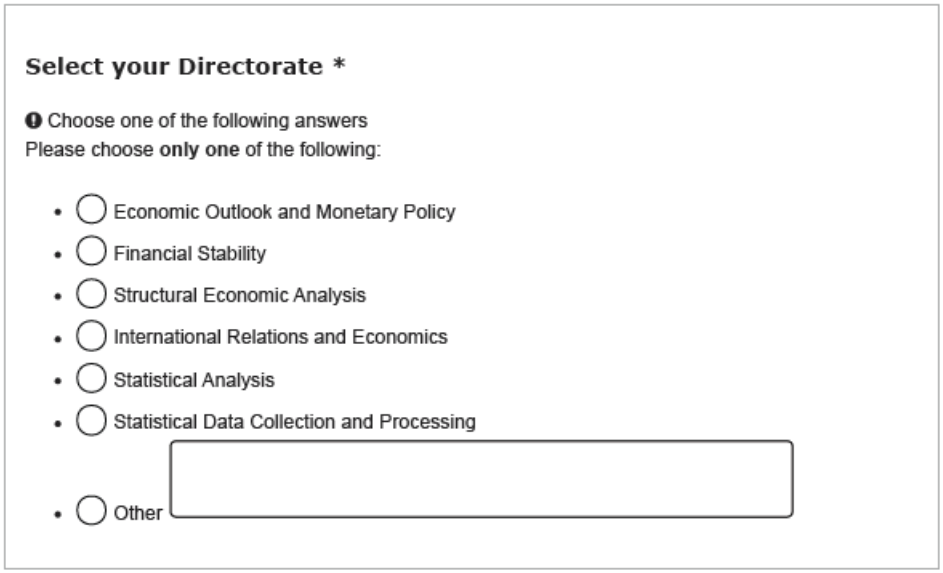
ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Yes
- ◯ No

How often do you read Bank of Italy's Economic Bulletin? *

ℹ Choose one of the following answers
Please choose **only one** of the following:

- ◯ Never
- ◯ Rarely
- ◯ Sometimes
- ◯ Often
- ◯ Regularly



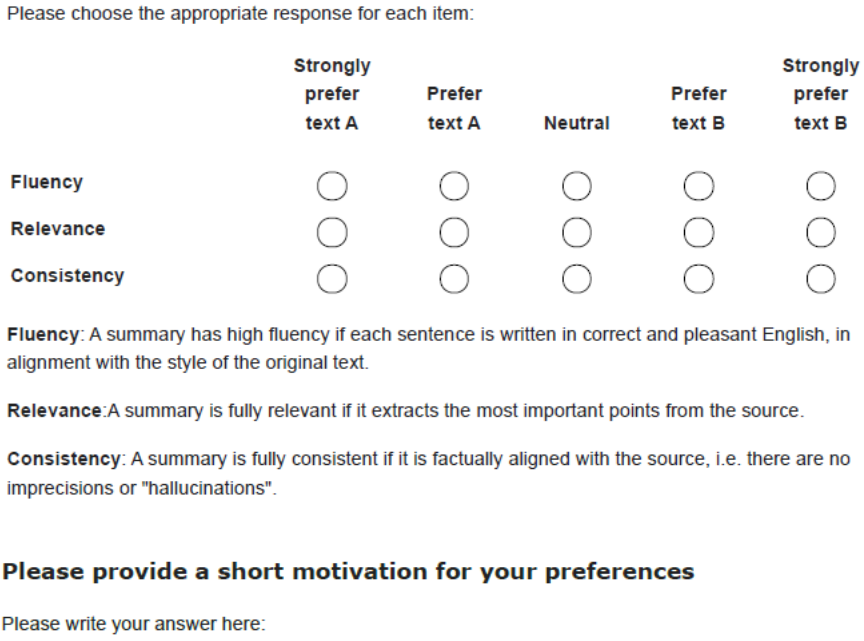
Text comparison 1

Text A and text B are two alternative summaries of chapter 1.2 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (Human)	Euro-area GDP stagnated in the last quarter of 2022, following a prolonged period of growth. The available indicators point to a slight expansion in economic activity in the early months of 2023. The fall in consumer price inflation continued, driven by the energy component; core inflation, which already stood at the highest levels, rose again. The Governing Council of the European Central Bank further raised the key interest rates, indicating that the high level of uncertainty is making it all the more important to adopt future decisions meeting by meeting and based on the data that become available. The European Union's plan to reduce energy dependence on Russia and to step up the green transition (REPowerEU) was approved.
Option 2 (GPT-3.5 Turbo)	The Euro Area witnessed stagnant GDP in Q4 2022 but hints at a mild upturn in early 2023. Despite falling consumer price inflation due to energy prices, core inflation rose. The European Central Bank responded by further raising key interest rates, emphasizing the need for adaptive decisions based on evolving data amid pervasive uncertainty. The Euro Area's plan, REPowerEU, aimed at reducing energy reliance on Russia and advancing the green transition, gained approval. GDP growth remains cautious, inflation forecasts revised downward, and although wage growth strengthens, risks of an inflationary spiral appear subdued. The ECB's decisive monetary tightening continues alongside reduced asset purchases and moderate credit to firms amidst increased loan costs. The Council of the EU endorsed the REPowerEU initiative to diminish energy reliance on Russia and promote sustainability.

Option 3 (LlaMa2 70B)	The eurozone economy experienced nil GDP grow that the end of 2022, but it slightly improved in the first quarter. Household consumption and gross fixed investment declined while net foreign demand contributed to GDP growth. The inflation rate fell for the fifth consecutive month, driven by energy product prices, while core inflation continued to rise. Wage growth strengthened, but risks of an upward wage-price spiral remain low. The ECB tightened monetary conditions by increasing key interest rates and announced that it will continue monitoring ongoing financial market tensions. Credit to firms decreased, while lending to households continued to grow at a slower pace. The cost of loans rose following the increase in key interest rates. The European Commission approved the REPowerEU plan to reduce energy dependence on Russia and step up the green transition.

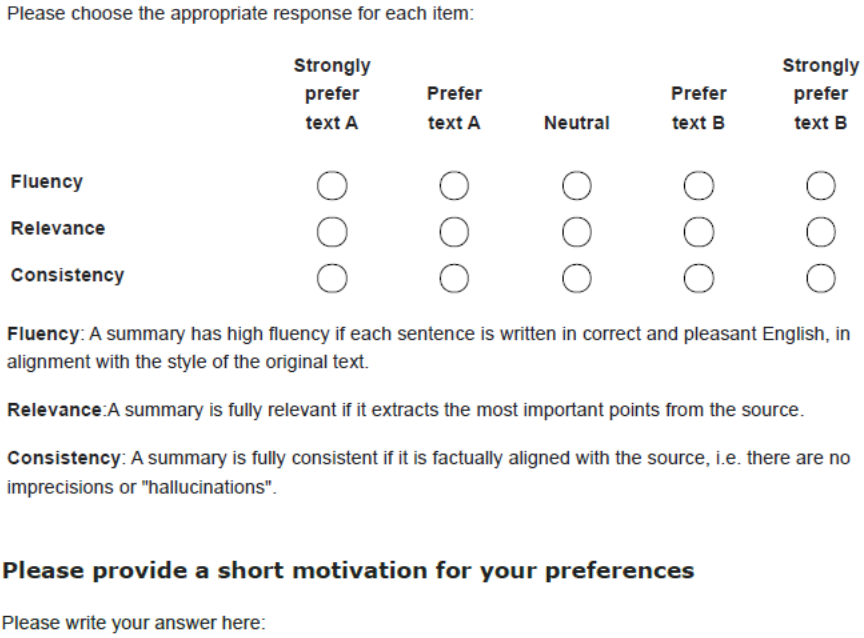


Text comparison 2

Text A and text B are two alternative summaries of chapter 2.2 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (GPT-3.5 Turbo)	Manufacturing activity has resumed growth in Q1 2023, while service sector indicators suggest a recovery. Business confidence has improved across all sectors, and investment has strengthened. However, the real estate market remains weak, with declining house sales and prices rising less than consumer price inflation. The debt-to-GDP ratio of Italian non-financial corporations has declined further.
Option 2 (LlaMa2 70B)	Industrial production appears to return to growth in Q1, despite a drop in capital goods and intermediate goods. Business confidence improved across all sectors, with positive signs from PMIs. Investment accelerated in Q4, while the real estate market weakened. Firms' debt-to-GDP ratio declined further, with liquidity remaining high.
Option 3 (Human)	After having declined fortwo consecutive quarters, manufacturing activity appears to have resumed its expansion in the first three months of 2023. The qualitative indicators for firms in the service sector point to a recovery inactivity. According to firms, capital accumulation will continue this year. Real estate market conditions remain weak.



Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

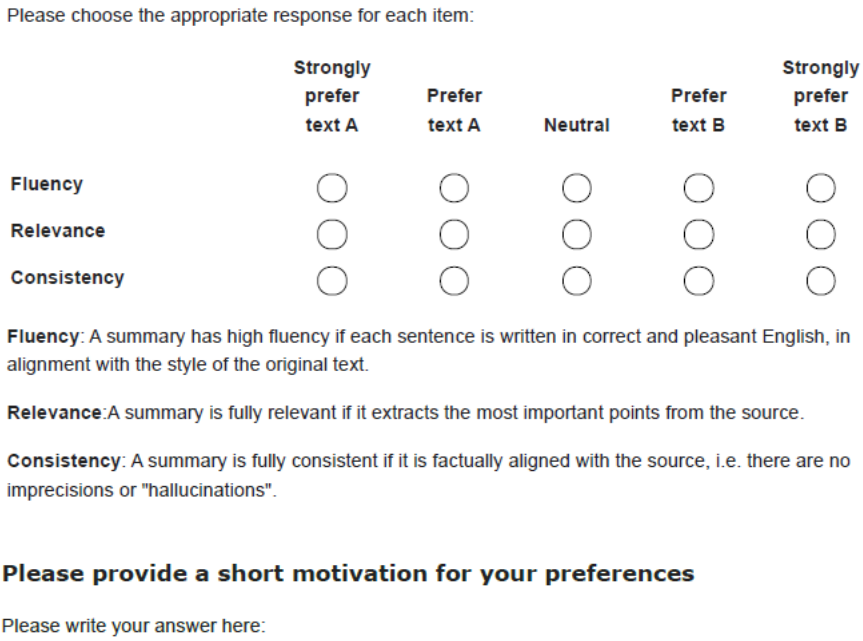
Please write your answer here:

Text comparison 3

Text A and text B are two alternative summaries of chapter 2.6 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (LlaMa2 70B)	Harmonized consumer price inflation in Italy decreased to 8.2% in March 2023 from its peak of 12.6% last autumn, primarily due to a drop in energy prices. Food prices continued to rise sharply, driven by the pass-through of past energy price increases. Producer price pressures eased, and wage pressures on prices remained subdued. Inflation expectations among firms and households continue to decline.
Option 2 (Human)	After peaking at the end of last year, consumer price inflation started to decline, reflecting the sharp drop in the energy component, while core inflation increased. Households' and firms' expectations point to continued signs of inflationary pressures easing.
Option 3 (GPT-3.5 Turbo)	Consumer price inflation in Italy decreased to 8.2% in March due to a reduction in energy prices, while core inflation increased. Food prices continued to rise sharply, driven by the pass-through of energy price increases. Producer price pressures continue to ease, and wage pressures on prices remain muted. Inflation expectations of firms and households continue to decline.



Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

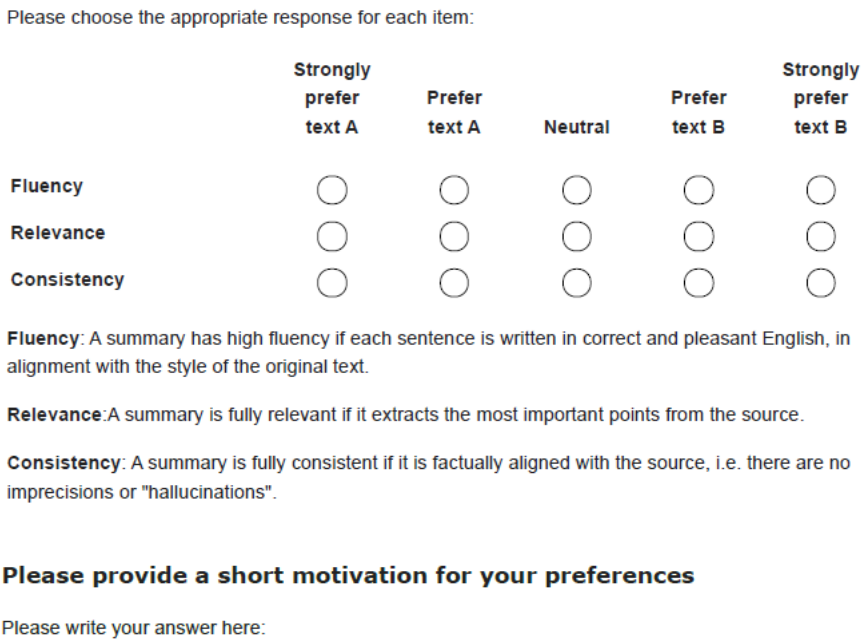
Please write your answer here:

Text comparison 4

Text A and text B are two alternative summaries of chapter 2.7 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (LlaMa2 70B)	Lending to non-financial corporations declines due to tighter credit standards and higher funding costs. Household lending is also slowing down, mainly due toa slowdown in home mortgages. The average interest rate on new bank loans to firms and households increases, while bank funding decreases. Despite tensions across the international banking system, Italian banks are in a better position compared to past crises. The non-performing loan rate remains low, and the profitability of significant banking groups increases. Capital ratios rise driven by reduction in weighted assets and Tier-1 capital growth.
Option 2 (GPT-3.5 Turbo)	The Italian banking sector saw a decline in lending to businesses due to tightened credit standards and higher funding costs, notably impacting firms' borrowing and repayment behaviors. The interest rate hikes continue to influence borrowing costs for both firms and households. Bank funding contracted due to repayments and reduced liabilities to the Eurosystem, as deposits decreased while funding costs rose. Despite global banking tensions post US and Swiss bank failures, Italian banks stand in a relatively stronger position. Non-performing loan rates remain low, and significant Italian banking groups exhibit improved profitability and increased capital ratios, attributed to reduced loan loss provisions and capital growth.
Option 3 (Human)	Between November and February, the growth rate of bank loans decreased sharply, particularly for loans to firms, reflecting higher funding costs and tighter credit standards. Key interest rate hikes continue to pass through to the cost of lending. Despite recent tensions following the failures of some banks in the United States and Switzerland, the Italian banking system is in a better position compared with crises in the past. The non-performing loan rate is still low. The profitability and capital ratios of Italian banking groups classified as significant improved in2022.



Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

Please write your answer here:

Text comparison 5

Text A and text B are two alternative summaries of chapter 2.8 of the Economic Bulletin published on April 2023.

Please rate your preferences according to fluency, relevance, and consistency.

Option 1 (Human)	In Italy as in other major financial markets, the tensions sparked by the collapses of some banks in the US and in Switzerland starting from 9 March have influenced the trends in asset prices. After rising in January and February, government bond yields and share prices fell sharply amid the first signs of instability in the international banking sector. Volatility increased as well, while yield spreads between Italian ten-year government bonds and the corresponding German Bunds held stable.
Option 2 (LlaMa2 70B)	Yields on Italian government bonds edged up 9 basis points to 4.1% in March, driven by increased concerns about policy rate hikes and risks to financial stability. Bond funding became more expensive, with yields on non-financial corporate and bank bonds rising 9 and 22 basis points respectively. Equity prices gained 4.8% for the Italian stock market index and5.2% for the banking index in January and February, but have since been impacted by increased volatility and downward pressure following the collapse of US banks and Credit Suisse.
Option 3 (GPT-3.5 Turbo)	Italian government bond yields edged up by 9 basis points to 4.1% at the end of March, while yields on bonds issued by Italian non-financial corporations and banks also rose, albeit to a lesser extent than in the euro area. Equity prices have gained since mid-January, but have been impacted by the collapses of some US banks and Credit Suisse, resulting in a spike in volatility and strong downward pressure on share prices, especially in the financial sector.

Please choose the appropriate response for each item:

	Strongly prefer text A	Prefer text A	Neutral	Prefer text B	Strongly prefer text B
Fluency	◯	◯	◯	◯	◯
Relevance	◯	◯	◯	◯	◯
Consistency	◯	◯	◯	◯	◯

Fluency: A summary has high fluency if each sentence is written in correct and pleasant English, in alignment with the style of the original text.

Relevance:A summary is fully relevant if it extracts the most important points from the source.

Consistency: A summary is fully consistent if it is factually aligned with the source, i.e. there are no imprecisions or "hallucinations".

Please provide a short motivation for your preferences

Please write your answer here:

Appendix C

C.1 The Bradley-Terry Model: formal specification and estimation

The Bradley-Terry model (Bradley and Terry 1952) is a probabilistic model for analysing paired comparison data, which we use in this study to convert respondents' pairwise preferences into a comprehensive ranking of summary sources. Let $i, j \in \{1, \ldots, m\}$ index the alternative summaries under comparison. The Bradley–Terry model assigns each summary a latent score $\beta_k \in R$ such that the probability that i is preferred to j is:

$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j},$$

with $Pr(j \succ i)$ obtained by swapping the indices. This can be equivalently expressed in logistic form:

$$Pr(i \succ j) = \frac{1}{1 + e^{-(\beta_i - \beta_j)}},$$

where $\beta_i = \log \pi_i$. β_i and β_j are positive-valued parameters which can be thought of as representing 'ability', and in the context of the study the ability to produce valid summaries of the *Economic Bulletin*.

The key advantage of the Bradley-Terry model is that it allows us to derive a complete ranking of multiple sources (in our case, human-generated, GPT-3.5 Turbo, and LLaMA2 70B summaries) from pairwise comparison data. While each individual comparison only involves two sources, the model can estimate ability parameters for all three sources by combining information across all pairwise contests. For identifiability, one source must be set as the reference category with its parameter constrained to zero. The estimated ability parameters for the other sources represent log-odds relative to this reference. These parameters can then be transformed back to the original scale, providing the basis for an absolute ranking of all sources.

To estimate the parameters of the Bradley-Terry model from our data, we use maximum likelihood estimation. We perform the estimation using the *BradleyTerry2* package in R (Turner and Firth 2012). To account for the 5-point Likert scale used in our survey, we weighted the comparisons as follows: strong preferences were given a weight of 2, regular preferences a weight of 1, and neutral responses were treated as 0.5 preference for each option. The model was first applied to each individual across all the answers given, in order to derive a general preference indication, and then separately to each quality dimension (fluency, consistency, and relevance), yielding dimension-specific preference parameters for each source.

C.2 Pairwise comparisons and the Bradley Terry Model in LLM evaluations

Pairwise comparison serves as an effective methodology to collect human preferences on outputs generated by Large Language Models. On one hand, pairwise preference data have been utilized to obtain data for the "human alignment" stage

of LLM training (Ouyang et al. 2022). On the other hand, they have been used to compare performance across different models, with statistical frameworks such as Bradley-Terry and Elo enabling the extraction of global rankings from these comparisons. Chiang et al. (2024) employed this latter approach in the Chatbot Arena framework, framework, which has emerged as a standard for human evaluation of LLM outputs.

The pairwise comparison approach reduces the cognitive load on evaluators, who need only compare two texts directly rather than assign absolute scores. It also eliminates scale biases that may arise when different evaluators interpret rating scales differently.

BankGPT: the use of Large Language Models in official communications

CLAUDIA BIANCOTTI, CAROLINA CAMASSA, MARCO FRUZZETTI, **LUIGI PALUMBO**, MYRIAM PORTALURI

LLMs and official communications

The **use of LLMs is increasingly pervasive** in business and official settings.

Potential efficiency gains and lagging guidelines for official use.

Our research questions:

- Is there a **quality gap** between text generated by LLMs and human experts?
- Are **preferences** related to specific **characteristics of the audience**?

Some references

OECD institutions' staff

23% use GenAI
76% do not have guidelines
(Mitton, 2023)

LLMs for economic analysis

Financial sentiment analysis
(Fatouros et al., 2023)
Investment signals
(Fatouros et al., 2024)

LLMs and CB communication

Deciphering Fedspeak
(Hansen and Kazinnik , 2023)
RBA monetary policy statements
(Smales, 2024)

CB Language Models

Special-purpose text models
LLMs for CB workflows
(Gambacorta et al., 2024)

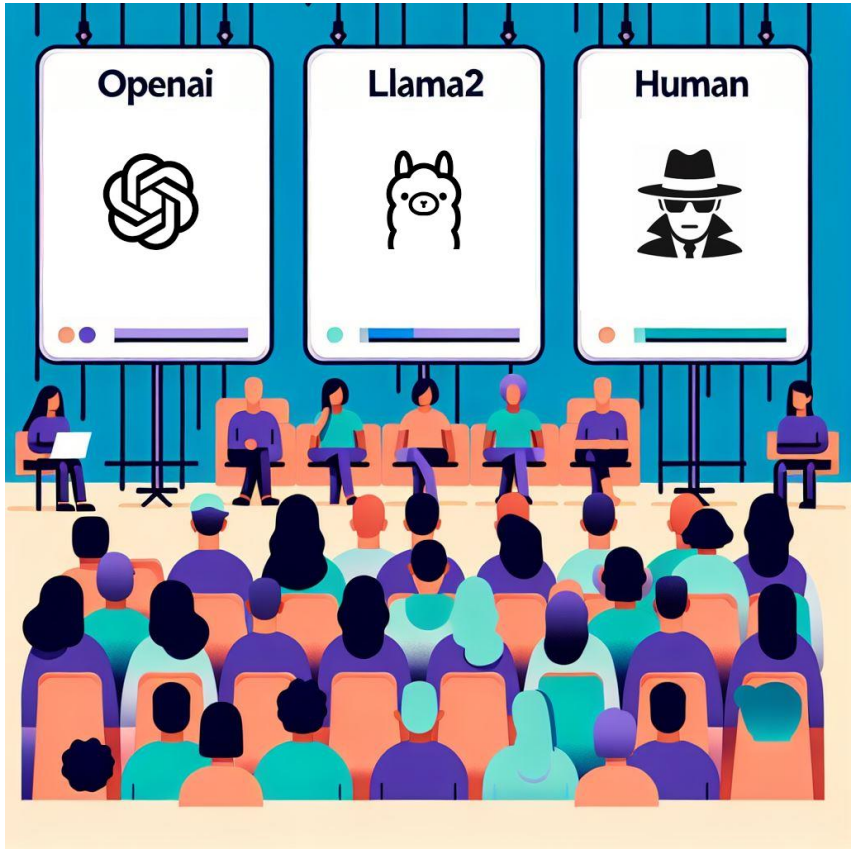
LLMs and summarization

LLMs on par with humans on news… (Zhang et al., 2024)
…but falling short on specific tasks (Lui et al., 2023)

GenAI vs. Human

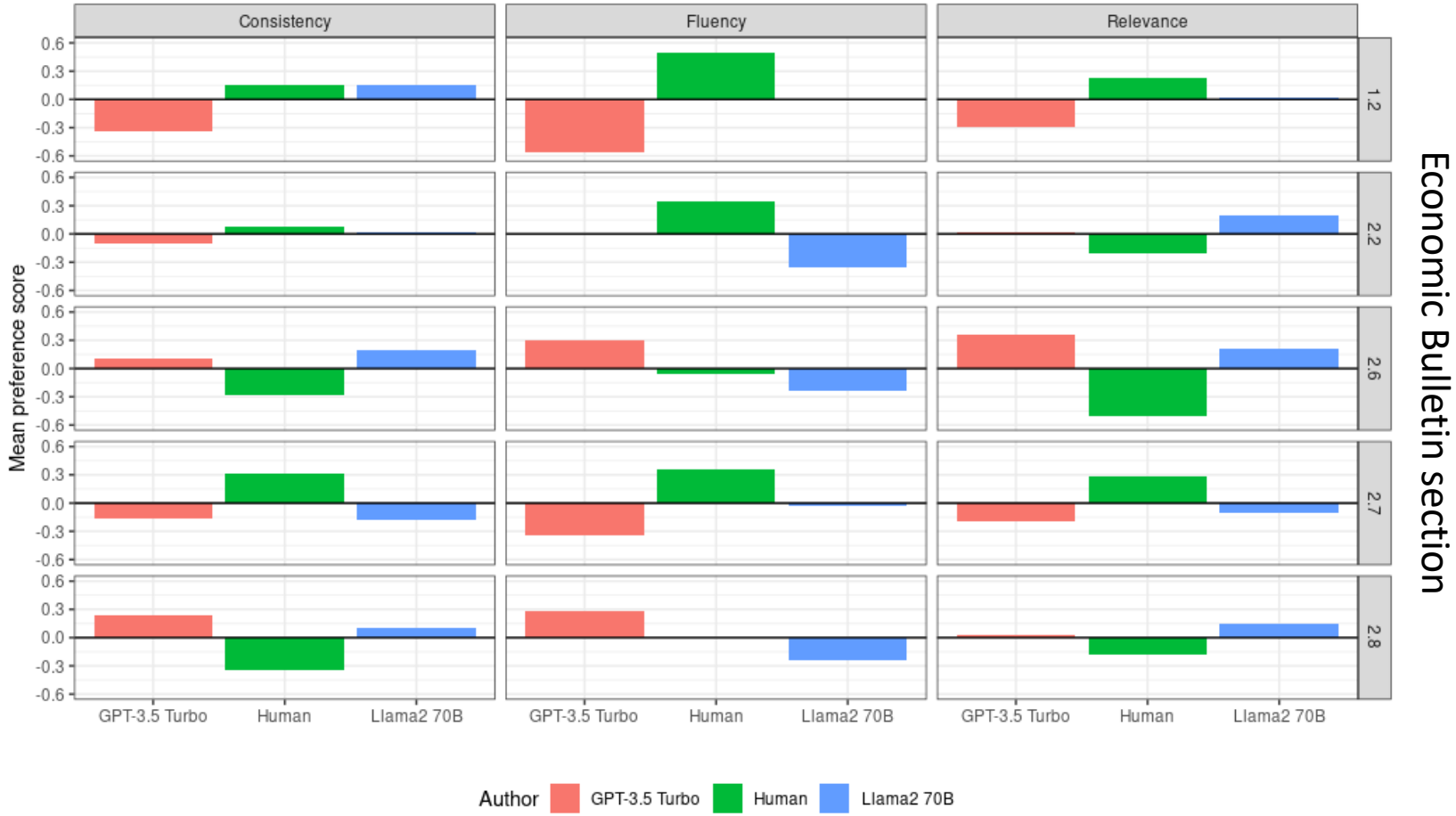
GenAI text better than human
(Porter and Machery, 2024)
Cannot distinguish, and bias
(Grassini and Koivisto, 2024)

Our experiment

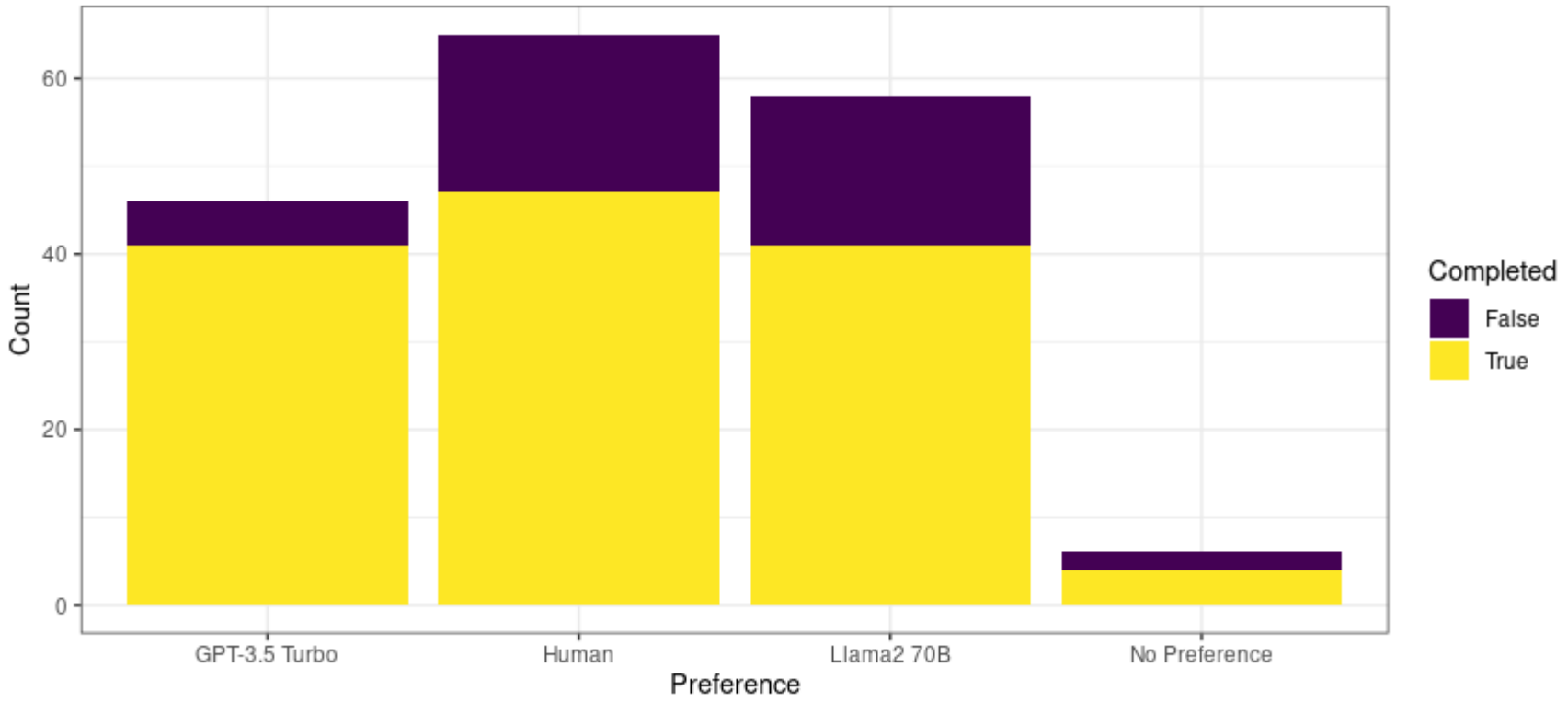


- Target: Bank of Italy's **Economics and Statistics Directorate General**
- Good participation, with **175** answers out of about 600 colleagues
- Pairwise comparison **of *Economic Bulletin* summaries**, all presented as LLM-generated
- Three evaluation metrics: **fluency, consistency and relevance** (Fabbri et al., 2021)
- Bradley-Terry model to convert comparisons into **comprehensive ranking of models** (Bradley and Terry, 1952)

Preferences are varied across sections and metrics



Human authors are (slightly) preferred



Age, gender, and education

	(1)		(2)	
Intercept	-2.064 **	*(0.638)*	-2.175 ***	*(0.528)*
Age: Over35	1.255 *	*(0.627)*	1.117 **	*(0.429)*
Gender: Male	-1.008 *	*(0.395)*	-0.968 *	*(0.389)*
Authorship: Yes	0.793 *	*(0.397)*	0.797 *	*(0.360)*
Motivation: Yes	0.952 *	*(0.414)*	0.858 *	*(0.352)*
Highest degree: PhD	1.119 *	*(0.447)*	1.052 **	*(0.396)*
English level: Advanced	-0.660	*(0.479)*		
Professional Experience: Senior	0.016	*(0.557)*		
Use of LLMs: Yes	0.247	*(0.428)*		
Readership: High	0.164	*(0.406)*		
Time	-0.000	*(0.000)*		
N. obs.	175		175	
AIC	218.931		211.088	

*** p < 0.001; ** p < 0.01; * p < 0.05.

Conclusions

- LLMs are capable of producing text of **comparable quality** to that produced by expert humans…
- …but stronger **domain expertise** steers readers' preferences towards human-generated text.
- Salience of **demographic factors** such as age and gender for preferences' prediction raise questions about **potential bias** in LLMs' training material and process.
- Implications for institutional communications: organizations should carefully **consider their audience composition** when determining optimal content generation strategies.

Q&A

THANKS!