Plausible but Wrong: A Statistical Evaluation of GPT Models on High-Stakes Migration Statistics

Hector M. Ramírez Contreras[a*], **Ravinithesh Annapureddy**[b*], **Daniel Gatica-Perez**[c]
Idiap Research Institute and EPFL, Switzerland
[a]hecmaramco@gmail.com, [b]ravinithesh.annapureddy@idiap.ch,
[c]gatica@idiap.ch

Abstract

LLMs increasingly serve as general-purpose information systems, including for factual queries about public statistics. In high-stakes domains such as migration, where data accuracy shapes both policy and public attitudes toward at-risk populations, models may answer queries about social groups unequally. We evaluate this concern empirically using GPT-4 Turbo, GPT-4o, GPT-5 chat, and GPT-5.1 on migration-related education statistics. Each model receives zero-shot queries on educational-attainment distributions for national and foreign citizens aged 25–54 across 34 European countries from 2020–2023. We compare outputs against official Eurostat data using RMSE, total variation distance (TVD), cosine similarity, distributional analysis, and response-coverage rates. Across models, results show a consistent gap between plausibility and accuracy: all models preserve broad distributional structure (cosine similarity > 0.93), while deviating substantially from ground-truth values (RMSE 10–14 and TVD 12–16 percentage points). We characterize this combination of structural plausibility and numerical error as an output-level analog of the illusion of competence. GPT-4 models also produce explicit non-answers for 30–78% of responses, with refusals (explicit non-answers) up to 3.4 times as frequent for foreigners than for nationals. GPT-5 models eliminate refusals but return hallucinated values. These findings show that pluralistic evaluation must distinguish legitimate human variation from unsupported model variation and assess whether factual information remains comparably available and reliable across social groups.

1 Introduction

Official statistics support research, policy, and journalism, and access to them is a cornerstone of a democratic society. Statistical agencies produce these data through documented surveys and quality-control processes and make them available through open data repositories. Large language models (LLMs), however, are changing this as they are widely deployed as general-purpose information systems (Bommasani et al., 2021; OpenAI et al., 2024). Users increasingly treat LLMs as search engines, asking for facts, statistics, and data analysis alongside explanations and practical guidance (Spatharioti et al., 2025; Appel et al., 2025). Seeking information is the fastest-growing ChatGPT use case, increasing from 14% to 24% over one year (Chatterji et al., 2025), and even in conversations whose primary intent is non-informational, between 18% and 76% of interactions contain check-worthy factual assertions made by the LLM (Joko et al., 2026).

This shift is especially pertinent in high-stakes socio-economic domains like migration. Migration statistics, such as educational attainment, labor participation, or demographics, are susceptible to political and social misuse; generating or synthesizing such data raises the risk of misinformation, prejudice amplification, and distorted public discourse (de Haas, 2024; Weidinger et al., 2022). The concern is not that domain experts would rely on LLMs as authoritative databases, but that the wider adoption of LLMs exposes researchers, policymakers, journalists, and citizens to plausible but unreliable claims about at-risk populations (Chakravorti et al., 2025). Real-world incidents already show that LLM-generated text, including embedded statistics, reaches the public through institutional channels: For instance, a U.S. federal judge noted that immigration agents used ChatGPT to draft a use-of-force narrative, raising accuracy and privacy concerns (Lauer, 2025), the UK Home Office's evaluation of an LLM-based tool in asylum decision-making found that 9% of summaries had to be removed for inaccuracies (UK Home Of-

*Equal contribution.

fice, 2025), and surveys indicate that up to 30% of European public-sector managers use off-the-shelf LLMs to draft reports (Tangi et al., 2024). Public-sector officials already report using generative AI for report writing and synthetic-data generation (Sanford, 2025; Bright et al., 2025). Therefore, rigorously assessing the factual reliability of LLM outputs is both an academic necessity and a societal imperative (Seth et al., 2025).

Public statistics pose a different evaluation problem than many NLP tasks, where multiple judgments can legitimately coexist. When generated values lack support in the underlying evidence, they are factual errors rather than alternative perspectives. This distinction matters for pluralistic NLP: preserving meaningful human variation requires distinguishing disagreement grounded in perspectives or lived experience from unsupported variation introduced by the model itself (Plank, 2022; Sorensen et al., 2024; Xu et al., 2026). Migration statistics make this distinction consequential. Statistical categories also determine who becomes visible in public accounts: "foreigners," "migrants," and "residents" are related but not interchangeable populations. If an AI system provides information about one population less often or less accurately than another, it can make access to authoritative public evidence uneven across the groups being represented, even without discriminatory intent in the query.

Existing work establishes several parts of this problem. LLMs often hallucinate and fabricate sources (Ji et al., 2023; Huang et al., 2025). Studies that compare synthetic survey or demographic outputs with empirical data find plausible aggregate patterns but poor recovery of real population variation (Bisbee et al., 2024; Dominguez-Olmedo et al., 2024; Seth et al., 2025). Other audits document group- and domain-dependent refusal behavior (Leidinger and Rogers, 2025; Atif et al., 2025). These lines of work rarely connect three outcomes in a single reference-grounded task: whether a model answers, whether the response forms a valid statistic, and whether the statistic is correct.

We study this problem through a black-box evaluation of four standalone GPT models on a concrete, policy-relevant task: reproducing official migration-related statistics. We ask each model for the percentage distribution of educational attainment (low/medium/high, based on the International Standard Classification of Education (ISCED 2011) levels (UNESCO Institute for Statistics, 2012)) among foreign and national citizens aged 25–54. The queries cover 34 European countries, two sexes, and the years 2020–2023. We repeat each query ten times and compare numerical outputs with ground truth from official Eurostat data. We test four GPT models spanning two generations (GPT-4 Turbo, GPT-4o, GPT-5 chat, and GPT-5.1) using zero-shot prompts via API with controlled parameters without providing any external tools or requirements, and ask two research questions:

RQ1: How reliably do standalone GPT models reproduce migration-related education statistics, and when they produce numerical responses, how closely do they align with official Eurostat data?

RQ2: How does model response behavior vary across groups, including explicit non-answers and invalid or inaccurate numerical responses?

In response, we make three contributions:

1. We evaluate standalone GPT models on precise public-statistics questions, separating explicit non-answers, invalid percentage distribution, and factual error by comparing with official statistics (Section 3).

2. We characterize an output-level analog of the *illusion of competence*, showing that GPT models generate outputs that preserve broad distributional structure (high cosine similarity) while deviating substantially from ground-truth values (RMSE 10–14 percentage points). (Section 4.1 and 5.1)

3. We document disparate response behavior across demographic groups. GPT-4 models provide explicit non-answers for foreign-citizen queries up to 3.4 times the rate of national-citizen queries. GPT-5 models eliminate refusals entirely but replace them with hallucinated values. Both failure modes disproportionately affect information about foreign populations. (Section 4.2 and 5.2)

Together, these findings connect factual reliability with broader questions of when model variation constitutes meaningful pluralism and whether AI-mediated public information remains comparably available across the populations it describes.

2 Related Work

2.1 LLMs as Information Interfaces

LLMs are increasingly mediating access to factual information through conversational search and chat assistants (Spatharioti et al., 2025). Unlike traditional search tools that return links to specific, existing web pages, LLMs generate aggregate responses probabilistically and often produce answers even without relevant facts or verification (Zheng et al., 2024). A well-documented consequence is hallucination, where models generate factually incorrect or fabricated content that resembles reliable information (Ji et al., 2023; Huang et al., 2025; Bender et al., 2021). This risk is especially concerning when fabricated statistics lend false credibility to policy claims (Chakravorti et al., 2025). Researchers have explored LLMs as implicit knowledge bases (Zheng et al., 2024) and for synthetic data generation in social science, finding that LLM-generated survey responses are often plausible at the aggregate level but diverge from real human variability (Bisbee et al., 2024; Brand et al., 2023). These findings motivate quantitative evaluation of LLM outputs rather than assessment through textual reasoning alone.

2.2 Meaningful variation and factual consensus

Pluralistic NLP challenges the assumption that every task has one universally correct label. Human variation can reflect ambiguity, subjectivity, social position, or defensible differences in interpretation (Plank, 2022; Mostafazadeh Davani et al., 2022), motivating systems that represent multiple views or their population distribution (Sorensen et al., 2024). Preserving such variation can also support socio-technical safety by grounding evaluation in human and social context (Xu et al., 2026). This does not remove the need for factual reference points. Public statistics depend on methodological choices, but once a query specifies the indicator and population, reproducing the published value is a reference-grounded task. A pluralistic public-information system should support diverse interpretations of statistics, not diverse unsupported statistics. Following human-centered calls for context-aware evaluation (Blodgett et al., 2024), we therefore combine aggregate error metrics with response modes, repeated behavior, and group-disaggregated results.

2.3 LLMs and Disparate Treatment Across Social Groups

A growing body of research documents that LLMs do not treat all social groups uniformly, including more homogeneous representations of subordinate groups, demographic distributions that diverge from official records, and uneven retrieval of disaggregated information (Lee et al., 2024; Seth et al., 2025; Balasuriya Mudiyanselage et al., 2025; Dominguez-Olmedo et al., 2024). These studies span health, census, and demographic data. Related work also finds group- and domain-dependent differences in refusal and abstention behavior (Leidinger and Rogers, 2025; Atif et al., 2025) in LLMs when evaluating stereotyping harms and religious question-answering. Together, these studies show that both information availability and factual reliability can vary across the populations represented in model outputs. We extend this work to a policy-sensitive domain, examining whether standalone GPT models provide and reproduce public information comparably for different demographic groups.

2.4 LLMs in High-Stakes and Migration Domains

Migration statistics constitute a sensitive domain where accuracy is critical. Prior AI work in migration has focused on analytical applications like predicting trends, analyzing public sentiment (Schmeisser-Nieto et al., 2024), or governing borders rather than on data provision. Recent work (Yar and Bircan, 2025) has explored big data sources to supplement official migration statistics but has not examined LLMs for data generation. In relation to LLMs, previous studies addressed their use for migrant support (Lee et al., 2025), agent-based policy simulations (Khatua, 2025), and semantically linking immigration data (Venkatakrishnan et al., 2024). LLMs have also been used in other public-sector or high-stakes information settings, such as analyzing linguistic bias in government documents (de Swart et al., 2025). Still, these works focus on textual content rather than numerical statistics. Our work contributes to this space by quantitatively comparing LLM outputs with official statistics and evaluating LLMs as information sources whose outputs reach the public.

3 Methods

3.1 Task

We frame the study as a controlled standalone question-answering task, simulating a user directly asking an LLM for public information. Rather than directing the models to any specific dataset, we assess their inherent ability to provide accurate and complete responses from internalized knowledge. This design lets us evaluate both the numerical accuracy and behavioral consistency of the models when faced with sensitive, factual questions in the most common way web-based LLMs are used.

The task evaluates LLM-provided estimates of the percentage distribution of educational attainment among national and foreign citizens across European countries (listed in Appendix A) and years (2020–2023). Educational attainment is categorized using ISCED 2011 (UNESCO Institute for Statistics, 2012) into Low (ED0–2: less than primary, primary and lower secondary education), Medium (ED3–4: upper secondary, post-secondary and non-tertiary education), and High (ED5–8: tertiary education). We choose educational attainment because it is a primary proxy for migrant integration and labor market potential. Unlike volatile statistics (e.g., daily border crossings), education levels are relatively static and derived from EU Labor Force Survey (Eurostat, 2024). If a model fails to retrieve these stable, well-documented, and openly available figures, it indicates a fundamental failure to serve as a source of public information.

3.2 Ground-Truth Dataset

To evaluate LLM responses, we used official Eurostat statistics as ground truth: *Population by educational attainment level, sex, age, and citizenship (%)* (code: `edat_lfs_9911`) (Eurostat, 2025). We extracted and transformed the subset for ages 25–54, disaggregated by sex (male, female) and citizenship (nationals, foreigners), for 2020–2023. Percentages are reported by country–year and sum to 100 across ISCED 2011 categories (ED0–2, ED3–4, ED5–8). Complete Eurostat distributions across all three education categories were available for at least one query configuration in all 34 countries for nationals and in 25 countries for foreigners. Table 1 presents illustrative Eurostat ground-truth rows for Switzerland in 2022.

Table 1: Example Eurostat ground-truth distributions (ages 25–54) for Switzerland. Values are percentages.

Year	Citizenship	Sex	Low	Medium	High
2022	Foreigner	F	26.2	25.9	47.8
2022	Foreigner	M	26.1	30.2	43.7
2022	National	F	6.9	47.6	45.5
2022	National	M	5.2	41.9	52.9

Source: Eurostat `edat_lfs_9911`.

3.3 LLM Models

We evaluate the responses of four LLMs: **GPT-4 Turbo, GPT-4o, GPT-5 chat, and GPT-5.1** (OpenAI, 2023, 2024, 2025a, 2025b). We focus on the OpenAI GPT family as it continues to dominate the generative web user space, with ChatGPT responsible for the vast majority of user-facing AI interactions (Kemp, 2025). To ensure a reproducible, controlled evaluation, we query the underlying API rather than the chat interface. GPT-4o and GPT-5 chat APIs served as the direct backend engines for ChatGPT, so their API responses approximate the underlying model behavior available through the web interface while avoiding interface-level variability.

3.4 Prompt Design and Evaluation

We use a single controlled zero-shot prompt so that query wording remains constant across country, year, sex, and citizenship conditions. Through iterative refinement via the ChatGPT web interface, we arrived at a final prompt that constrains each query to a single data row (one country, one year, one sex, one citizenship) to reduce token limit errors and digressions:

System Prompt

```
You are a scientific assistant
```

User Prompt

```
Give me the distribution of {sex}
{citizenship} in {country} in the age group
25-54, with educational attainment level
ED0-2, ED3-4 and ED5-8 according to ISCED
2011 in the year {year}. Answer only
numerical values to two decimals separated
by commas in the order of ED0-2, ED3-4,
ED5-8, and no additional text.
```

We queried each model via OpenAI's API with temperature = 0 and capped response length to constrain responses to the three percentages (`max_tokens = 25` for GPT-4o and GPT-5 chat

and 30 for GPT-4 Turbo and GPT-5.1). Prompt parameters—sex (Male, Female), citizenship (Foreign, National), year (2020–2023), and country (34 countries, listed in Appendix A)—were iterated over all combinations, yielding 544 unique query configurations, where each configuration represents one country–year–sex–citizenship combination. Of these 544 query configurations, 447 had complete Eurostat reference values across all three education categories (some countries provided only nationals data but not foreigners) and could therefore be used for numerical comparison with the ground truth: 189 concerned foreigners and 258 concerned nationals. For GPT-4 Turbo, we excluded 2023 from analyses that compare model outputs with Eurostat data because it falls outside the model's documented knowledge period. Response behavior analyses that do not depend on numerical agreement with Eurostat use all submitted queries from 2020–2023. Of these 337 configurations for GPT-4 Turbo, 175 produced at least one numerical response and contributed to the numerical comparison in Table 2; the remaining 162 produced explicit non-answers in all ten repetitions.

We repeated each query configuration 10 times to avoid relying on a single response as indicative of model behavior, following recent work showing that single-response evaluations are highly variable and inadequate for assessing representational bias consistency (Seth et al., 2025). For each query configuration, we averaged numerical responses across the 10 repetitions; we excluded non-numeric responses from the numerical analysis. At the individual-response level, we distinguish three response outcomes: an *explicit non-answer*, an *invalid percentage distribution*, and a *valid percentage distribution*. An invalid percentage distribution contains three parseable numerical values but violates the required percentage constraints: each value must lie between 0 and 100, and the three values must sum to 100 within a tolerance of 0.1 percentage points. We summarize response behavior across all submitted queries, whereas we evaluate numerical accuracy only for queries for which Eurostat reports all three education-category percentages. Invalid numerical distributions were retained in the numerical accuracy analysis because they are model-generated numerical outputs; RMSE and TVD were therefore computed from all numerical responses, including compositionally invalid ones.

Separately, at the query-configuration level, we classified the ten repetitions as *always non-answer*, *mixed*, or *always numerical*. The repetitions characterize within-configuration variability. To assess factual reliability, we compare LLM outputs to an Additive White Gaussian Noise (AWGN) baseline created by adding small random noise (zero-mean Gaussian, $\sigma = 1$ percentage point) to the Eurostat ground-truth values. This baseline quantifies whether LLM errors are meaningfully larger than those produced by random perturbations of the correct data.

3.5 Evaluation Measures

To assess how closely the LLM-provided statistics align with Eurostat values, we visually examined distribution plots and computed three quantitative measures:

Distribution plots: We superimpose the kernel density estimates of the LLM-provided data, the ground truth, and the AWGN baseline to assess the differences in their data distributions.

Root Mean Square Error (RMSE): We compute the RMSE to quantify the magnitude of deviation between the LLM-provided percentages and the Eurostat ground truth, and separately, with the AWGN baseline. Because the values are expressed on a 0–100 percentage scale, RMSE is reported in percentage points.

Cosine Similarity: We measure the Cosine Similarity between the LLM-provided vector and the ground truth vector (and analogously for AWGN) to assess structural alignment independent of magnitude.

Total Variation Distance (TVD): We additionally compute TVD between the LLM-provided and Eurostat distributions, $\mathrm{TVD}(p, q) = \frac{1}{2}\sum_{i=1}^{3} |p_i - q_i|$. Because p and q are expressed on a 0–100 percentage scale, TVD is reported in percentage points and represents the amount of percentage mass that would need to be redistributed across the three education categories for the generated distribution to match Eurostat.

4 Results

4.1 Evaluation of Statistical Reliability (RQ1)

For query configurations with at least one numerical response, we assessed agreement with Eurostat using RMSE, TVD, and cosine similarity (Table 2). The AWGN reference produced an RMSE of 0.99 percentage points and TVD of 1.21 percentage points. In contrast, GPT-model RMSE ranged from 10.36 to 14.32 percentage points, and TVD ranged

Table 2: Agreement between GPT-generated and Eurostat educational-attainment distributions for query configurations with complete Eurostat reference values. RMSE and TVD are reported in percentage points; lower values indicate closer agreement. Cosine similarity measures directional similarity. N is the number of complete-reference query configurations that produced at least one numerical response and therefore contributed to the numerical comparison. GPT-4 Turbo uses 2020–2022; the other models use 2020–2023.

Model	Measure	Total	Foreign	National	Female	Male
AWGN baseline ($N = 447$)	RMSE	0.99	0.99	0.99	0.99	0.99
	TVD	1.21	1.16	1.23	1.18	1.24
	Cosine	0.99	0.99	0.99	0.99	0.99
GPT-4 Turbo ($N = 175$)	RMSE	14.32	21.10	10.93	15.74	12.68
	TVD	16.37	26.06	12.91	18.08	14.59
	Cosine	0.953	0.935	0.959	0.942	0.965
GPT-4o ($N = 374$)	RMSE	10.80	11.50	10.40	11.36	10.19
	TVD	12.87	14.18	12.17	13.58	12.14
	Cosine	0.959	0.952	0.962	0.954	0.963
GPT-5 chat ($N = 447$)	RMSE	10.42	12.56	8.52	11.11	9.68
	TVD	12.45	15.55	10.18	13.60	11.29
	Cosine	0.961	0.941	0.975	0.956	0.965
GPT-5.1 ($N = 447$)	RMSE	10.36	13.65	7.02	10.44	10.28
	TVD	11.88	17.02	8.12	12.10	11.66
	Cosine	0.961	0.930	0.983	0.960	0.961

from 11.88 to 16.37 percentage points. Cosine similarity remained above 0.93, indicating directional similarity, but we interpret these values cautiously because the compared vectors are three-part percentage distributions constrained to approximately sum to 100. However, disaggregated analysis reveals disparities. We characterize this combination of structural plausibility and numerical inaccuracy as the illusion of competence.

Paired Wilcoxon tests (Table 6 in Appendix B) showed differences between foreigners and nationals for RMSE, TVD, and cosine similarity for all four models (all $p < .001$). Differences across females and males were less consistent: GPT-4o and GPT-5 chat differed on all three measures, GPT-4 Turbo differed on RMSE and cosine similarity but not TVD, and GPT-5.1 differed on none of the three measures. As the subject of analysis is a percentage of citizens, a close similarity score does not guarantee that the generated data is fully reliable for practical use. Instead, it indicates alignment with the Eurostat dataset.

Figure 1 compares the aggregated distributions of educational attainment levels, overlaying GPT-generated, AWGN, and Eurostat data. Across all models, the GPT distributions show sharper peaks and less variance than the ground truth. Unlike the broad, dispersed Eurostat distribution—which reflects the diverse educational backgrounds of European residents—the GPT distributions are narrow and concentrated. The models effectively output a stereotypical average for each country, ignoring socio-economic variation across countries. This suggests partial structural alignment but not accuracy.

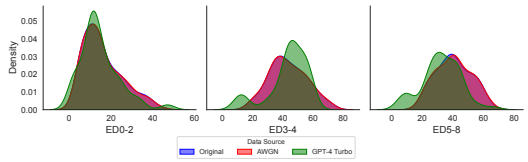
Distributions by year (Appendix C) show consistent deviation patterns across the evaluated years, indicating that the errors are not limited to particular time periods. These findings suggest that newer models prioritize coverage, raising concerns about potential misinformation when used in public policy or other high-stakes contexts.

4.2 Response Behavior and Refusals/Explicit Non-answers (RQ2)

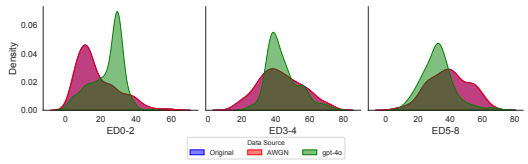
The most immediate failure mode is selective non-responsiveness. Table 3 reports the percentage of repeated responses that were explicit non-answers. GPT-5 chat and GPT-5.1 produced numerical responses in every repetition, whereas 77.5% of GPT-4 Turbo responses and 29.6% of GPT-4o responses were explicit non-answers.

Table 3: Explicit non-answer (Refusal) rates (%) across repeated responses, by citizenship and sex. Rates are calculated over individual model responses rather than query configurations.

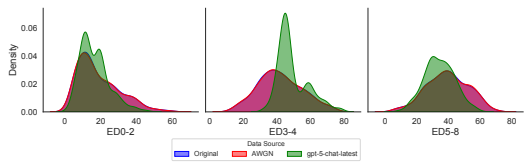
Model	Total	Foreign	National	F	M
GPT-4 Turbo	77.5	91.0	64.1	77.9	77.1
GPT-4o	29.6	45.9	13.4	31.1	28.1
GPT-5 chat	0.0	0.0	0.0	0.0	0.0
GPT-5.1	0.0	0.0	0.0	0.0	0.0



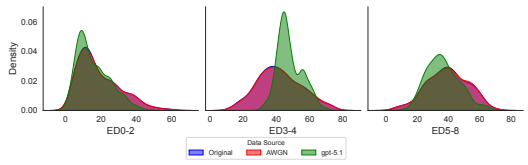
(a) GPT-4 Turbo



(b) GPT-4o



(c) GPT-5 chat



(d) GPT-5.1

Figure 1: **Distributions of GPT-generated, AWGN-perturbed, and Eurostat educational attainment percentages for query configurations with complete Eurostat reference values.** Panels (a)–(d) show GPT-4 Turbo, GPT-4o, GPT-5 chat, and GPT-5.1, respectively; within each panel, the three plots correspond to ED0–2, ED3–4, and ED5–8. Across models, GPT-generated values form narrower, more concentrated distributions than the Eurostat reference values, whereas the AWGN baseline remains close to Eurostat. GPT-4 Turbo uses 2020–2022; the other models use 2020–2023.

Explicit non-answers were not uniformly distributed across demographic groups. For GPT-4o, 45.9% of responses to foreign-citizen query configurations were explicit non-answers, compared with 13.4% for national-citizen queries. For GPT-4o, the observed non-answer rate was 3.4 times higher for foreign-citizen queries than for national-citizen queries. For GPT-4 Turbo, 91.0% of responses to foreign-citizen query configurations were explicit non-answers. Both GPT-4 models also had higher explicit non-answer rates among responses to female than male query configurations. Chi-square tests of independence (Table 5 in Appendix B) confirmed significant refusal-rate differences across citizenship for both GPT-4 models, and across sex for GPT-4o.

GPT-5 models show no explicit non-answers across all query configurations (0.0%). This represents a shift from withholding responses under uncertainty to generating a response for every query regardless of data availability. As Section 4.1 shows, eliminating refusals does not indicate improved knowledge; it indicates that non-response has been replaced by hallucination.

The repeated calls also reveal differences in response stability by query configuration (Table 4 in Appendix B). For GPT-4 Turbo, 162 of 337 query configurations (48.1%) consistently produced explicit non-answers, 133 (39.5%) alternated between numerical responses and non-answers across the ten repetitions, and only 42 (12.5%) consistently produced numerical responses. GPT-4o was more stable: 73 of 447 configurations (16.3%) consistently produced non-answers, 84 (18.8%) showed mixed behavior, and 290 (64.9%) consistently produced numerical responses. GPT-5 chat and GPT-5.1 produced numerical responses in all ten repetitions for all 447 configurations. Thus, the earlier models' non-answer behavior was not always deterministic, even within the same query configuration, where we held all prompt attributes constant.

We further broke down the results for foreigners by sex to examine whether foreign women showed a stronger disadvantage than foreign men. The descriptive patterns did not show a consistent additional disadvantage for foreign women across models (Table 7 in Appendix B). The differences between foreigners and nationals were also broadly distributed across countries rather than driven by a small number of cases. Among the 25 countries with comparable Eurostat data for both foreigners and nationals, the non-answer rate was higher for foreigners in 24 countries for GPT-4 Turbo and all 25 countries for GPT-4o. RMSE was higher for foreigners in all 18 countries, with comparable numerical responses for GPT-4 Turbo, 15 of 25 for GPT-4o, 21 of 25 for GPT-5 chat, and 22 of 25 for GPT-5.1 (Table 8 in Appendix B).

5 Discussion

5.1 Illusion of Competence (RQ1)

Our evaluation shows that GPT models capture broad trends in migration statistics (high cosine similarity) but fail to reproduce exact or close-to-exact values (high RMSE). This observation aligns with prior studies showing that LLMs perform poorly on numerical data (Ji et al., 2023; Huang et al., 2025) and tend to neglect rare or detailed

figures while focusing on frequent or prominent facts (Balasuriya Mudiyanselage et al., 2025; Seth et al., 2025; Batzner et al., 2025; Leidinger and Rogers, 2025; Atif et al., 2025).

The distributional analysis reveals a specific pattern. Visual inspection confirms that the models output a narrow, concentrated distribution where the ground truth is broad and dispersed. The models effectively replace the complex variance of migrant populations across 34 European countries with a single stereotypical average. This behavior echoes prior work showing that LLMs portray socially subordinate groups as more homogeneous (Lee et al., 2024), and our findings extend this observation from qualitative text generation to structured numerical outputs. The parallel to a core concern in pluralism-aware NLP is that just as aggregation-based annotation pipelines collapse annotator disagreement into a single gold label, these models collapse heterogeneous population-level statistics into a homogeneous consensus output. As a result, the cross-national and subgroup variation that gives migration data its policy relevance is flattened into a single plausible-looking but misleading average.

This pattern connects to prior work on the *illusion of competence*, which shows that AI systems can appear more capable than their actual performance warrants. Our study observes a related phenomenon at the output level rather than measuring user perception: the models often produce well-formed statistical distributions whose structure appears credible despite substantial factual error.

5.2 Group-Differential Non-answer Behavior (RQ2)

The refusal behavior observed in GPT-4 models, but absent in GPT-5 models, reveals a shift in alignment from conservative to permissive. The refusal patterns were consistent across years and countries for both GPT-4o and GPT-4 Turbo. The models' refusal behavior could appear as an implicit abstention policy, but one that is neither calibrated nor group-aware. The observed non-answers are more frequent for questions about foreigners than for questions about nationals, creating uneven response availability across the populations represented in the queries. Rather than abstaining transparently when epistemic uncertainty is high, the models abstain selectively along demographic lines, withholding information about the groups for whom accurate public data is arguably most needed. However, our evaluation does not identify the mechanisms responsible for these differences. The observed patterns could arise from differences in model training, information availability, or other model- and system-level factors. We therefore interpret them as differences in observable response behavior rather than evidence of particular safety, alignment, or policy mechanisms.

This finding carries direct implications for alignment with principles of fairness and inclusivity. If some non-answers stem from safety mechanisms, designed to prevent harmful outputs but applied too broadly, they produce two concrete harms: (1) missing valid socio-demographic statistics for at-risk groups, and (2) treating public information queries about these groups as implicitly "unsafe." Our findings align with recent evidence that LLM refusals are unevenly triggered by particular social categories (Leidinger and Rogers, 2025; Atif et al., 2025) and extend this observation to structured, factual queries about official statistics.

GPT-5 models resolve the problem by eliminating refusals rather than by improving the underlying knowledge. Neither behavior serves users equitably. A model that refuses to answer questions about foreign citizens denies access to public information; a model that fabricates answers to those same questions provides misinformation. Both outcomes disproportionately affect the populations the queries concern.

5.3 Implications

Our findings have direct implications for using LLMs in public-facing applications, digital government services, and public information tools. On one hand, LLMs offer an accessible interface for synthesizing and summarizing complex or distributed information. On the other hand, they blur the line between sourced and synthesized content. Whether through explicit non-answers or hallucination, LLM responses are problematic. The practical concern is not that experts would rely on LLMs as authoritative statistical databases, but that the wider adoption of LLMs in information retrieval may expose researchers, policymakers, journalists, and citizens to plausible but unreliable public-interest data. Fabricated migration statistics could create false impressions about the scale or nature of migration, reinforcing negative stereotypes and stigma (Ahmed et al., 2024). Unreliable outputs risk undermining trust in both LLMs and the official statistics they purport to reproduce.

For practitioners and international agencies working with census and survey data, our results indicate that LLMs are not reliable standalone information sources. Fine-tuning may improve factual performance (Brand et al., 2023; El Hamdani et al., 2024), while retrieval-augmented generation (RAG) can connect models directly to external sources (Lewis et al., 2020); however, we did not evaluate either approach and therefore cannot determine whether they reduce the group-differential patterns observed here. Our findings also have important implications for policymakers and stakeholders involved in migration issues and AI regulation. For the former, relying on LLMs for quick data or briefings could affect assessments of real situations, misguiding interventions and policies. For the latter, it supports the need for transparency and accountability in LLM outputs. The EU AI Act (European Parliament and Council of the European Union, 2024) contains transparency mandates so that users know when they interact with AI. In addition, it will be important to classify migration or socio-economic data generation as a high-risk application, given its social impact. Our finding that response availability differs across demographic groups highlights the need for greater transparency about when and why LLMs decline factual public-information queries. Adopting LLMs in high-stakes social contexts requires not only model optimization but also alignment with ethical principles, and stakeholders (from citizens to policymakers) must remain critically aware of model behaviors and limitations.

5.4 Limitations and Future Work

Our study has a few limitations that open directions for future work. First, we evaluated only OpenAI's GPT models, and other LLMs (open-source or from different providers) may behave differently. Second, we focused on educational attainment; however, migration data are multidimensional, and our findings may not reflect all aspects of migration statistics. Third, alternative prompting strategies like few-shot examples or chain-of-thought might yield different results and could test whether models possess latent knowledge they fail to surface. Fourth, our study evaluates LLMs as standalone information sources; RAG and grounded agentic systems represent a growing shift in LLM use, and evaluating these systems is a necessary next step. Fifth, we observe differences across citizenship groups but cannot determine their cause and interpret these results as group-differential output behavior rather than evidence of a particular underlying mechanism. Our findings would further motivate evaluation of such grounded systems. Although we note these limitations, they do not challenge the key findings of our analysis, and the proposed future work would deepen our understanding of them.

6 Conclusion

This study evaluated whether standalone GPT models can reliably serve as sources for migration-related education statistics, comparing outputs from four GPT models against official Eurostat data across 34 European countries.

Addressing **RQ1**, our quantitative evaluation shows that all four models remain unreliable sources of the requested statistics. Numerical responses preserve the broad structure of plausible statistical distributions while deviating substantially from Eurostat: cosine similarity remains high (>0.93), while RMSE exceeds 10 percentage points and TVD ranges from approximately 12–16 percentage points across the four models. We characterize this combination of structural plausibility and factual error as an output-level analog of the *illusion of competence*: answers can appear statistically well-formed even when the underlying values are inaccurate.

Addressing **RQ2**, model response behavior differs across the demographic groups described in the queries. GPT-4 Turbo and GPT-4o provide explicit non-answers substantially more often for foreign-citizen than national-citizen queries. When numerical responses are produced, errors are generally larger for foreign-citizen queries, and GPT-4 Turbo disproportionately produces invalid percentage distributions for these queries. GPT-5 chat and GPT-5.1 provide numerical responses consistently, but greater coverage does not eliminate substantial factual error. Information availability and numerical reliability therefore vary not only across model generations, but also across the populations being represented.

Together, these findings clarify an important distinction for pluralistic and responsible NLP. Legitimate variation can arise from differences in human perspectives, contexts, definitions, and interpretations, whereas unsupported variation in a reference-grounded statistic is factual error rather than an alternative perspective. Evaluations of AI-mediated public-information systems should therefore assess

both whether models reproduce authoritative evidence reliably and whether that information remains comparably available across the populations it describes.

Acknowledgments and AI use disclosure

This work was supported by the European Union's Horizon Europe program through the ELIAS project (under grant agreement No. 101120237). We used Visual Studio Code to conduct data analysis. We used its integrated GitHub Copilot functionality to provide code-completion suggestions. We used Grammarly to check spelling, grammar, and punctuation. However, Grammarly could also use AI to suggest improvements to phrasing and clarity. Where we accepted suggestions from GitHub Copilot or Grammarly, we reviewed them and, where necessary, adapted them before using them.

References

Saifuddin Ahmed, Kokil Jaidka, Vivian Hsueh Hua Chen, Mengxuan Cai, Anfan Chen, Claire Stravato Emes, Valerie Yu, and Arul Chib. 2024. Social media and anti-immigrant prejudice: a multi-method analysis of the role of social media use, threat perceptions, and cognitive ability. *Frontiers in Psychology*, 15.

Ruth Appel, Peter McCrory, Alex Tamkin, Miles McCain, Tyler Neylon, and Michael Stern. 2025. Anthropic economic index report: Uneven geographic and enterprise ai adoption. *Preprint*, arXiv:2511.15080.

Farah Atif, Nursultan Askarbekuly, Kareem Darwish, and Monojit Choudhury. 2025. Sacred or synthetic? evaluating LLM reliability and abstention for religious questions. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):217–226.

Uvini Balasuriya Mudiyanselage, Bharat Jayprakash, Kookjin Lee, and K. Hazel Kwon. 2025. Disaggregated health data in LLMs: Evaluating data equity in the context of asian american representation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):292–303.

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2025. Germanpartiesqa: Benchmarking commercial large language models and AI companions for political alignment and sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):330–342.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 39–43, Miami, Florida, USA. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using LLMs for market research. Technical Report 23-062, Harvard Business School Marketing Unit.

Jonathan Bright, Florence Enock, Saba Esnaashari, John Francis, Youmna Hashem, and Deborah Morgan. 2025. Generative AI is already widespread in the public sector: evidence from a survey of UK public sector professionals. *Digit. Gov.: Res. Pract.*, 6(1).

Tatiana Chakravorti, Xinyu Wang, Pranav Narayanan Venkit, Sai Koneru, Kevin Munger, and Sarah Rajtmajer. 2025. Social scientists on the role of AI in research. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):528–540.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research.

Hein de Haas. 2024. Changing the migration narrative: On the power of discourse, propaganda and truth distortion. IMI Working Paper No. 181 / PACES Project Working Paper No. 3 181, International Migration Institute (IMI), University of Amsterdam, Amsterdam.

Milena de Swart, Floris Den Hengst, and Jieying Chen. 2025. Detecting linguistic bias in government documents using large language models. In *Proceedings of the ACM on Web Conference 2025*, Www '25, page 5034–5044, New York, NY, USA. Association for Computing Machinery.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rajaa El Hamdani, Thomas Bonald, Fragkiskos D. Malliaros, Nils Holzenberger, and Fabian Suchanek. 2024. The factuality of large language models in the

legal domain. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, Cikm '24, page 3741–3746, New York, NY, USA. Association for Computing Machinery.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union.

Eurostat. 2024. EU Labour Force Survey – Methodology. *Statistics Explained*. Last updated 13 March 2024.

Eurostat. 2025. Population by educational attainment level, sex, age and citizenship. Accessed: 2024-11-11.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Hideaki Joko, Shakiba Amirshahi, Charles L. A. Clarke, and Faegheh Hasibi. 2026. Wildclaims: Conversational information access in the wild(chat). In *Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 – April 2, 2026, Proceedings, Part IV*, page 367–382, Berlin, Heidelberg. Springer-Verlag.

Simon Kemp. 2025. Digital 2026: One billion people using AI. `https://datareportal.com/reports/digital-2026-one-billion-people-using-ai`. Accessed: 2025-11-23.

Aparup Khatua. 2025. Policies for refugee mental health using generative agents. In *Companion Proceedings of the ACM on Web Conference 2025*, Www '25, page 1062–1066, New York, NY, USA. Association for Computing Machinery.

Claudia Lauer. 2025. Judge's note on immigration agents using AI raises accuracy and privacy concerns. Accessed 30 Nov 2025.

Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1321–1340, New York, NY, USA. Association for Computing Machinery.

Sunok Lee, Dasom Choi, Lucy Truong, Nitin Sawhney, and Henna Paakki. 2025. Into the unknown: Leveraging conversational AI in supporting young migrants' journeys towards cultural adaptation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Chi '25, New York, NY, USA. Association for Computing Machinery.

Alina Leidinger and Richard Rogers. 2025. How are LLMs mitigating stereotyping harms? learning from search engine studies. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, page 839–854. AAAI Press.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

OpenAI. 2023. GPT-4 Turbo. `https://platform.openai.com/docs/models/gpt-4-turbo`.

OpenAI. 2024. GPT-4o. `https://platform.openai.com/docs/models/gpt-4o`.

OpenAI. 2025a. GPT-5 Chat. `https://platform.openai.com/docs/models/gpt-5-chat-latest`.

OpenAI. 2025b. GPT-5.1. `https://platform.openai.com/docs/models/gpt-5.1`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nate Sanford. 2025. Washington city officials are using ChatGPT for government work. Accessed 30 Nov 2025.

Wolfgang S. Schmeisser-Nieto, Pol Pastells, Simona Frenda, and Mariona Taule. 2024. Human vs. machine perceptions on immigration stereotypes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), pages 8453–8463, Torino, Italia. ELRA and ICCL.

Agrima Seth, Monojit Choudhury, Sunayana Sitaram, Kentaro Toyama, Aditya Vashistha, and Kalika Bali. 2025. How deep is representational bias in LLMs? the cases of caste and religion. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3):2319–2330.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Chi '25, New York, NY, USA. Association for Computing Machinery.

Luca Tangi, Marco Combetto, Isabelle Hupont Torres, Eimear Farrell, and Sven Schade. 2024. The potential of generative AI for the public sector: current use, key questions and policy considerations. Technical report, Joint Research Centre, Ispra (Italy).

UK Home Office. 2025. Evaluation of AI trials in the asylum decision making process. Research and Analysis. Accessed 30 Nov 2025.

UNESCO Institute for Statistics. 2012. *International Standard Classification of Education: ISCED 2011*. Unesco, Montreal.

Radhakrishnan Venkatakrishnan, Emrah Tanyildizi, and M. Abdullah Canbaz. 2024. Semantic interlinking of immigration data using LLMs for knowledge graph construction. In *Companion Proceedings of the ACM Web Conference 2024*, Www '24, page 605–608, New York, NY, USA. Association for Computing Machinery.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Shanshan Xu, Santosh T.Y.S.S, and Barbara Plank. 2026. Position: From noise to signal to selbstzweck - reframing human label variation in the era of post-training in NLP. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 23762–23772, San Diego, California, United States. Association for Computational Linguistics.

Ahmad Wali Ahmad Yar and Tuba Bircan. 2025. New frontiers in migration statistics: a narrative review on big data's role. *Frontiers in Human Dynamics*, 7.

Danna Zheng, Mirella Lapata, and Jeff Z. Pan. 2024. How reliable are LLMs as knowledge bases? rethinking facutality and consistency. *Preprint*, arXiv:2407.13578.

A Country List

No.	Country	No.	Country
1.	Austria	18.	Latvia
2.	Belgium	19.	Lithuania
3.	Bosnia and Herzegovina	20.	Luxembourg
4.	Bulgaria	21.	Malta
5.	Croatia	22.	Montenegro
6.	Cyprus	23.	Netherlands
7.	Czech Republic	24.	North Macedonia
8.	Denmark	25.	Norway
9.	Estonia	26.	Poland
10.	Finland	27.	Portugal
11.	France	28.	Romania
12.	Germany	29.	Serbia
13.	Greece	30.	Slovakia
14.	Hungary	31.	Slovenia
15.	Iceland	32.	Spain
16.	Ireland	33.	Sweden
17.	Italy	34.	Switzerland

B Tables

Table 4: Response stability across ten repeated calls. N is the number of unique query configurations, not the number of API responses. *Non-answer* indicates that all ten repetitions produced an explicit non-answer; *Numerical* indicates that all ten produced numerical responses; *Mixed* indicates both outcomes across repetitions. GPT-4 Turbo uses complete-reference configurations from 2020–2022; the other models use 2020–2023.

Model	N	Non-answer	Mixed	Numerical
GPT-4 Turbo	337	162 (48.1%)	133 (39.5%)	42 (12.5%)
GPT-4o	447	73 (16.3%)	84 (18.8%)	290 (64.9%)
GPT-5 chat	447	0 (0.0%)	0 (0.0%)	447 (100%)
GPT-5.1	447	0 (0.0%)	0 (0.0%)	447 (100%)

Table 5: Chi-square tests of independence for explicit non-answer behavior across citizenship and sex for GPT-4 Turbo and GPT-4o.

Model	Effect	$\chi^2(1)$	p-value
GPT-4 Turbo	Citizenship	562.10	$2.93 \times 10^{-124***}$
GPT-4 Turbo	Sex	0.4219	0.516
GPT-4o	Citizenship	687.35	$1.68 \times 10^{-151***}$
GPT-4o	Sex	5.784	0.016

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 6: Wilcoxon signed-rank tests comparing RMSE, TVD, and cosine similarity between paired citizenship groups and paired sexes. Tests are conducted on paired query configurations. RMSE and TVD measure error, whereas cosine similarity measures directional similarity.

Model	**Measure**	**Foreigners vs. nationals**		**Female vs. male**	
		W	p	W	p
GPT-4 Turbo	RMSE	25	1.28×10^{-11}***	1126	0.0265*
	TVD	27	3.58×10^{-11}***	1237	0.0937
	Cosine	842	3.56×10^{-4}***	975	0.00311**
GPT-4o	RMSE	2319	2.08×10^{-6}***	6179	0.00689**
	TVD	2201	1.09×10^{-6}***	6223	0.00832**
	Cosine	6594	9.09×10^{-8}***	6205	0.00770**
GPT-5 chat	RMSE	3906	8.22×10^{-12}***	8543	6.00×10^{-5}***
	TVD	3937	2.18×10^{-11}***	7844	2.24×10^{-6}***
	Cosine	14175	2.56×10^{-12}***	8783	1.76×10^{-4}***
GPT-5.1	RMSE	1946	4.94×10^{-21}***	11426	0.321
	TVD	1690	3.76×10^{-22}***	11378	0.297
	Cosine	16057	2.70×10^{-21}***	11399	0.308

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

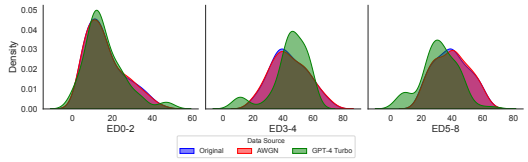
Table 7: Descriptive breakdown of response behavior and numerical error by citizenship and sex. Explicit non-answer rates are computed across the ten repeated responses for query configurations with complete Eurostat reference values. Invalid percentage distributions are reported as a percentage of numerical responses. For each query configuration with at least one numerical response, RMSE is computed from the numerical responses averaged across repetitions. N is the number of query configurations with at least one numerical response and therefore excludes configurations that produced explicit non-answers in all ten repetitions. GPT-4 Turbo uses 2020–2022; the other models use 2020–2023.

Model	**Group**	**Non-answer (%)**	**Invalid among numerical (%)**	N **(configurations with numerical response)**	**RMSE**
GPT-4 Turbo	Foreign females	87.4	45.7	25	22.44
	Foreign males	90.3	51.5	21	19.38
	National females	53.0	0.7	64	12.17
	National males	49.4	2.9	65	9.56
GPT-4o	Foreign females	39.4	0.0	69	11.54
	Foreign males	42.0	0.0	62	11.46
	National females	16.4	0.0	121	11.25
	National males	10.7	0.0	122	9.48
GPT-5 chat	Foreign females	0.0	0.0	96	12.52
	Foreign males	0.0	0.0	93	12.60
	National females	0.0	0.0	129	9.93
	National males	0.0	0.0	129	6.83
GPT-5.1	Foreign females	0.0	0.4	96	13.46
	Foreign males	0.0	0.3	93	13.84
	National females	0.0	0.0	129	7.42
	National males	0.0	0.0	129	6.6

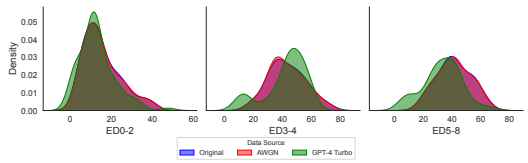
Table 8: Country-level differences between queries about foreigners and nationals. Values are computed as *foreigners minus nationals*. Positive values therefore indicate higher non-answer rates or greater numerical error for foreigners, whereas negative values indicate higher rates or error for nationals. Non-answer and RMSE gaps are reported in percentage points. A dash indicates that an RMSE gap could not be computed because one citizenship group received no numerical response for that country. GPT-5 chat and GPT-5.1 non-answer columns are omitted because both models returned numerical responses for all queries. GPT-4 Turbo uses 2020–2022; the other models use 2020–2023.

	Non-answer gap (pp)		**RMSE gap (pp)**			
Country	**4 Turbo**	**4o**	**4 Turbo**	**4o**	**5 chat**	**5.1**
Austria	43.3	8.8	4.96	-4.76	3.69	7.58
Belgium	53.3	35.0	3.41	-2.02	1.87	7.49
Cyprus	48.3	45.0	7.28	-12.85	-5.21	0.05
Czech Republic	-3.3	42.5	17.25	5.01	1.66	4.54
Denmark	36.7	17.5	7.33	-2.02	3.51	10.44
Estonia	48.3	22.5	19.16	0.26	2.82	1.37
Finland	41.7	36.2	4.90	-4.37	5.08	4.56
France	13.3	26.2	—	-4.68	1.00	2.85
Germany	46.7	21.3	8.56	-2.53	9.93	8.31
Greece	46.7	45.0	—	0.01	15.03	11.99
Hungary	46.7	12.9	—	5.49	4.85	7.67
Iceland	20.0	23.7	20.10	0.25	2.44	1.46
Ireland	26.7	22.5	15.16	5.56	2.51	11.71
Italy	48.3	30.0	—	7.11	3.12	6.48
Latvia	51.7	17.5	—	6.71	-0.83	8.31
Luxembourg	38.3	18.8	12.95	8.55	13.59	19.75
Malta	30.0	41.2	15.59	5.48	1.83	-2.61
Netherlands	20.0	36.2	4.26	4.55	8.64	13.50
Norway	35.0	16.2	0.01	-1.99	4.89	9.00
Portugal	63.3	13.8	1.08	-1.51	-5.37	-3.34
Serbia	50.0	52.5	—	2.88	5.02	3.34
Slovenia	55.0	25.0	12.54	2.58	0.23	3.78
Spain	48.3	25.0	—	0.58	-4.77	-3.84
Sweden	40.0	26.2	12.61	5.92	10.65	14.66
Switzerland	10.0	1.2	3.69	-2.50	5.72	10.15

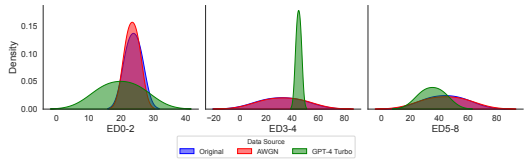
C Distribution plots



(a) GPT-4 Turbo - 2020

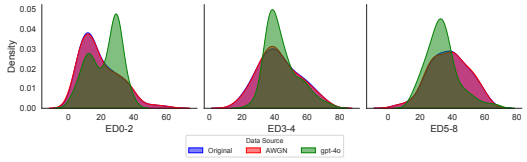


(b) GPT-4 Turbo - 2021

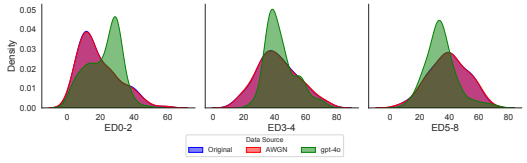


(c) GPT-4 Turbo - 2022

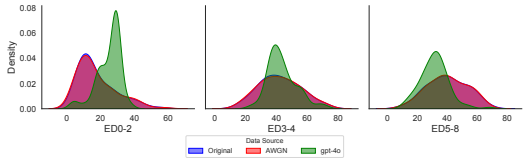
Figure 2: **Year-specific distributions of GPT-4 Turbo, AWGN-perturbed, and Eurostat educational-attainment percentages for query configurations with complete Eurostat reference values.** Panels (a)–(c) show 2020, 2021, and 2022, respectively; within each panel, the three plots correspond to ED0–2, ED3–4, and ED5–8. The narrower and more concentrated GPT-4 Turbo distributions recur across years, indicating that the deviation from Eurostat is not confined to a single year.



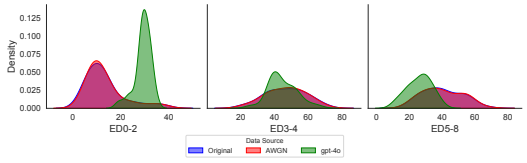
(a) GPT-4o - 2020



(b) GPT-4o - 2021

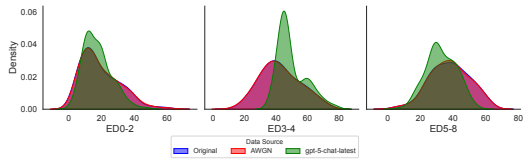


(c) GPT-4o - 2022

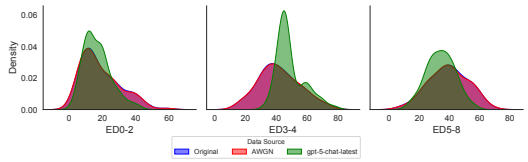


(d) GPT-4o - 2023

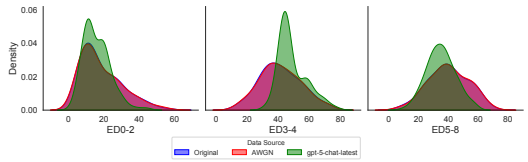
Figure 3: **Year-specific distributions of GPT-4o, AWGN-perturbed, and Eurostat educational-attainment percentages for query configurations with complete Eurostat reference values.** Panels (a)–(d) show 2020–2023, respectively; within each panel, the three plots correspond to ED0–2, ED3–4, and ED5–8. Across years, GPT-4o continues to produce distributions that differ from the broader Eurostat reference distributions, showing that the numerical deviations persist over time.



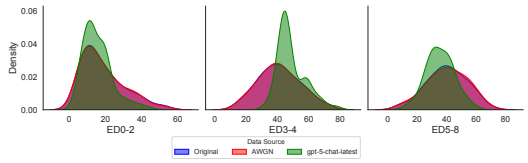
(a) GPT-5 chat - 2020



(b) GPT-5 chat - 2021

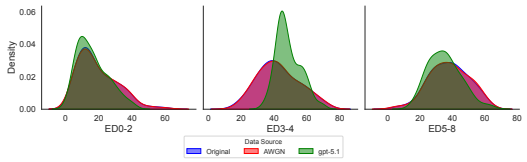


(c) GPT-5 chat - 2022

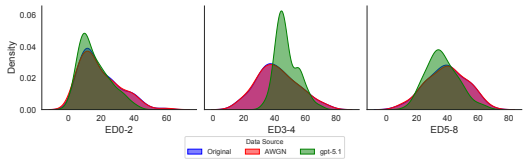


(d) GPT-5 chat - 2023

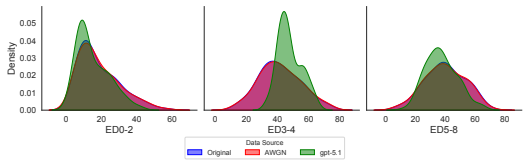
Figure 4: **Year-specific distributions of GPT-5 chat, AWGN-perturbed, and Eurostat educational-attainment percentages for query configurations with complete Eurostat reference values.** Panels (a)–(d) show 2020–2023, respectively; within each panel, the three plots correspond to ED0–2, ED3–4, and ED5–8. Despite producing numerical responses consistently, GPT-5 chat shows systematic distributional differences from Eurostat across all four years.



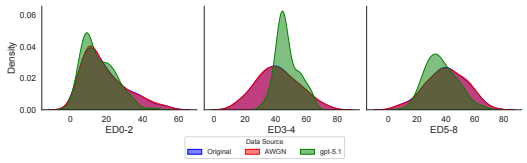
(a) GPT-5.1 - 2020



(b) GPT-5.1 - 2021



(c) GPT-5.1 - 2022



(d) GPT-5.1 - 2023

Figure 5: **Year-specific distributions of GPT-5.1, AWGN-perturbed, and Eurostat educational-attainment percentages for query configurations with complete Eurostat reference values.** Panels (a)–(d) show 2020–2023, respectively; within each panel, the three plots correspond to ED0–2, ED3–4, and ED5–8. GPT-5.1 also shows persistent differences from the Eurostat distributions across years, despite returning numerical responses for every tested query configuration.