LLMs Do Not See Age: Assessing Demographic Bias in Automated Systematic Review Synthesis

Favour Yahdii Aghaebe[*][†], **Tanefa Apekey**[‡], **Elizabeth Williams**[§][†], **Nafise Sadat Moosavi**[*]
University of Sheffield, UK
{fyaghaebe1, t.apekey, e.a.williams, n.s.moosavi}@sheffield.ac.uk

Abstract

Clinical interventions often hinge on age: medications and procedures safe for adults may be harmful to children or ineffective for older adults. However, as language models are increasingly integrated into biomedical evidence synthesis workflows, it remains uncertain whether these systems preserve such crucial demographic distinctions. To address this gap, we evaluate how well state-of-the-art language models retain age-related information when generating abstractive summaries of biomedical studies. We construct **DemogSummary**, a novel age-stratified dataset of systematic review primary studies, covering child, adult, and older adult populations. We evaluate three prominent summarisation-capable LLMs, Qwen (open-source), Longformer (open-source) and GPT-4.1 Nano (proprietary), using both standard metrics and a newly proposed **Demographic Salience Score (DSS)**, which quantifies age-related entity retention and hallucination. Our results reveal systematic disparities across models and age groups: demographic fidelity is lowest for adult-focused summaries, and under-represented populations are more prone to hallucinations. These findings highlight the limitations of current LLMs in faithful and bias-free summarisation and point to the need for fairness-aware evaluation frameworks and summarisation pipelines in biomedical NLP.

1 Introduction

The use of large language models (LLMs) to enhance the efficiency of scientific research and clinical practice has become increasingly common. In particular, LLMs have shown promise in accelerating labour-intensive processes such as systematic reviews. These models are now being explored at various stages of the review pipeline, including abstract identification and screening (Delgado-Chaves et al., 2025), offering the potential to reduce time and cost in evidence synthesis.

Despite these advances, concerns are emerging about the potential biases introduced by LLMs when applied to domains involving sensitive or underrepresented populations. Recent studies suggest that LLMs may exhibit demographic bias in their outputs (Kamruzzaman et al., 2024). In the context of systematic reviews, which often inform high-stakes decisions in medicine and policy, such biases pose serious risks. The process by which LLMs synthesise and preserve critical demographic information, specifically with narrative synthesis, where findings from individual studies are summarised, remains largely unexamined.

This presents a critical gap: while LLMs are increasingly used in summarising biomedical literature, little is known about their ability to retain demographic and population information, particularly age-related descriptors which are integral aspects of the systematic review. Misrepresentation or omission of such details can compromise both clinical relevance and health equity, reinforcing the very disparities such systematic reviews are designed to reduce.

Building on existing work that treats narrative synthesis as a form of multi-document summarisation (DeYoung et al., 2024; Wallace et al., 2021), we investigate how well LLMs preserve demographic integrity, specifically age-related descriptors, when approximating a systematic review abstract from the abstracts of included primary studies. To support this investigation, we construct DemogSummary, an age-stratified dataset of systematic reviews and primary studies grouped by population, and propose the Demographic Salience Score (DSS), a composite metric that quantifies entity-level retention, omission, and hallucination of age-related information. Using this framework,

[*] School of Computer Science
[†] Healthy Lifespan Institute
[‡] Sheffield Centre for Health and Related Research
[§] Department of Oncology and Metabolism

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1815–1833
December 20-24, 2025 ©2025 Association for Computational Linguistics

we assess the summarisation performance of three state-of-the-art LLMs across child, adult, and older adult populations. Our results reveal that fidelity to demographic information is not uniform. Summaries concerning adult populations show the lowest retention of age-related content and the highest incidence of hallucinated descriptors, while those focused on children and older adults show greater accuracy. Across models, GPT-4.1 Nano (OpenAI, 2023) and Longformer (Beltagy et al., 2020) demonstrate stronger demographic preservation than Qwen-2.5 (Yang et al., 2025), though all exhibit limitations in faithfully synthesising age-specific biomedical content. These findings highlight the need to move beyond generic evaluation metrics and toward assessments capturing demographic fidelity in biomedical summarisation. Our contributions are threefold:

- Presenting a novel age-stratified dataset of systematic review primary studies (**DemogSummary**), grouped by population (children, adults, older adults), designed to support demographic-specific evaluation of summarisation models in biomedical domains.
- Identifying systematic disparities in how LLMs preserve age-related information during multi-document summarisation, highlighting representational gaps that are obscured by conventional evaluation metrics.
- Introducing the *Demographic Salience Score*, a targeted metric that quantifies the retention, omission, and hallucination of demographic entities, enabling a fairness-aware assessment of summarisation fidelity.

2 Related Work

2.1 Automatic Systematic Reviews

Data extraction and synthesis represent the most resource-intensive and error-prone phases within the systematic review workflow, often requiring significant manual effort and rigour. Surveys of professionals involved in systematic reviews reinforce this notion; 45% of respondents in one study of 194 professionals identified data extraction as the most time-consuming stage (Scott et al., 2021). Consequently, there has been an increased focus on automating these steps, particularly using LLMs (Ge et al., 2024; Schmidt et al., 2023). While LLMs have shown promise in biomedical information extraction, their accuracy remains inconsistent across tasks and domains. More broadly, concerns have emerged about their tendency to amplify societal biases, especially in high-stakes fields like medicine. Biases in training data, model design, and linguistic priors can lead to unequal treatment of groups based on attributes such as race (Yang et al., 2024) and gender (Bajaj et al., 2024; Tang et al., 2024), resulting in harms like stereotyping and misassociation (Gallegos et al., 2024). Yet demographic attributes like *age* remain underexplored in bias evaluations, particularly in biomedical synthesis. Automatic summarisation selects and integrates key content from one or more documents to produce a concise, informative output (Nenkova and McKeown, 2011). This task is increasingly important in the biomedical domain, where the exponential growth of publications makes manual synthesis difficult to scale (Pawar et al., 2023). In particular, multi-document summarisation offers a promising way to approximate the narrative synthesis found in systematic reviews. As LLMs are increasingly adopted for this purpose, it becomes essential to examine not only their efficiency, but also how reliably and equitably they preserve critical population-specific information.

2.2 Bias in Automated Synthesis and Summarisation

Despite the promise of LLMs for multi-document biomedical summarisation, their outputs often reflect biases that can distort or omit critical demographic information. This is especially concerning in systematic reviews, where under-representation or misrepresentation of population groups may have direct clinical and policy implications. Prior work has examined bias in summarisation across domains such as news (Steen and Markert, 2024a), opinion generation (Huang et al., 2023, 2024), and radiology reports (Seyyed-Kalantari et al., 2021; Nguyen et al., 2024). However, biomedicine has received little attention, particularly in narrative synthesis of systematic reviews, where omissions or hallucinations of population-specific information can compromise the validity of clinical evidence. Fairness and bias have been long-standing concerns in NLP (Bolukbasi et al., 2016; Bender and Friedman, 2018; Blodgett et al., 2020; Tang et al., 2024), though their definitions are often inconsistent across tasks and domains. We draw on the framework posited by Crawford (2017), which dis-

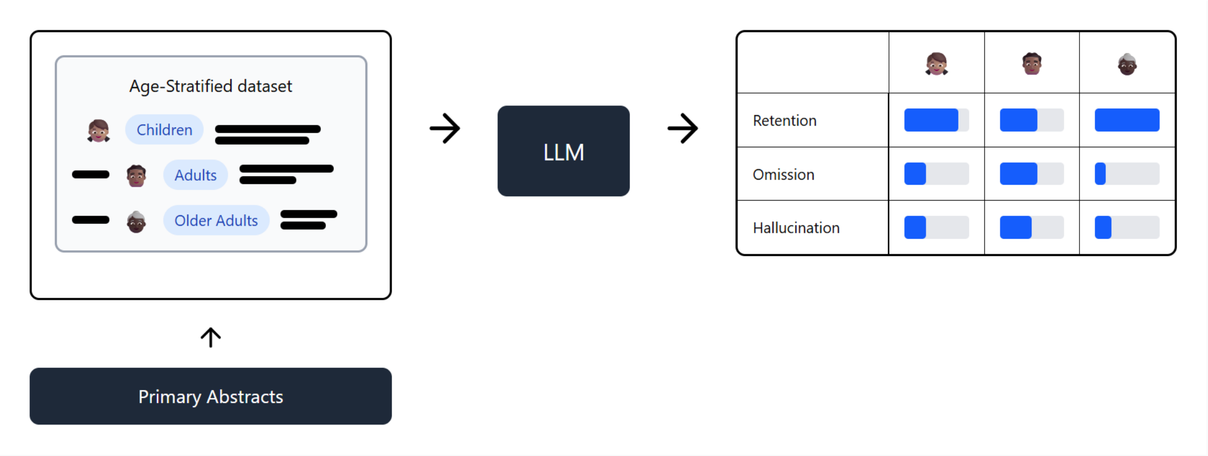


Figure 1: Age-stratified primary study abstracts are summarised by LLMs, and the outputs are compared to systematic review abstracts. Summaries are evaluated by age group for demographic fidelity using retention, omission, and hallucination metrics.

tinguishes between two forms of algorithmic bias: allocative bias, which affects access to resources, and representational bias, which misrepresents or excludes certain groups (Sun et al., 2019; Suresh and Guttag, 2021). This study focuses on representational bias, particularly the loss or distortion of age-related information in generated summaries. We adopt the definition of an unbiased summariser proposed by Steen and Markert (2024b), which emphasises faithful and complete representation of relevant input content.

Evaluating bias in summarisation presents several challenges. In certain domains, researchers can generate synthetic or controlled input texts to isolate the effects of bias in model outputs. In biomedical summarisation, however, such input manipulation is impractical, as real clinical studies must be used to reflect the actual conditions and constraints of systematic reviews. To address this, we curate **DemogSummary**, a real-world, age-stratified dataset that allows for population-specific evaluation without altering the original inputs. This design lets us examine representational bias as it occurs naturally, while controlling for demographic focus through corpus structure and targeted evaluation. Another challenge lies in how summaries are evaluated: standard automatic metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) measure surface or semantic similarity but fail to capture representational distortions or omissions (Deutsch et al., 2022; Gao and Wan, 2022). To address this limitation, we complement existing metrics with a novel evaluation method: the *Demographic Salience Score*, which quantifies how well age-related demographic entities are preserved in generated summaries. This allows us to assess summarisation fidelity through the lens of demographic representation, which is an essential but often overlooked dimension in biomedical NLP.

3 The DemogSummary Dataset

Existing systematic review datasets, such as Synergy (De Bruin et al., 2023), do not support stratification by population age group, which is essential for our analysis. We therefore construct a new dataset that enables explicit categorisation of primary studies and reviews into child, adult, and older adult populations.

3.1 Dataset Construction

Systematic reviews were selected based on the presence of demographic terms (e.g., 'Aged', 'Older Adult', 'Adult', 'Child') in titles or Medical Subject Headings (MeSH) annotations. Searches were conducted in three open-access biomedical databases: PubMed (National Center for Biotechnology Information (NCBI), 1988), Cochrane (Cochrane Database of Systematic Reviews, 1992), and Web of Science (Clarivate, 2025), using combinations of 'systematic review' and demographic keywords. This yielded reviews across medical domains, categorised into three population groups: children (aged <18 years), adults (aged 28-59 years), and older adults (60+ years). Inclusion required that reviews (i) focus on a single, well-defined demographic group and (ii) cite at least three primary studies with accessible abstracts. Any reviews that did not comply with (i) and (ii) were excluded. For each included review, cited primary studies were retrieved via PubMed and PubMed Central identifiers (PMIDs/PMCIDs); if unavailable, in-text hyperlinks were used. The final dataset qualifies under the UK Parliament (2014) exemption for non-commercial text and data analysis[1].

3.2 Demographic Annotation

We identified demographic information using a combination of rule-based methods and LLM-based named entity recognition; details are included in Appendix B. The resulting dataset in Table 1 includes 14 adult, 15 child, and 15 older adult systematic reviews, each comprising primary studies spanning multiple medical domains as described in Table 2. In total, these reviews encompass approximately 1,200 primary studies. Although the dataset consists of 44 systematic reviews, this scale is appropriate given the nature of systematic reviews, which are intended to synthesise large bodies of evidence and typically include many primary studies each.

Age Group	Reviews	Avg. Studies per Review
Adults	14	23
Children	15	32
Older Adults	15	25

Table 1: Overview of the DemogSummary Dataset.

Medical Domain	Number of Reviews
Public Health	14
Frailty	2
Mental Health	8
Nutrition and Digestive Health	9
Cognitive Health	1
Cardiovascular Health	5
Pain	2
Dental and Gut Health	2
Respiratory Health	1
Total	44

Table 2: Breakdown of the DEMOGSUMMARY Dataset by Medical Domain.

4 Experimental Design

4.1 Model Selection

We selected three large language models that reflect a range of architectures, context-handling capabilities, and access modalities:

QWEN (Qwen/Qwen2.5-14B-Instruct-1M) (Yang et al., 2025): a state-of-the-art open-source autoregressive transformer model with extended context length support. QWEN was selected due to its ability to process long sequences (up to 1M tokens), which is critical for working with multiple primary study abstracts without truncation.

GPT (OpenAI/gpt-4.1-nano) (OpenAI, 2023): a proprietary model accessed via the OpenAI API. We included this model as a strong commercial baseline, selected for its favourable trade-off between speed and performance in general-purpose language understanding tasks.

Longformer (allenai/led-large-16384-arxiv) (Beltagy et al., 2020): a transformer-based encoder-decoder model pretrained on scientific papers from ArXiv and designed specifically for scientific long-document processing, making it particularly suitable for tasks involving structured academic language and domain-specific content.

4.2 Task Setup

Each model is tasked with generating an abstractive narrative synthesis based on the full set of primary study abstracts from a systematic review. The target output is a concise summary approximating the published systematic review abstract. We evaluated two prompt conditions: a *regular prompt*, which did not refer to patient demographics, and an *age-aware prompt*, which explicitly informed the model of the population group involved. Both prompts are outlined in (Table 3). This setup enabled assessment of whether demographic cues influence summarisation behaviour.

4.3 Implementation Details

All experiments were conducted over approximately 90 GPU hours on a single NVIDIA L4. Inference with GPT-4.1-nano via the OpenAI API cost approximately $5 for all reviews, while the open-source models (QWEN and Longformer) were accessed via Hugging Face. More details, including runtime environment, inference cost, and hyperparameters, are provided in Appendix A.

[1] We release the PubMed IDs for systematic reviews in DEMOGSUMMARY alongside the code at: https://github.com/Favour-Yahdii/lllms_dont_see_age to support transparency, reproducibility, and future research.

> You are an experienced and objective biomedical systematic reviewer. Your task is to draft a concise, structured abstract that approximates a systematic review abstract, using the set of provided biomedical research abstracts.**The provided research abstracts involve studies conducted specifically in {POPULATION GROUP} populations. Keep this in mind as you complete the task.** Ensure to produce your summary abstract based only on the provided abstracts. Do not include any external information or personal opinions. Your summary should be a synthesis of the provided abstracts, not a critique or evaluation of them.

Table 3: Prompt used for summary generation. The text in **bold** was included only in the age-aware prompt.

5 Evaluation Metrics

We conducted a supervised evaluation using the published systematic review abstracts as gold standards. In addition to standard summarisation metrics, we introduced the *Demographic Salience Score* to assess retention of demographic content.

5.1 Standardised Metrics

We employed a suite of complementary evaluation metrics to capture different aspects of summarisation quality. BLEU (Papineni et al., 2002) measures surface-level n-gram overlap with the gold-standard abstract, while BERTScore (Zhang et al., 2019) evaluates semantic similarity using contextual embeddings from a pre-trained transformer. To assess fluency, informativeness, and alignment with expert-authored content, we used BARTScore (Yuan et al., 2021) in a supervised setup, where the model computes the likelihood of the generated summary given the gold standard. Lastly, FactCC (Kryscinski et al., 2020) was used to assess factual consistency. It classifies each sentence in the generated summary by whether it is supported by the gold summary or not.

5.2 Demographic Salience Score (DSS)

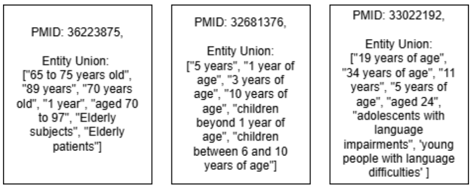
To evaluate how effectively demographic information is preserved and conveyed in multi-document summarisation, we introduce the Demographic Salience Score (DSS). While this study focuses on age-related demographic content, the metric is generalisable to other demographic dimensions (e.g., gender, ethnicity). DSS captures two key aspects of demographic fidelity: (1) inclusion of salient demographic entities, and (2) penalisation of unsupported (hallucinated) content.

Entity Extraction. We construct a gold-standard set of demographic entities by parsing each systematic review. Entities related to age are extracted using a combination of rule-based patterns and LLM-assisted methods[2]. These extracted entities form the reference set Ent_{gold}, which serve as the evaluation target.

To evaluate the reliability of the demographic annotation process, we manually reviewed a stratified random sample of 60 annotated abstracts, 20 from each age group (children, adults, older adults), representing approximately 5% of the total dataset.

In this subset, 59 out of 60 annotations (98%) were accurate. Based on this high accuracy rate and the diversity of the sample, we considered the annotation pipeline to be sufficiently reliable for downstream analysis.

Each primary study was subsequently categorised into one of the following age groups; **Child**: <18 years, **Adult**: 18–59 years, **Older Adult**: 60+ years.



PMID: 36223875,

Entity Union:
["65 to 75 years old", "89 years", "70 years old", "1 year", "aged 70 to 97", "Elderly subjects", "Elderly patients"]

PMID: 32681376,

Entity Union:
["5 years", "1 year of age", "3 years of age", "10 years of age", "children beyond 1 year of age", "children between 6 and 10 years of age"]

PMID: 33022192,

Entity Union:
["19 years of age", "34 years of age", "11 years", "5 years of age", "aged 24", "adolescents with language impairments", 'young people with language difficulties']

Figure 2: Subset of Gold standard Demographic Entities from Reviews

Scoring Components. The DSS has two main components, the Entity Retention Score (ERS) and the Hallucination Penalty (HP). Components such as the omission rate are derivatives of these main components. We establish $Ent_{summary}$ as the set of demographic entities in the generated summary.

Entity Retention Score (ERS): The proportion of gold entities preserved in the generated summary. This score reflects how comprehensively the summary captures the reference demographic content.

$$ERS = \frac{|Ent_{\text{summary}} \cap Ent_{\text{gold}}|}{|Ent_{\text{gold}}|} \tag{1}$$

Omission Rate. As the complement of ERS, we define omission as the proportion of gold entities missing from the summary:

$$Omission = 1 - ERS \tag{2}$$

Although not part of the DSS formulation, we report omission rates in our results to provide an additional perspective on demographic coverage.

[2] Details of these techniques are provided in Appendix B.

Hallucination Penalty (HP): The proportion of entities in $Ent_{summary}$ that do not match any entity in Ent_{gold}. A generated entity is considered a match if it either exactly matches a gold entity or exceeds a specified cosine similarity threshold[3].

$$HP = \frac{|Ent_{\text{summary}} \setminus \text{Match}(Ent_{\text{gold}})|}{|Ent_{\text{summary}}|} \tag{3}$$

Over-length Penalty (OP): In multi-document summarisation, excessively long outputs can artificially deflate the hallucination penalty by increasing the total number of extracted entities—thereby reducing the proportion of hallucinated content. To account for this, we introduce an over-length penalty (OP) that activates when the number of generated tokens exceeds a predefined threshold. This adjustment ensures that the hallucination penalty remains sensitive to unsupported content, regardless of summary length:

$$OP = \frac{\max(0, T_{\text{gen}} - T_{\text{max}})}{T_{\text{max}}} \tag{4}$$

where T_{gen} is the number of generated tokens, and T_{max} is a predefined token limit. The adjusted hallucination penalty is then defined as:

$$HP_{adj} = HP + OP \tag{5}$$

DSS: The Demographic Salience Score is computed as a weighted combination of retention and penalty terms:

$$\text{DSS} = \alpha \times \text{ERS} - \gamma \times \text{HP}_{\text{adj}} \tag{6}$$

Where α, and γ are non-negative weighting coefficients. To normalise the score to the $[0, 1]$ range, we divide by the maximum achievable score $\alpha \cdot N$ (i.e., a perfect score with full retention and no hallucination penalty), and clip negative values, where N is the number of systematic reviews:

$$\text{DSS}_{\text{normalised}} = \max\left(0, \frac{\alpha \times \text{ERS} - \gamma \times \text{HP}_{\text{adj}}}{\alpha \times N}\right) \tag{7}$$

This normalisation ensures interpretability while bounding the score, rewarding summaries that retain salient demographic content and penalising unsupported or excessive additions.

Implementation Details. We identified matched demographic entities using cosine-based semantic similarity, applying a threshold of 0.7 to determine matches[4]. The same similarity threshold was applied inversely to identify hallucinations,i.e., entities with similarity below the threshold were considered unsupported. Weighting parameters $\alpha = \gamma = 2$ were used in Equations 6 and 7, balancing the contribution of entity retention and hallucination penalties. These settings were chosen to prioritise demographic salience while discouraging unsupported content [5].

Metric Comparison and Interpretation. To contextualise the behaviour of DSS, we examined its relationship with the standard evaluation metrics introduced in Section 5.1. Pearson correlation analysis showed strong positive correlations between DSS and BLEU (0.970), BERTScore (0.923), and BARTScore (0.892), indicating that summaries with higher demographic fidelity often also exhibit strong lexical and semantic alignment with reference abstracts. However, despite this alignment, DSS remains conceptually distinct. Unlike standard metrics, which assess surface-level similarity, DSS explicitly measures the retention of demographic entities and penalises unsupported or omitted demographic information. In contrast, metrics such as BERTScore do not distinguish between which information is preserved or hallucinated, nor do they prioritise content relevance tied to specific population groups.

DSS also showed a negative correlation with FactCC (-0.99), a factual consistency metric, suggesting that DSS captures a complementary dimension of quality, namely, demographic salience and inclusion, that is not the primary focus of factuality-oriented metrics. Taken together, these comparisons indicate that DSS aligns with general measures of summary quality but contributes a focused, domain-relevant perspective that standard metrics do not directly capture.

5.3 Model-Level Performance Analysis

To assess differences in model performance across the three age groups, we used the Kruskal–Wallis test (Daniel, 2008), a non-parametric method appropriate for comparing three or more independent

[3] see Section 5.2 for details

[4] We also explored prompting an LLM for entity matching but found it inconsistent and less reproducible.

[5] We performed sensitivity analysis tests for these parameters; details can be found in Appendix C.

groups without assuming normality. When significant effects were found, Dunn's post-hoc tests (Sedgwick, 2012) with Bonferroni correction were applied to adjust for multiple comparisons. Effect sizes were calculated using epsilon-squared (ε^2).

6 Results and Discussions

Surface Metrics Are Demographically Insensitive; FactCC Shows Partial Sensitivity. Based on the results in Table 4, BLEU, BERTScore, and BARTScore exhibit minimal variation across models and age groups, indicating a lack of sensitivity to population-specific fidelity and limits their utility in demographic evaluation. While limited, FactCC reveals greater differentiation. GPT consistently scores higher, with age-aware prompts improving performance in the child group. However, scores remain low for older adults across QWEN and Longformer, despite similar gains, indicating ongoing challenges in summarising this group. The divergence between high FactCC scores and poor demographic fidelity underscores the need for metrics that capture representational accuracy.

DSS Highlights Model-Specific Tradeoffs. DSS, designed to assess demographic fidelity, reveals clearer distinctions between models. Under regular prompting, GPT achieves high DSS across all groups by balancing entity retention with low hallucination and length penalties. In contrast, QWEN, despite strong ERS, suffers from high hallucination and overly verbose summaries, resulting in sharply reduced DSS, especially in adults and children. Longformer exhibits fewer hallucinations overall but is notably uneven across age groups: it performs relatively well for older adults and children, where omission rates are lower, but struggles with the adult group, where retention is markedly weaker. Overall, in the older adult group, DSS is consistently higher.

Simple Fixes Are Non-Trivial: Age-Aware Prompting Yields Mixed Effects on Fidelity and Stability. The impact of age-aware prompting varies considerably across models and age groups. For GPT, it leads to modest gains in factual consistency (FactCC) but slightly reduces DSS due to lower entity coverage. QWEN shows high sensitivity to age aware prompting, achieving near-perfect ERS in some settings, but at the cost of increased hallucination and verbosity, leading to substantial drops in DSS. Longformer's behaviour is more consistent: It shows limited improvement overall but performs relatively well in the older adult group, where DSS increases without major penalty tradeoffs. These results suggest that while demographic specific age-aware prompting can shift model behaviour, it does not consistently enhance demographic fidelity and often introduces new sources of instability, thereby highlighting the non-triviality of improving model's retention of salient demographic entities.

6.1 Retention and Hallucination Patterns Across LLMs.

While Table 4 summarises model-level averages, it obscures instance-level variability that can affect the reliability of summarisation in practice. Figure 3 shows the distribution of Entity Retention Score (Figure 3a) and Hallucination Score (Figure 3b) for the adult group under the regular prompt. The ERS distribution highlights GPT's consistently strong demographic coverage, with minimal variation across reviews. In contrast, Longformer displays a wide spread and lower median, suggesting inconsistent retention of key population information. QWEN achieves comparable retention but at the cost of highly variable hallucination scores, including extreme outliers, indicating instability in content faithfulness. These patterns suggest that even when models appear comparable on average, their behaviour can diverge substantially across instances, a risk especially critical in biomedical settings. Additional distributions for other age groups and prompt conditions are provided in Appendix F, along with a token-level qualitative analysis in Section 6.3.

6.2 Statistical Analysis of Inter-Model Differences

We assessed whether model-level differences across age groups and evaluation metrics were statistically significant using the Kruskal–Wallis and Dunn's tests. Results below summarise key findings for each prompt condition.

Regular Prompt. Under the regular prompt, significant differences appeared mainly in the adult group. ERS and omission scores differed across models ($H = 10.30, p = 0.0058, \varepsilon^2 = 0.20$), with Longformer underperforming relative to GPT ($p = 0.013$) and QWEN ($p = 0.021$); differences between GPT and QWEN were not significant ($p = 1.00$). Hallucination scores also var-

Group	Model	BERT ↑	BART ↑	BLEU ↑	FactCC ↑	ERS ↑	HP ↓	Omission ↓	OP ↓	DSS ↑
Regular Prompt										
Child	GPT	**0.84**	**-2.26**	2.40	**0.76**	0.84	**0.12**	0.16	**0.00**	**0.72**
	QWEN	0.82	**-2.26**	1.68	0.52	**0.97**	0.58	**0.02**	0.95	0 (-0.55)
	Longformer	0.81	-2.30	**2.46**	0.44	0.91	0.33	0.09	**0**	0.63
Adult	GPT	**0.83**	-2.31	2.05	**0.82**	**0.81**	**0.12**	**0.19**	**0**	**0.69**
	QWEN	0.81	**-2.30**	1.76	0.51	0.78	0.74	0.22	1.02	0 (-0.98)
	Longformer	0.80	-2.37	**2.16**	0.60	0.45	0.18	0.50	**0.00**	0.27
Older Adult	GPT	**0.83**	-2.00	**3.75**	**0.71**	0.92	0.14	0.08	**0**	0.78
	QWEN	0.82	**-1.99**	2.25	0.43	**0.98**	0.11	**0.02**	**0**	**0.79**
	Longformer	0.81	-2.03	2.81	0.41	0.95	**0.07**	0.05	**0**	0.78
Age-Aware Prompt										
Child	GPT	**0.84**	**-2.25**	**3.04**	**0.83**	0.87	**0.17**	0.13	**0**	**0.69**
	QWEN	0.81	-2.31	0.72	0.63	0.89	2.06	0.11	1.32	0(-2.49)
	Longformer	0.82	-2.31	2.04	0.59	**0.91**	0.33	**0.09**	**0**	0.58
Adult	GPT	**0.83**	-2.31	**2.60**	**0.82**	0.62	**0.12**	0.38	**0**	**0.5**
	QWEN	**0.83**	**-2.27**	1.62	0.59	**0.93**	1.14	**0.07**	1.57	0(-1.78)
	Longformer	0.81	-2.35	2.18	0.63	0.45	0.18	0.55	**0**	0.27
Older Adult	GPT	**0.84**	-2.00	**3.70**	**0.67**	0.85	**0.04**	0.15	**0**	0.81
	QWEN	0.83	**-1.98**	2.27	0.42	**0.98**	0.11	**0.02**	1.33	0(-0.46)
	Longformer	0.81	-2.04	1.93	0.50	0.95	0.07	0.05	**0**	**0.88**

Table 4: Comprehensive automatic evaluation results across demographic groups and summarisation models. The top section reports scores under the regular prompt, and the bottom under the age-aware prompt. ERS, HP, Omission and OP are reported as averages across reviews. DSS is reported as the normalised value, with negative unnormalised scores in parentheses.

ied ($H = 7.23, p = 0.0269, \varepsilon^2 = 0.12$), with QWEN exceeding GPT ($p = 0.042$). In the child group, only hallucination differences were significant ($H = 6.14, p = 0.046, \varepsilon^2 = 0.10$), with QWEN hallucinating more than GPT ($p = 0.040$). No significant differences were found in the older adult group ($p > 0.90$).

Age-Aware Prompt. With the age-aware prompt, inter-model differences became more pronounced for adults. ERS and omission scores again differed significantly ($H = 11.40, p = 0.0034, \varepsilon^2 = 0.22$), with Longformer performing worse than QWEN ($p = 0.0027$); GPT and QWEN did not differ significantly ($p = 0.088$). Hallucination scores also varied ($H = 13.10, p = 0.0014, \varepsilon^2 = 0.26$), with QWEN exceeding both GPT ($p = 0.0023$) and Longformer ($p = 0.013$). Among children, hallucination scores showed strong differences ($H = 19.27, p = 6.54 \times 10^{-5}, \varepsilon^2 = 0.41$), with QWEN again exceeding GPT ($p = 0.00012$) and Longformer ($p = 0.0020$); ERS and omission were not significant. Older adults showed no significant differences on any metric ($p > 0.10$).

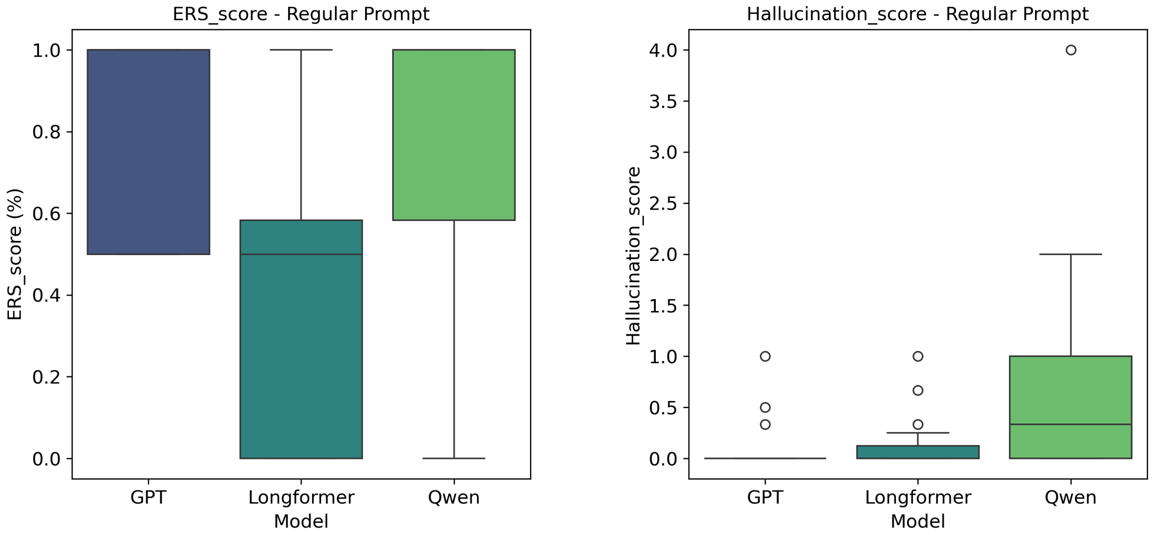
Overall, model variability was most pronounced in adults, particularly for ERS and hallucinations. QWEN consistently hallucinated more than GPT and Longformer, while Longformer's lower ERS and higher omission were concentrated in the adult group.

6.3 Qualitative Evaluation and Token Analysis

Entity-level behaviour was examined across three dimensions: entities retained, hallucinations, and omissions. These were evaluated across demographic groups and under both regular and age-aware prompts.

Entity Retention. Retained entities spanned a mix of canonical age-group labels (e.g., *adults*, *children*, *older adults*) and more granular descriptors such as *young adults*, *adults aged 18–35*, *early childhood*, and *community-dwelling older adults*. The age-aware prompt tended to elicit more specific and contextually rich terms, including detailed demographic subgroups (e.g., *healthy elderly men and women*, *Midwestern young adults*, *adult cardiac patients*). These findings suggest that models were generally sensitive to age-related cues and capable of surfacing relevant synonyms and descriptive variants, especially when the prompt was explicitly age-conditioned.

Hallucinations and Omissions. Hallucinated entities, while less frequent overall, often involved tangential or demographically unrelated popula-



(a) Entity Retention Across Models (Adult Group) - Regular Prompt

(b) Hallucinations Across Models (Adult Group) - Regular Prompt

Figure 3: Comparison of Entity Retention and hallucinations across Models for the Adult Group - Regular Prompt.

tions not mentioned in the gold document. Examples include the inclusion of *prisoners* in the population group of a review on smoking cessation interventions for young adults. Hallucination frequency was higher for QWEN and more prominent under the age-aware prompt, reflecting a tendency to overgenerate plausible but unsupported population descriptors. These patterns mirror the quantitative hallucination scores and indicate model-specific susceptibility to fabrication and generation of unsupported content, particularly when prompted to focus on demographic detail. Omissions included both general and highly specific age-related terms that were present in the gold documents but not captured in the generated outputs. These, similar to retained entities, ranged from descriptive phrases (e.g., *mean age 73.66 ± 14.67 years*, *residents in aged care settings*) to developmental or life stage terms such as *infancy*, *childhood*, and *young adulthood*. Across prompts, omissions were most pronounced for Longformer, consistent with its lower quantitative entity retention scores. The omission of precise subgroups may reflect limitations in entity grounding, particularly for complex or compound descriptors. There seemed to be no obvious pattern in the models decision on which entities to retain and which to exclude.

6.4 Effect of Prompt Design

The age-aware prompt systematically increased the specificity and range of demographic terms produced, particularly for adults and older populations. However, this increase in granularity was accompanied by greater lexical variability and, for QWEN in particular, more hallucinations and unsupported entities. GPT retained a better balance between fidelity and specificity, while Longformer exhibited lower entity coverage across conditions.

Overall, the age-aware prompt improved coverage of fine-grained descriptors but also amplified model-specific failure modes, such as hallucinations or omissions, highlighting the non-triviality of improving demographic entity retention in LLMs. These qualitative observations support the quantitative findings, particularly regarding QWEN's hallucination rate and Longformer's lower entity retention performance.

7 Conclusion

Our study shows that current LLMs exhibit systematic disparities in age-related information retention in biomedical summaries. Using **DemogSummary** and the **Demographic Salience Score (DSS)**, we quantify these biases, finding adult summaries particularly error-prone and underrepresented populations more likely to be hallucinated. These results highlight the need for demographic-aware evaluation and fair summarisation pipelines, paving the way for more equitable and transparent biomedical NLP systems.

Limitations

This study offers a focused evaluation of age-related fairness in LLM-based summarisation but has several limitations. First, the dataset covers a limited set of medical domains, which may constrain generalisability. Second, the causes of observed disparities remain under explored. Lastly, the study does not investigate bias mitigation methods, limiting its prescriptive value for fairness-oriented applications.

Acknowledgements

This work was supported by the University of Sheffield's Healthy Lifespan Institute Research Scholarship. We also acknowledge the University of Sheffield IT Services for providing access to High Performance Computing resources. We are grateful to the reviewers and area chairs for their thoughtful feedback and constructive discussions, and especially to Jasivan Sivakumar and Joe Stacey for their invaluable insights and guidance.

References

Divij Bajaj, Yuanyuan Lei, Jonathan Tong, and Ruihong Huang. 2024. Evaluating gender bias of LLMs in making morality judgements. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15804–15818, Miami, Florida, USA. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604. Place: Cambridge, MA Publisher: MIT Press.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Clarivate. 2025. Web of Science [Internet]. `https://www.webofscience.com`. Certain data included herein are derived from Clarivate™ (Web of Science™). © Clarivate 2025. All rights reserved. [cited 2025 May 19].

Cochrane Database of Systematic Reviews. 1992. Cochrane Database of Systematic Reviews [Internet]. `https://www.cochranelibrary.com/cdsr/about-cdsr`. [cited 2025 May 19].

Kate Crawford. 2017. The trouble with Bias.

Wayne W. Daniel. 2008. *Kruskal-Wallis Test*, pages 288–290. Springer New York, New York, NY.

Jonathan De Bruin, Yongchao Ma, Gerbrich Ferdinands, Jelle Teijema, and Rens Van de Schoot. 2023. SYNERGY - Open machine learning dataset on study selection in systematic reviews.

Fernando M. Delgado-Chaves, Matthew J. Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M. Mamdouh, Jan Baumbach, and Linda Baumbach. 2025. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences*, 122(2):e2411962122.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Jay DeYoung, Stephanie C. Martinez, Iain J. Marshall, and Byron C. Wallace. 2024. Do multi-document summarization models synthesize? *Transactions of the Association for Computational Linguistics*, 12:1043–1062.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Lixia Ge, Rupesh Agrawal, Maxwell Singer, Palvannan Kannapiran, Joseph Antonio De Castro Molina, Kiok Liang Teow, Chun Wei Yap, and John Arputhan Abisheganaden. 2024. Leveraging artificial intelligence to enhance systematic reviews in health research: advanced tools and challenges. *Systematic Reviews*, 13(1):269.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pre-training to adaptation: A case study in political bias. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1055, St. Julian's, Malta. Association for Computational Linguistics.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Mahammed Kamruzzaman, Md. Shovon, and Gene Kim. 2024. Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8940–8965, Bangkok, Thailand. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

National Center for Biotechnology Information (NCBI). 1988. National Center for Biotechnology Information (NCBI) [Internet]. `https://www.ncbi.nlm.nih.gov/`. [cited 2025 May 19].

Ani Nenkova and Kathleen McKeown. 2011. *Automatic Summarization*, volume 5. Now Publishers. Journal Abbreviation: Foundations and Trends in Information Retrieval Publication Title: Foundations and Trends in Information Retrieval.

Derek L. Nguyen, Yinhao Ren, Tyler M. Jones, Samantha M. Thomas, Joseph Y. Lo, and Lars J. Grimm. 2024. Patient characteristics impact performance of ai algorithm in interpreting negative screening digital breast tomosynthesis studies. *Radiology*, 311(2):e232286. PMID: 38771177.

OpenAI. 2023. Gpt-4 technical report. `https://arxiv.org/abs/2303.08774`. ArXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dipti Pawar, Shraddha Phansalkar, Abhishek Sharma, Gouri K. Sahu, Chun K. Ang, and Wei H. Lim. 2023. Survey on the Biomedical Text Summarization Techniques with an Emphasis on Databases, Techniques, Semantic Approaches, Classification Techniques, and Similarity Measures. *Sustainability*, 15(5).

Lena Schmidt, Ailbhe N. Finnerty Mutlu, Rebecca Elmore, Babatunde K. Olorisade, James Thomas, and Julian P. T. Higgins. 2023. Data extraction methods for systematic review (semi)automation: Update of a living systematic review.

Anna M Scott, Caitlin Forbes, Jasmine Clark, Matt Carter, Paul Glasziou, and Zachary Munn. 2021. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. *Journal of Clinical Epidemiology*, 138:80–94. Epub 2021 Jul 7.

Philip Sedgwick. 2012. Multiple significance tests: the bonferroni correction. *BMJ (online)*, 344:e509–e509.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182.

Julius Steen and Katja Markert. 2024a. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2024b. Bias in News Summarization: Measures, Pitfalls and Corpora. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, – NY USA. ACM.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and NengHai Yu. 2024. GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pages 1196–1210, New York, NY, USA. Association for Computing Machinery. Event-place: Salt Lake City, UT, USA.

UK Parliament. 2014. Copyright, Designs and Patents Act 1988, Section 29A.

https://www.legislation.gov.uk/ukpga/1988/48/section/29A? Section 29A inserted (1 June 2014) by The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 (S.I. 2014/1372), regs. 1, 3(2). [cited 2025 May 19].

Byron C. Wallace, Shibam Saha, Frank Soboczenski, and Ian J. Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Joint Summits on Translational Science Proceedings*, 2021:605–614.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025. Qwen2.5-1m technical report. *Preprint*, arXiv:2501.15383.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(1):176.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Hyperparameters

The hyperparameters in Table 5 were selected to encourage highly deterministic, concise outputs (temperature = 0) and to reduce redundancy. The maximum output token limit (750) was chosen both to provide sufficient token space for generating detailed summaries that approximate the length and content of gold-standard review abstracts, as well as a guide for penalising overly lengthy syntheses that provide weak and shallow aggregations of biomedical systematic review abstracts.

Model	Temperature	Max Tokens	FP
GPT-4.1 Mini	0	750	NA
GPT-4.1 Nano	0	750	NA
Qwen	0	750	1.1
Longformer	NA	750	1.1

Table 5: Hyperparameters used for different models in the experiment. FP = Frequency Penalty

B Demographic Annotation Procedure

B.1 Rule-Based Age Entity Extraction

Listing 1: Pseudocode for rule-based age extraction

```
def extract_demographics(text):
    demographics = {}

    # Define regular expression patterns
        to identify age-related
        expressions
    age_patterns = [
        # Matches "45 years old", "30-50
            years"

        r'(?<!\d)(\d{1,3})\s*(?:(?:-|to)
            \s*(\d{1,3}))?\s*(?:years?|
            yrs?)\s*(?:old|of age)?',

        # Matches "aged 40 to 65", "aged
            70"
        r'aged?\s*(\d{1,3})(?:\s*(?:to
            |-|-)s*(\d{1,3}))?',
        # Matches "6 month-old", "12 mo
            old"
        r'(\d{1,2})\s*(?:month[- ]old|mo
            old)',

        # Matches "age of 20 to 40", "
            aged 60-85"
        r'(?:age|aged)\s*(?:of\s*)?(\d
            {1,3})\s*(?:to|-|-)|s*(\d
            {1,3})'
    ]

    def find_matches(patterns):
        return [match.group().strip()
            for p in patterns for match
            in re.finditer(p, text, re.
            IGNORECASE)]

    demographics["age"] = find_matches(
        age_patterns)
    return demographics
```

B.2 LLM-Based Named Entity Recognition Prompt

Listing 2: LLM prompt for structured demographic extraction

```
You are an intelligent assistant tasked
    with extracting structured
    information from academic PDF
    documents.

Step 1: Extract the following fields:
- "title"
- "firstauthor"
- "year"
- "abstract" (until the "Keywords"
    section)
- "systematicreviewpmid"
- "stype"
- "populationdemographics"

Step 2: Return a JSON object in the
    following format:
```

```
{
  "title": "...",
  "firstauthor": "...",
  "year": "...",
  "abstract": "...",
  "systematicreviewpmid": "...",
  "stype": "...",
  "populationdemographics": [...]
}

Notes:
- Read the full abstract across chunks
    if needed.
- Do not stop early or include extra
    commentary.
```

C Sensitivity Analysis of DSS Parameters

To evaluate the robustness of the Demographic Salience Score (DSS), we conducted a sensitivity analysis on key hyperparameters: (i) the semantic similarity and hallucination thresholds, and (ii) the weighting parameters α (ERS weight) and γ (hallucination penalty).

Threshold analysis showed a consistent trend: DSS scores were highest at lower, more permissive thresholds and declined gradually as thresholds increased, reflecting stricter evaluation. We selected 0.7 for both thresholds, yielding a normalized DSS of 0.64—a balance between semantic precision and penalization of unsupported content. DSS varied minimally in the surrounding parameter space, indicating local stability.

For α and γ, we tested values between 1 and 2. We avoided assigning 0 to any of the parameters to avoid cancelling the contribution of the corresponding component, which is undesirable. In our tests, increasing α slightly improved DSS, whereas higher γ values reduced it due to stronger hallucination penalties. For experimentation, researchers may adjust these weights depending on which aspect they wish to emphasise, assigning higher weight to γ to prioritise hallucination control, or to α to strengthen DSS. The final configuration ($\alpha = 2.0$, $\gamma = 2.0$) achieved a DSS of 0.66, one of the highest observed, and exhibited robustness to parameter variation.

Figure 5 displays normalised DSS scores over a grid of semantic similarity and hallucination thresholds, while figure 5 shows DSS values as a function of α and γ, varied from 1.0 to 2.0.

D Source-Level Demographic Entity Extraction

Listing 3: Pseudocode for Demographic Extraction Function

```
Function extract_demographics(text):

    Initialise an empty dictionary:
        demographics

    Define regex patterns for:
        - Age (e.g., "45 years old", "
            aged 40 to 65", "6 month-old
            ")

    Define helper function find_matches(
        patterns):
        For each pattern in patterns:
            Use case-insensitive regex
                search on text
            Collect all matching
                substrings
        Return list of matches

    demographics["age"] = find_matches(
        age_patterns)

    Return demographics
```

Listing 4: Pseudocode for Age Entity Extraction

```
Function extract_entities(text):

    Define system prompt:
        "You are a helpful assistant.
            Given the abstract, extract
            all age related demographic
            entities.
        You should extract entities
            related to age.
        Your job is to extract these
            entities only, do not add to
             or subtract from the
            provided text."

    Send prompt and input text to
        language model (e.g., GPT):
        - Model: "gpt-4.1-nano"
        - Instructions: system prompt
        - Input: "Here is the abstract
            set: \n{text}"

    Receive response from model
    Return the output text as extracted
        entities
```

E Demographic Salience and Entity Retention

Listing 5: Pseudocode for Entity Retention Evaluation and DSS Computation

```
Function
    compute_semantic_similarity_and_ers(
    records, threshold=0.7,
        hallucination_threshold=0.7,
    alpha=2, gamma=2, overlength_penalty
        =None):

    Initialise:
        exact_matches := empty dict
```

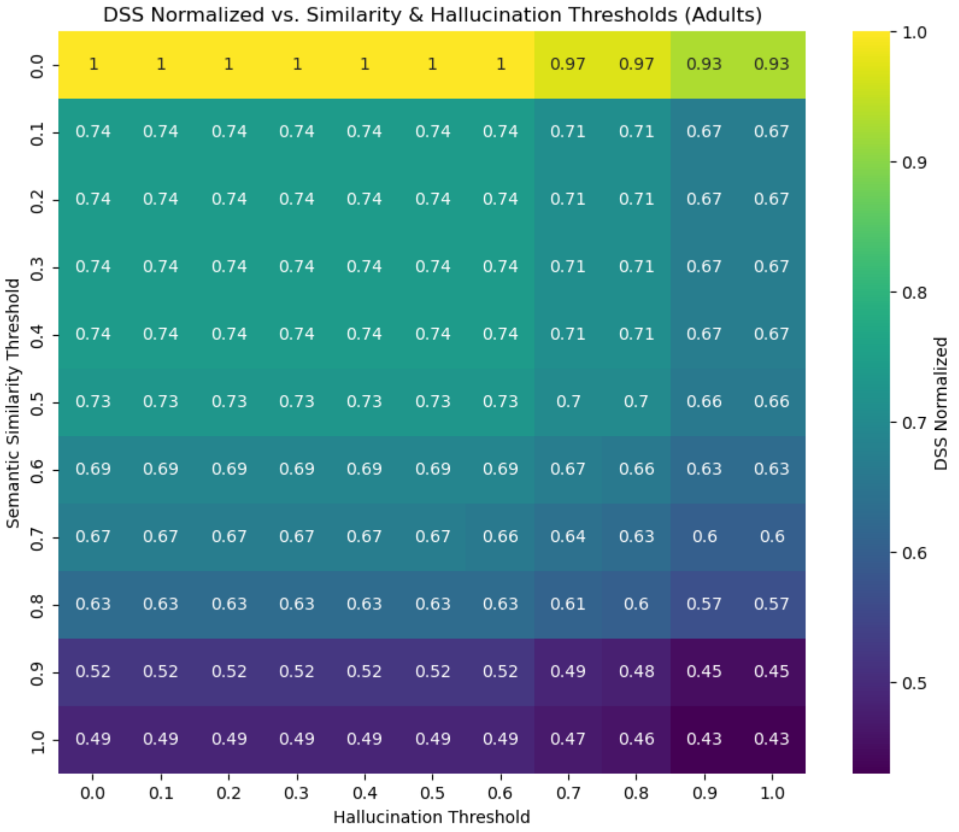


Figure 4: Semantic Similarity Threshold

```
    similar_matches := empty dict
    hallucinations := empty dict
    omissions := empty dict
    omission_scores := empty dict
    ers_scores := empty dict
    hallucination_scores := empty dict

For each record in records:
    reference_entities := list of gold entities
    generated_entities := list of predicted entities
    ID := record identifier

    Compute:
        exact := intersection of reference and generated entities
        embeddings_ref := embeddings for reference_entities
        embeddings_gen := embeddings for generated_entities

    For each reference entity not in exact:
        If cosine similarity with any generated entity >= threshold:
            Append to similar_matches
            Count as similar match

    For each generated entity not in reference_entities:
        If cosine similarity with all reference entities < hallucination_threshold:
            Add to hallucinations

    For each reference entity not matched:
        If cosine similarity with all generated entities < threshold:
            Add to omissions

    Compute:
        omission_score := omitted / reference_entities
        ers := 1 - omission_score
        hallucination_score := hallucinations / reference_entities

    Store scores for ID

Compute group-level metrics:
    ers_sum := sum of all ERS scores
    hall_sum := sum of all hallucination scores

    If overlength_penalty is provided:
        Add penalties to hall_sum

    DSS := alpha x ers_sum - gamma x
```

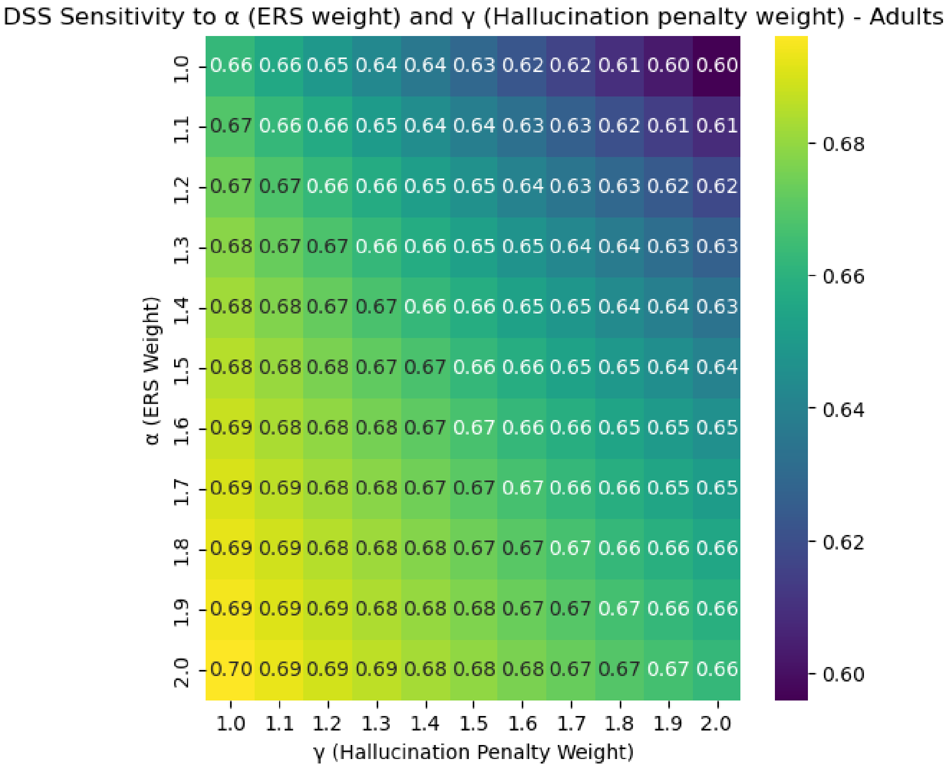


Figure 5: Alpha-Gamma Grid

```
         hall_sum
    DSS_normalised := max(0, DSS / (
        alpha x number of records))

Return:
    exact_matches, similar_matches,
        hallucinations,
        hallucination_scores,
    omissions, omission_scores,
        ers_scores, DSS,
        DSS_normalised,
    [overlength_penalty if provided]
```

F Entity Retention, Hallucinations and Omissions

Here we present the full set of results across both prompt regimes and demographic groups tested.

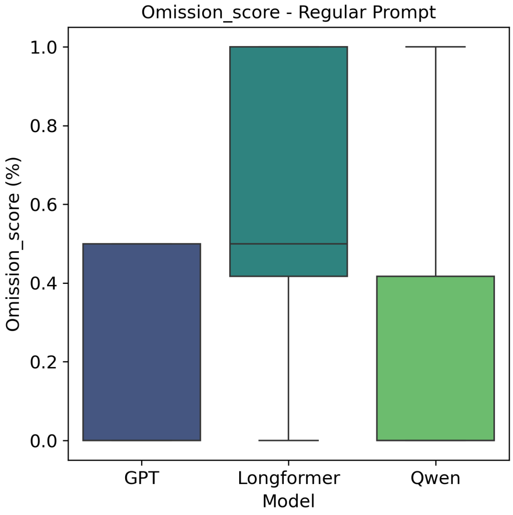


Figure 6: Entity Omissions Across Models (Adult Group) - Regular Prompt

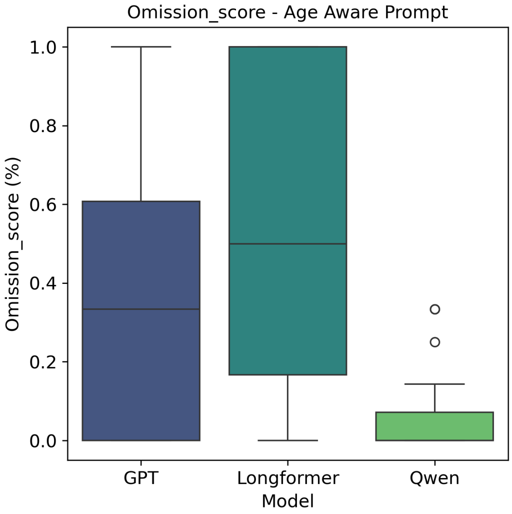


Figure 8: Entity Omissions Across Models (Adult Group) - Age Aware Prompt

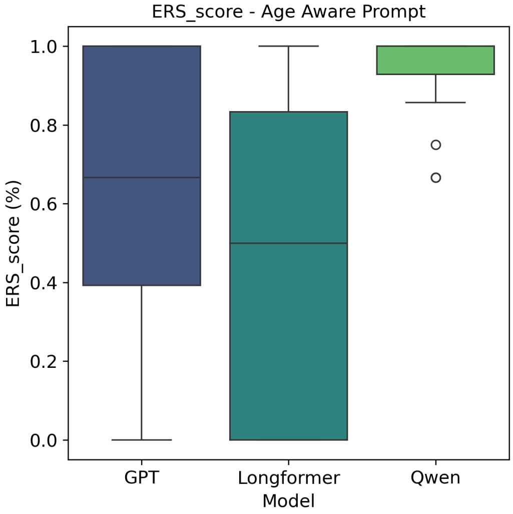


Figure 7: Entity Retention Across Models (Adult Group) - Age Aware Prompt

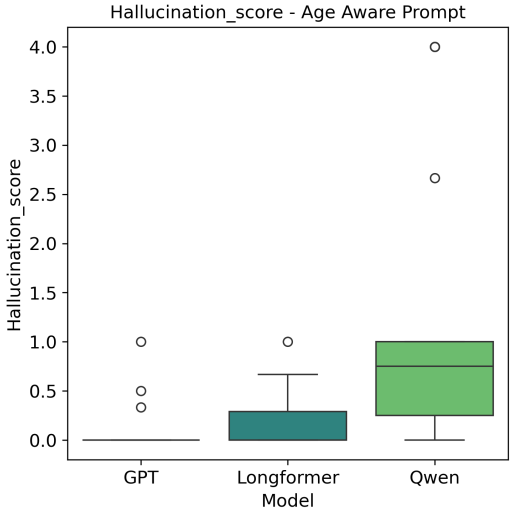


Figure 9: Hallucinations Across Models (Adult Group) - Age Aware Prompt

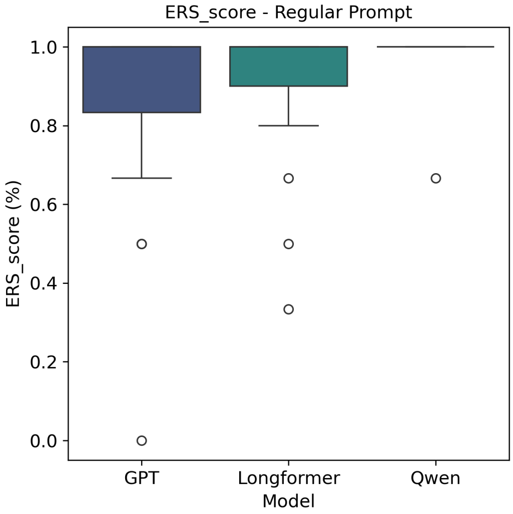


Figure 10: Entity Retention Across Models (Child Group) - Regular Prompt

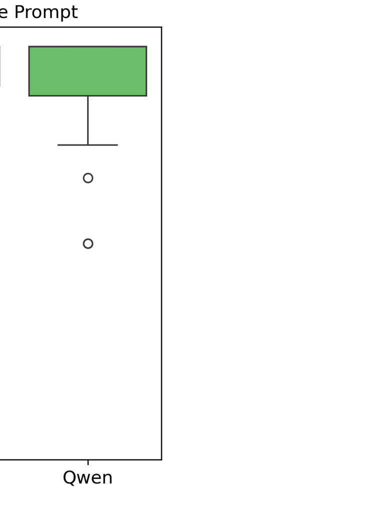


Figure 11: Entity Omissions Across Models (Child Group) - Regular Prompt

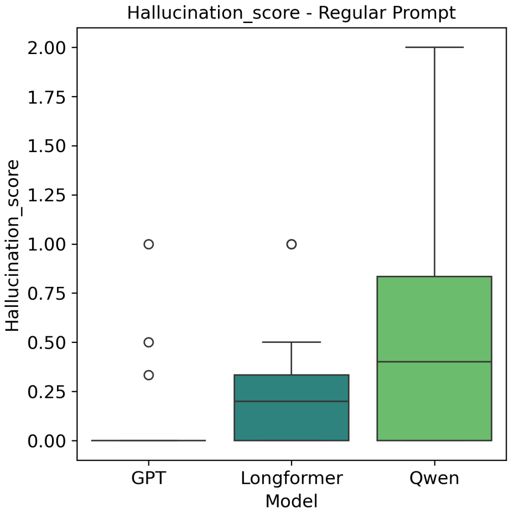


Figure 12: Hallucinations Across Models (Child Group) - Regular Prompt

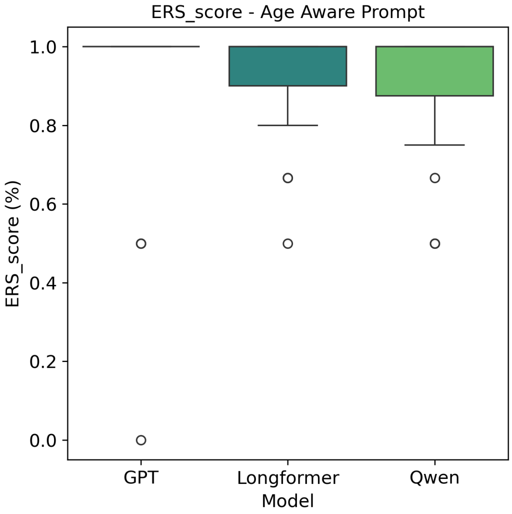


Figure 13: Entity Retention Across Models (Child Group) - Age Aware Prompt

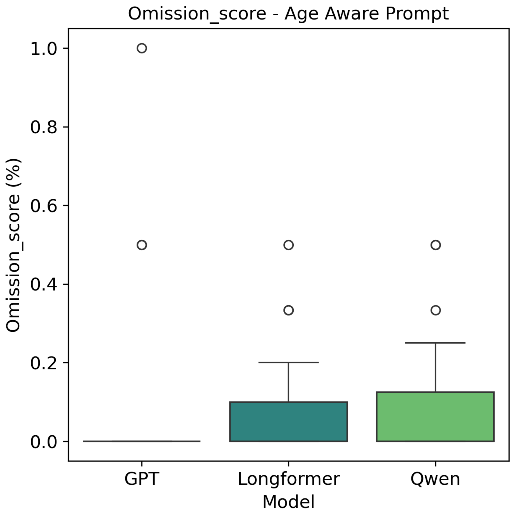


Figure 14: Entity Omissions Across Models (Child Group) - Age Awaare Prompt

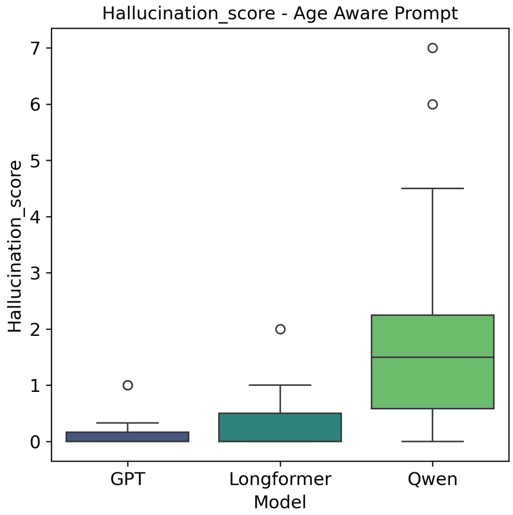


Figure 15: Hallucinations Across Models (Child Group) - Age Aware Prompt

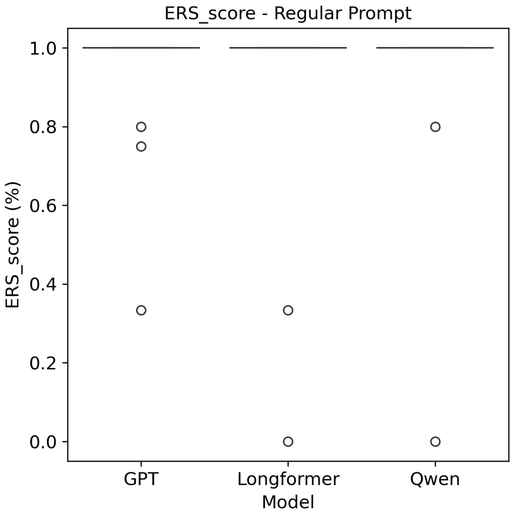


Figure 16: Entity Retention Across Models (Older Adult Group) - Regular Prompt

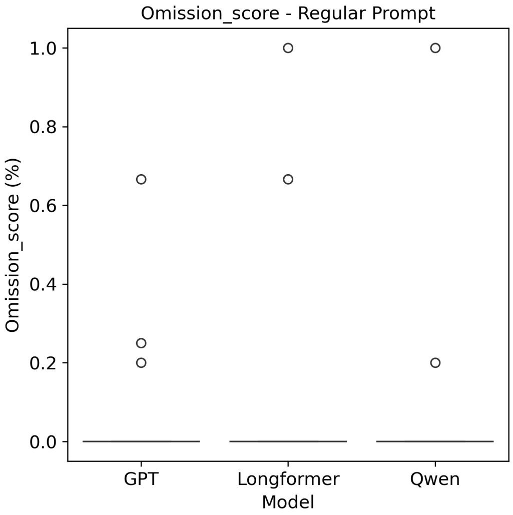


Figure 17: Entity Omissions Across Models (Older Adult Group) - Regular Prompt

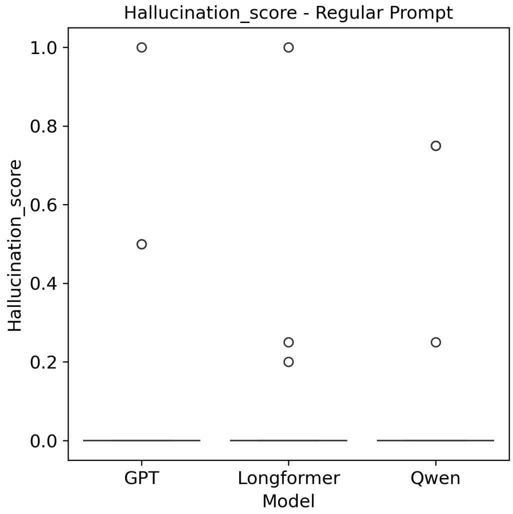


Figure 18: Hallucinations Across Models (Older Adult Group) - Regular Prompt

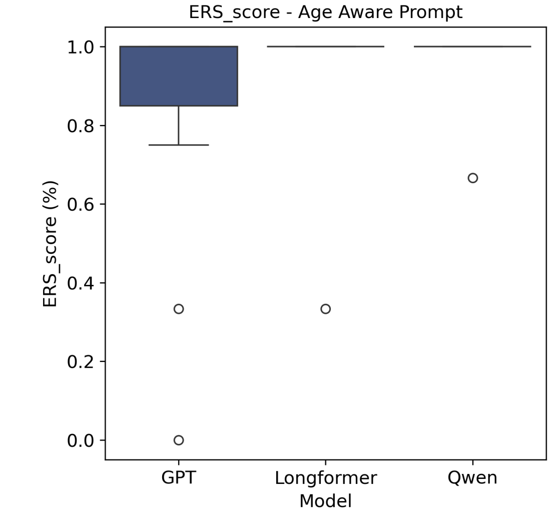


Figure 19: Entity Retention Across Models (Older Adult Group) - Age Aware Prompt

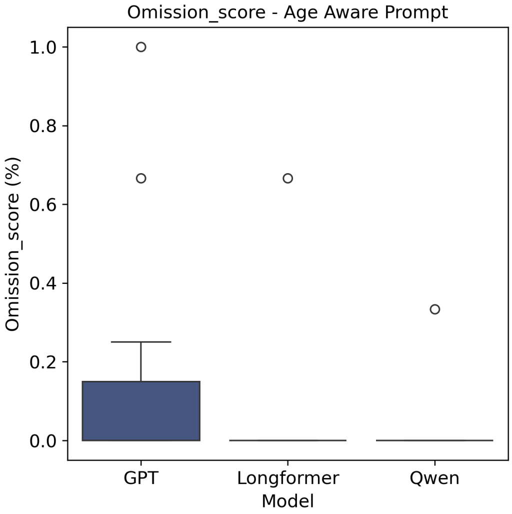


Figure 20: Entity Omissions Across Models (Older Adult Group) - Age Aware Prompt

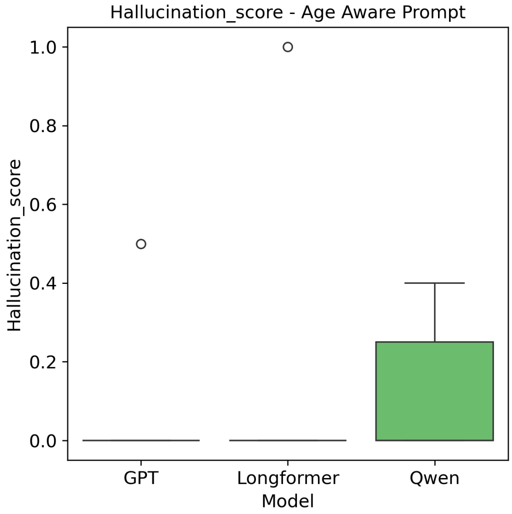


Figure 21: Hallucinations Across Models (Older Adult Group) - Age Aware Prompt