LLM Summaries Systematically Distort Topic Distributions

Shreeya Nallagonda[1], Melissa Torgbi[1], Harish Tayyar Madabushi[1], Steven Schockaert[2]

[1]University of Bath, [2]Cardiff University

Correspondence: htm43@bath.ac.uk

Abstract

LLM-generated summaries of customer feedback increasingly shape purchasing decisions and organisational strategy. However, we find that such summaries often do not stay faithful to the customer voice, potentially undermining the validity of any downstream decisions made from them. In particular, through a systematic evaluation of 50,000 synthetic customer reviews across 50 domains and 11 models, we demonstrate that the same topics are amplified across different models, regardless of their prominence in the actual reviews. More strikingly, models continue to amplify these same topics even when they are entirely absent from the input, and the topics that are amplified correlate with those the model independently rates as most important. Our findings were validated on real-world reviews from four domains. Furthermore, no intervention we tested, including scaling to frontier models, prompt engineering, and varying temperature, summary length, and input size, reduces the bias substantially.[1]

1 Introduction

Purchasing decisions are often influenced by the feedback that is reported in product reviews, while organisations crucially rely on such reviews to inform decisions on product development and resource allocation (Boytsov et al., 2025; Zhang et al., 2022). Due to the large volumes of customer feedback that are usually available, such decisions are increasingly based on LLM-generated summaries, rather than manual inspection of the reviews themselves (Malik and Bilal, 2024; Mohtadi et al., 2025). This practice crucially relies on the assumption that summaries faithfully represent the dominant themes from the given corpus. Recent work has started to challenge this assumption, showing that LLM-generated summaries imperfectly synthesise the overall sentiment of the reviews, being insufficiently sensitive to the ratio of positive to negative reviews (DeYoung et al., 2024) and over-representing majority opinions (Aghaebe et al., 2026), while also suffering from position bias (DeYoung et al., 2024; Sinha, 2025).

However, another fundamental question has received little attention: do LLM-generated summaries preserve the distribution of topics discussed in the source corpus? Consider a restaurant with several reviews that complain about high prices, but very few that discuss food quality. Then a summary which primarily focuses on food quality would clearly be misleading. In this paper, we study the quality of LLM-generated review summaries from this perspective. To this end, we construct a corpus of 50,000 synthetic reviews across 50 domains. For each domain, we consider 10 different topics, with each review covering exactly one topic. This setup allows us to precisely control the topic distribution in the source corpus and systematically analyse the impact of this distribution on the resulting LLM-generated summaries. The overall process is illustrated in Figure 1. Starting from a corpus in which the distribution of topics is uniform, we obtain an LLM-generated summary. We then use another LLM to predict the distribution of topics based only on the summary and analyse the divergence from the ground truth.

Our experiments reveal that LLM-generated review summaries significantly distort the topic distributions of the underlying corpus. While some information loss is inevitable when condensing many reviews (Peters and Chin-Yee, 2025), across 50 domains and 11 models, we find that *the same topics* are consistently amplified, and these are precisely the topics the model ranks as most important. In other words, the summaries reflect the models' priors about which topics are most important, rather

[1]The dataset and code to replicate our experiments are available at https://github.com/shreeya1510/llm-summarisation-bias.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 7110–7148
October 24-29, 2026 ©2026 Association for Computational Linguistics



Figure 1: An LLM summarises reviews covering ten topics in equal proportions and a judge estimates the topic distribution from the summary. The predicted proportions diverge substantially from the ground truth.

than what is actually covered by the reviews. We find that the extent to which a given topic is covered in the summary is almost completely independent of its prevalence in the original corpus. In fact, even when a topic is not mentioned at all, its prominence in the summary remains largely unchanged. While most of our analysis uses a synthetic corpus, we also validate our findings on real-world reviews from four domains.

2 Related Work

LLM-based summarisation of customer feedback has been extensively studied (Malik and Bilal, 2024; Mohtadi et al., 2025; Boytsov et al., 2025; Zhang et al., 2022; Belibasakis et al., 2025). However, to the best of our knowledge, the problem of topic distortion has not yet been studied in this context. Within a wider setting, Sim et al. (2022) found that abstractive multi-document summarisation systems distort topic proportions more than extractive approaches. Related work shows that models are undersensitive to changes in source composition when summarising movie reviews (DeYoung et al., 2024; Olabisi and Agrawal, 2024), and that varying summary length can increase topic diversity (Khandelwal, 2024). While these works have also reported some topic distortions, they have not studied whether the distortions are systematic and persistent across experimental conditions, which is what we focus on in this paper.

Other recent works have studied the limitations of LLMs for summarisation. For instance, LLMs have been found to generate factual inconsistencies and they have a tendency to overgeneralise beyond what the source documents support (Peters and Chin-Yee, 2025; Maynez et al., 2020; Guo et al., 2024; Huang et al., 2025). Xu and Durrett (2021) further show that abstractive models rely heavily on background language distributions absorbed during pre-training. For multi-document summarisation, even highly capable LLMs struggle with selective attention and information coverage (Ma et al., 2021; Fabbri et al., 2019; Ma, 2025; Aghaebe et al., 2026). Prior research suggests the bias we observe may partly stem from how LLMs process long inputs, where positional effects and instance count can amplify distortion (Liu et al., 2025; Cheng et al., 2023; Ma, 2025). LLMs tend to focus on information at the start and end of long contexts (Liu et al., 2023; Bertsch et al., 2025), and performance degrades as document count increases, even when individual document length is held constant (Chen et al., 2026; Levy et al., 2025). To rule these out as confounds, we shuffle reviews across runs, vary input size in our robustness checks and isolate the summarisation stage through ablation.

3 Methodology

In this section, we introduce the dataset we constructed and outline our core experimental pipeline.

Dataset Generation We constructed a number of synthetic datasets to ensure that we have precise control over ground truth topic distributions. We first selected 50 domains, such as *Restaurants & Cafés*, *Laptops* and *Airlines*. For each domain, we used Claude-Sonnet-4.5 to generate 10 topics of interest, where a topic in this context refers to a theme that might be addressed in a review. For instance, the *Restaurants & Cafés* domain involves topics such as *food quality*, *wait times* and *staff service*. A full list of the 50 domains and corresponding lists of topics can be found in Appendix B. For each domain and topic, we generated 100 reviews using GPT-5-mini, split evenly between positive and negative sentiment, giving a total set of 50,000 reviews. To avoid ambiguity in the analysis, every review focuses on a single topic. Furthermore, each review is short, consisting of one or two sentences. Finally, from the full set of reviews, we have sampled 10 smaller datasets for each domain, which we refer

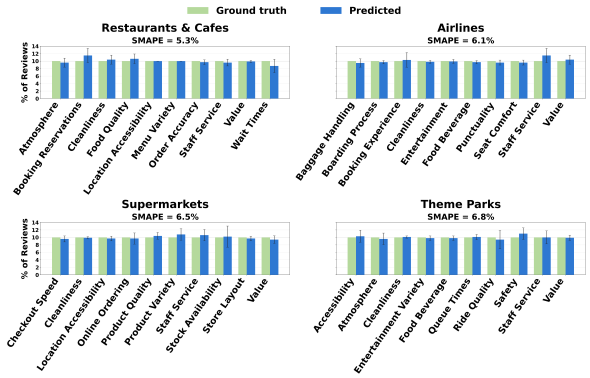


Figure 2: No-summarisation baseline results across four representative domains. The GPT-5-mini judge recovers the ground truth distribution closely from raw reviews, confirming that the LLM judge is not the source of the distortion.

Model	SMAPE	JSD
GPT-5-mini	40.8±4.7	0.030±0.007
GPT-5.4-mini	39.4±6.5	0.030±0.009
Llama-3-8B	59.2±10.0	0.064±0.023
Llama-3.3-70B	47.2±6.8	0.041±0.011
OLMo-3.1-32B	41.9±6.1	0.032±0.009
Qwen3-14B	34.8±5.2	0.022±0.006
Qwen3-80B	35.3±8.0	0.028±0.012
No-Sum	**16.3**±5.9	**0.009**±0.005

Table 1: Summary of results across all models using the same LLM for generating the summary and for estimating the topic distribution from this summary. The No-Sum baseline uses GPT-5-mini as the judge on raw reviews, without summarisation. SMAPE (%) and JSD are reported as mean ± std across 50 domains. Lower is better. Full per-domain results in Appendix E.

to as *surveys*. Each of these surveys consists of 40 reviews, namely 2 positive and 2 negative reviews for each of the 10 topics.

Pipeline Overview The aim of our experiments is to evaluate whether LLM-based multi-document summarisation faithfully captures the topic distribution of a collection of reviews. We achieve this through a two-stage approach (Figure 1). First, one of the generated synthetic surveys is passed to an LLM to produce a summary. Our prompt[2] for this step instructs the model that the summary should accurately reflect the distribution of topics that are discussed in the survey. The reviews from the given dataset are shuffled before being passed to the LLM to eliminate positional bias. In the second step, the summary is evaluated by an LLM judge, which is required to predict the importance of the 10 topics based on the summary. The 10 topics are listed as part of the prompt, along with a description of their meaning. We then assess the divergence between the ground truth (where the topic distribution is uniform) and the predicted percentages, which we primarily evaluate in terms of the Symmetric Mean Absolute Percentage Error (SMAPE).

4 Empirical Evaluation

In Table 1, we present the results of our main experiment, where we investigate whether LLM-generated summaries preserve the topic distribution of customer reviews. The table compares results for summaries generated by a number of different LLMs. We used the same LLM to act as summariser and as judge (which predicts the topic distribution from the summary).[3] To put the results in context, we also evaluate a no-summarisation baseline. For this variant, GPT-5-mini predicted the topic distributions from the original set of reviews, rather than from an LLM-generated summary. Examples of the resulting predictions are visualised in Figure 2. Our results show that summarisation dramatically increases SMAPE across all models and domains, ranging from 34.8% to 59.2%, compared to the no-summarisation baseline of 16.3%. We also report Jensen-Shannon Divergence, which measures how far the predicted distribution is from the ground truth as a whole rather than topic by topic. The two metrics agree closely. JSD ranges from 0.022 to 0.064 across models, compared to 0.009 for the no-summarisation baseline, confirming the trends seen in SMAPE scores. Figure 3 compares the predicted distribution with the ground truth for the Restaurants & Cafés domain across four representative models; a full comparison for all 50 domains can be found in Appendices E.1 and E.3. Despite a balanced ground truth of 10% per topic, *food quality* and *staff service* are consistently amplified while other topics are suppressed.

Consistency of Amplified Topics While some distortion is expected given the information loss inherent to summarisation, what is striking is the *consistency* with which topics get amplified. If distortion were a function of random information loss, we would expect different topics to surface as the

[2]Full prompts are provided in Appendix H.

[3]We also experimented with a variant in which GPT-5-mini is fixed as the judge, and only the summariser model is varied. Results for this variant were broadly similar and can be found in Appendix E.2.

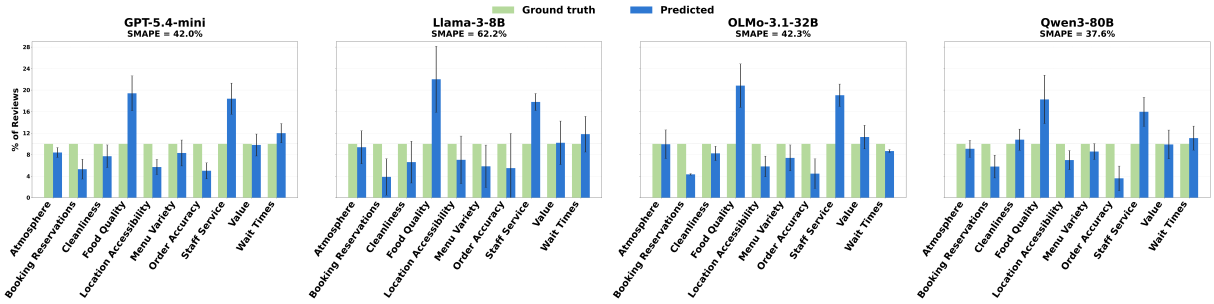


Figure 3: Predicted topic distributions for the Restaurants & Cafés domain across four representative models (same-model judge). Each bar represents the mean predicted proportion across 10 surveys.

Domain	Top amplified topic	Count / 70
Furniture Stores	Product Quality	62
Online Tutoring	Tutor Quality	57
Airlines	Staff Service	53
Insurance	Claims Process	50
Home Cleaning	Cleaning Quality	46
Laptops	Performance	46
Public Transport	Punctuality	46
Video Streaming	Content Library	46
Language Learning	Lesson Quality	45
Museums & Galleries	Exhibit Quality	45

Table 2: Domains where a single topic dominates as the most over-predicted, counted across 7 models and 10 surveys per domain (70 runs total). The chance-level expectation under uniform random over-prediction is 7.

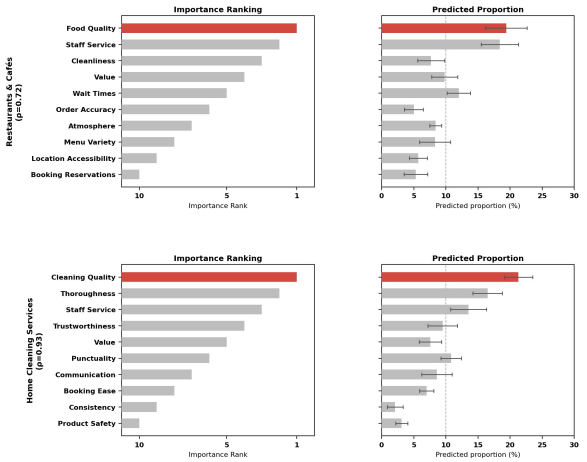


Figure 4: Importance ranking (left) vs predicted topic distribution (right) for two example domains. The topic the model independently rates as most important is also the topic it amplifies most heavily in its summaries (red), despite the ground truth being a balanced 10% per topic (dashed line).

most over-predicted across different models and surveys. Instead, the bias is clearly directional. Within a single model, the most over-predicted topic is identical across 65.0% of the 10 surveys per domain on average, far above the 10% expected by chance. Table 2 shows the strongest cases: in *Furniture Stores*, *product quality* is the top-amplified topic in 62 of the 70 runs (7 models $\times$ 10 surveys), against a chance-level expectation of 7, *tutor quality* dominates 57 of 70 *Online Tutoring* runs, *staff service* 53 of 70 *Airlines* runs and *claims process* 50 of 70 *Insurance* runs. Different models agree on which topic to amplify and the same model agrees with itself across runs.

Our hypothesis is that this bias towards particular topics reflects the model's pre-trained priors about which topics are most important in a given domain. Pre-training corpora contain large amounts of unstructured text describing every domain we test, and within each domain certain topics appear far more often than others (e.g. real-world restaurant reviews discuss food quality far more than booking systems). To test this hypothesis, we prompted GPT-5.4-mini to rank the 10 topics for each of the 50 domains by their perceived importance to customer satisfaction, *without showing any of the reviews* (Appendix F). We then rank each topic by its mean predicted proportion and compute the Spearman correlation between these two rankings across all 50 domains. We found an average Spearman $\rho = 0.68$, with 42 of the 50 domains showing $\rho > 0.5$. Figure 4 illustrates this alignment with model priors on two example domains. In *Home Cleaning Services* ($\rho = 0.93$), the model ranks *cleaning quality* as most important to customer satisfaction, and this is exactly the topic it amplifies most in summaries. The same pattern holds for *Restaurants & Cafés* ($\rho = 0.72$), where *food quality* dominates both the model's importance ranking and its predicted distribution. For a few domains, the correlation is weaker.

Domain	Summary excerpt
Restaurants & Cafés	*"Responses were mixed, with many comments focused on **food quality and menu** breadth..."*
Ride-Hailing	*"Responses were heavily focused on **pricing, fees, and transparency**..."*
Airlines	*"The responses are heavily centered on **service disruptions and operational reliability**..."*
Insurance	*"Across the responses, people most often commented on **claim handling**..."*
Online Tutoring	*"The strongest themes are about **tutor quality and learning outcomes**..."*
Furniture Stores	*"Responses focused heavily on **product quality and durability**..."*
Food Delivery	*"Responses were heavily centered on **delivery tracking and timing**..."*
Music Streaming	*"The responses are dominated by **playback and reliability experiences**..."*
Online Shopping	*"Responses were dominated by feedback about **marketplace reliability, especially shipping**..."*
Hair Salons	*"The responses are split between strong praise and serious complaints about **haircut quality and consistency**..."*
Home Cleaning	*"The responses are dominated by detailed feedback on **cleaning quality and consistency**..."*
Internet Providers	*"Responses were heavily centered on **service reliability and performance**..."*
GP Clinics	*"The responses were heavily focused on **access, timeliness, and coordination of care**..."*
Laptops	*"The responses are dominated by comments about **core laptop hardware quality and performance**..."*
Concert Venues	*"Responses were dominated by comments on **venue operations, accessibility, and crowd management**..."*
Washing Machines	*"The responses are dominated by feedback on **washer performance, reliability, and build quality**..."*
University Courses	*"Responses focused heavily on **teaching quality, assessment, and workload**..."*

Table 3: Examples of LLM-generated summaries that frame one theme as dominant across an underlying balanced set of reviews. Results for summaries generated with GPT-5.4-mini.

Qualitative Analysis To better understand the nature of the bias, we manually inspected a sample of LLM-generated summaries. For *Restaurants & Cafés*, we found that summaries explicitly emphasise food quality, in line with the findings from our main experiments. In fact, GPT-5.4-mini states in its summary *"comments focused on food quality and menu ..."*, directly asserting food as the dominant theme despite equal representation across all topics in the source dataset. Table 3 illustrates this effect across summaries that were generated by GPT-5.4-mini from balanced surveys across several domains. Llama-3.3-70B produced near-identical framing across four surveys, consistently describing *"a significant focus on the quality of food and beverages ..."*. This recurrent pattern in summary structure suggests boilerplate prior-driven phrasing rather than faithful reflection of the source data. Qwen3-80B similarly highlights food, describing *"...significant inconsistencies in food quality and preparation..."*. This confirms that the summaries themselves have misrepresented the source data and topic distortions do not (primarily) arise from mistakes by the LLM judge.

Figure 5 shows one such summary in full, illustrating how the framing persists across the entire generated summary.

Domain	GPT-5.4-mini	Llama-3.3-70B
Restaurants	50.3	76.3
Hotels	40.0	55.6
Laptops	27.6	62.0
Airlines	50.5	47.6
Mean ± Std	**42.1**±9.4	**60.4**±10.5

Table 4: Mean SMAPE (%) on real-world datasets (same-model judge).

5 Robustness Checks

If summarisation bias originates from pre-trained priors, it should persist regardless of surface-level conditions such as the choice of prompt, model scale or the number of reviews. We test this by conducting a series of targeted robustness checks, each designed to rule out a potential cause of bias. Unless stated otherwise, for this analysis we restrict the evaluation to 10 domains and we focus on GPT-5.4-mini and Llama-3.3-70B. We selected these 10 domains because they are relatively common and cover a broad range of products and services. Additionally, sentiment is balanced across all experiments, with the exception of the sentiment robustness check, where it is deliberately varied.

Generalisation to Real-world Data To validate our findings on real-world customer reviews, we applied our pipeline to two SemEval and one Airline Reviews benchmarks, spanning four domains, namely the SemEval-2014 Task 4 Restaurant and

Example LLM-Generated Summary - Restaurants & Cafés (GPT-5.4-mini)

Responses were mixed, with **many comments focused on food quality and menu breadth**. Positive remarks highlighted thoughtful accommodations for dietary needs such as low-FODMAP, gluten-free, vegan, and no-crouton requests, along with several well-executed dishes and desserts. Guests also praised good value deals, accurate spice customization, and generally strong items like avocado toast, pizza, panna cotta, and tiramisu. Negative food feedback centered on poor texture or flavor, undersized portions, and limited variety, especially for brunch and dairy-free/vegan options.

Service and ordering accuracy were another major theme. Several guests reported strong, efficient, and helpful service, including fast turnaround times, staff helping with lost items or complex bills, and reservations or app tools that worked well. However, there were also serious complaints about rude or dismissive staff, ignored manager concerns, and incorrect orders such as the wrong drink, wrong doneness, or missing requested modifications. Reservation handling was inconsistent, with reports of long waits, delayed seating, prepayment issues, and duplicate charges.

Comments also frequently addressed convenience and atmosphere. Positive notes mentioned easy access from transit, clear drop-off support, accessible parking assistance, quiet seating, and smooth booking features. On the negative side, guests cited unclear signage, restrictive parking clearance, greasy or unfinished-looking spaces, damp napkins, sticky high chairs, and harsh fluorescent lighting. Cleanliness was generally viewed positively in visible kitchen and prep areas, but there were some hygiene concerns in shared or family-use items.

Figure 5: Restaurants & Cafés summary (GPT-5.4-mini). Despite a balanced ground truth of 10% per topic, the summariser frames food quality and menu breadth as dominant from the opening line, leading the judge to over-predict food-related topics.

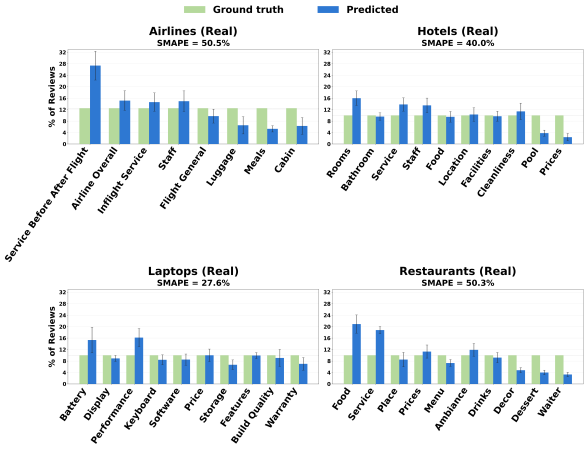


Figure 6: Ground truth vs predicted topic distributions on real-world datasets (GPT-5.4-mini, same-model judge). Despite balanced ground truth proportions, the same prior-driven distortion pattern observed in synthetic experiments persists across all four domains.

Domain	GPT-5.4	Gemini	DeepSeek	Claude
Restaurants & Cafés	32.0	22.5	30.9	30.5
Airlines	28.9	6.5	22.6	27.8
Supermarkets	23.6	15.0	36.0	28.1
Theme Parks	33.1	38.3	30.2	26.9
Online Shopping	34.5	26.6	32.9	33.7
Fitness Centres	30.6	33.5	32.3	35.3
GP Clinics	38.0	31.8	30.4	32.9
Cinemas	36.6	24.9	20.8	23.1
Laptops	34.6	31.3	35.7	39.2
University Courses	29.2	22.4	23.9	40.3
Mean ± Std	**32.1**±4.0	**25.3**±8.9	**29.6**±5.1	**31.8**±5.2

Table 5: Mean SMAPE (%) per domain for four frontier models: GPT-5.4, Gemini-3.1-Pro, DeepSeek-V3.2, Claude Sonnet 4.6.

Laptop datasets (Pontiki et al., 2014), the SemEval-2016 Task 5 Hotel dataset (Pontiki et al., 2016), and a publicly available Airline Reviews dataset (more details regarding datasets in Appendix G.1). The SemEval-2014 Task 4 dataset consists of over 3,000 English customer reviews for restaurants and laptops, annotated with aspect terms and coarse aspect categories such as food, service, and price, along with sentiment polarity (example review in Appendix G.1).

To enable an unambiguous analysis, we only include reviews which discuss a single aspect category, in line with how the synthetic dataset was constructed. Moreover, we retain only positive and negative sentiment, discarding neutral labels, allowing for a controlled like-for-like comparison with the synthetic datasets. In particular, we again create multiple datasets using the same balanced topics and sentiment distribution setup as before. Table 4 shows that both models exhibit high SMAPE scores across all four domains. Furthermore, the topics that are over-represented in the real data mirror those from our synthetic experiments, e.g. *food quality* and *service* in *Restaurants*, *battery life* and *performance* in *Laptops*, and *cabin service* in *Airlines* are consistently amplified. In the Airlines domain, *service before flight* is dramatically over-represented despite being one of eight equally distributed topics. Figure 6 illustrates this for all four domains.

Impact of Model Scale We examine whether frontier models are capable of providing more balanced summaries. For this analysis, we have used GPT-5.4, Gemini-3.1-Pro, DeepSeek-V3.2 and Claude Sonnet 4.6.

Table 5 shows all four models exhibit consider-

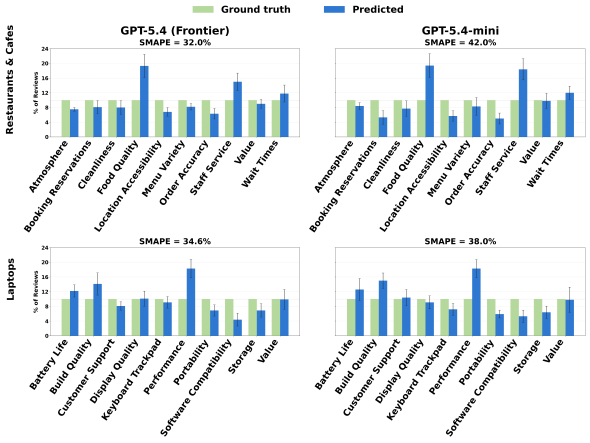


Figure 7: Predicted topic distributions for GPT-5.4 and GPT-5.4-mini across two example domains, Restaurants & Cafés and GP Clinics.

	GPT-5.4-mini		Llama-3.3-70B	
Domain	**Bal.**	**Skew.**	**Bal.**	**Skew.**
Restaurants & Cafés	42.0±8.5	56.6±7.5	48.0±12.6	62.6±8.5
Airlines	40.7±12.0	42.3±8.1	34.7±5.4	50.1±12.7
Supermarkets	30.2±7.8	44.7±7.8	40.6±8.3	45.0±17.3
Theme Parks	34.6±7.9	62.0±10.0	42.8±11.7	51.3±14.3
Online Shopping	34.4±6.9	51.5±6.5	50.9±11.8	57.8±6.9
Fitness Centres	34.7±7.3	53.3±9.8	44.6±10.4	67.7±10.1
GP Clinics	45.9±8.9	49.8±7.0	49.1±8.0	55.4±17.5
Cinemas	35.5±8.5	51.6±10.9	38.4±8.4	56.4±10.5
Laptops	38.0±9.0	54.5±10.1	64.6±21.8	77.0±16.3
University Courses	29.7±7.7	50.7±9.6	46.3±10.6	66.2±9.8
Mean	**36.7**	**51.7**	**46.0**	**59.0**

Table 6: SMAPE (%, mean ± std) for the 10-domain subset under balanced and skewed topic distributions. Bal. = balanced, Skew. = skewed.

able bias, with mean SMAPE ranging from 25.3% (Gemini-3.1-Pro) to 32.1% (GPT-5.4). While these values are slightly lower than what we obtained with smaller models in our main experiments, the summaries still exhibit clear topic distortion. GPT-5.4 (32.1%) shows only a slight improvement over GPT-5.4-mini (39.4%), suggesting model scale alone is not sufficient to eliminate summarisation bias. Figure 7 illustrates this directly: GPT-5.4 and GPT-5.4-mini produce near-identical distortion patterns across domains, with the same topics amplified in both models.

Impact of Topic Distribution In our experiments thus far, reviews were always distributed equally across all topics. We now analyse how starting from a skewed distribution affects the observed bias. We construct datasets using a fixed long-tail distribution of 25/20/15/10/10/7.5/5/2.5/2.5/2.5%, with the assignment of proportions to topics randomised. To accommodate this long-tail distribution, we increase the review count per dataset from 40 to 80. All other conditions remain unchanged. Table 6 shows both models perform worse under skewed conditions than balanced ones. GPT-5.4-mini rises from 36.7% to 51.7% and Llama-3.3-70B from 46.0% to 59.0%. Figure 8 reveals the underlying pattern: models essentially ignore the ground truth distribution. In Restaurants & Cafés, *value* has the highest ground truth proportion at 25% but is under-predicted, while *food quality*, consistently amplified in the balanced experiment, remains elevated despite holding only a 2.5% ground truth share. The *Laptops* domain shows the same pattern. *Performance* is over-predicted relative to its assigned ground truth share. The topics over-predicted in the balanced experiment remain amplified, even when assigned a low ground truth proportion, suggesting the model's prior beliefs persist regardless of the actual input distribution.

To probe this more carefully, we constructed surveys in which the ground truth proportions interpolate between the true, uniform distribution and the model's own biased distribution. At $\alpha = 0$, the ground truth is simply the honest uniform distribution (10% per topic). At $\alpha = 1$, the ground truth instead matches the distribution the model tends to predict when given a balanced input. Intermediate values of $\alpha \in \{0.25, 0.5, 0.75\}$ interpolate linearly between these two extremes. Figure 9 shows that predictions barely shift across these five conditions: SMAPE falls as α increases, but only because the ground truth is moving closer to what the model would have predicted anyway, not because the model is becoming more responsive to the actual input distribution.

Impact of Sentiment We test whether sentiment, rather than topic priors, drives the distortion by constructing surveys in which the first five topics receive only positive reviews and the last five only negative. The topic distribution remains balanced at 10% per topic, with all other conditions unchanged. The overall magnitude of the bias does not change, but a clear pattern emerges at the topic level: topics with only negative reviews are amplified by 1.4 percentage points on average, while topics with only positive reviews are suppressed by the same amount. The effect is strongest for *wait times* and *value*, which gain 2.8 and 2.6 percentage points respectively. Complaints therefore carry more weight than praise even when topic

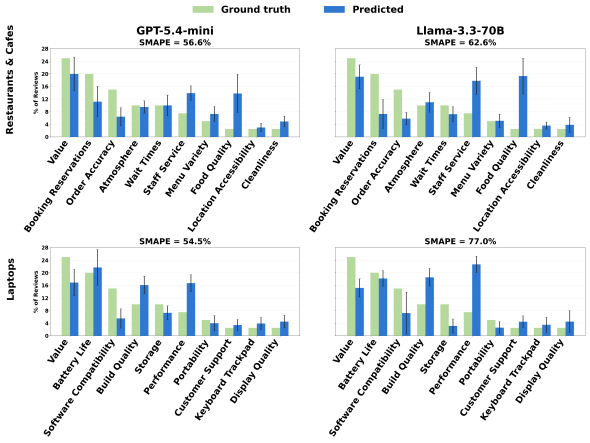


Figure 8: Ground truth versus predicted topic proportions under skewed input distribution for Restaurants & Cafés and Laptops (GPT-5.4-mini and Llama-3.3-70B). Topics ordered by ground truth frequency. Both models flatten the predicted distribution, with topics like food quality and performance remaining amplified despite low ground truth proportions.

proportions are held constant. Sentiment skew shifts which topics are emphasised without eliminating the underlying bias; full results are in Appendix G.5.

Impact of Topic Omission Given that the input distribution of topics was largely ignored by the models, we carried out an experiment in which the topic that was most over-predicted in the main experiment was omitted from the review set entirely. The surveys for this analysis are thus reduced from 40 to 36 reviews, and the ground truth is recalculated as a balanced distribution across the remaining 9 topics. Our results show that the omitted topics are still predicted to be prominent in the reviews in LLM summaries (Table 7). In domains such as *Restaurants & Cafés* (*food quality* - 14.2%/17.3%) and *GP Clinics* (*communication* - 14.1%/14.1%), both models allocate substantial proportion to the omitted topic. Figure 10 visualises this, with the omitted topic highlighted in red. The omitted topic consistently receives a non-zero predicted proportion.

Impact of Corpus Size Prior work has shown that LLM performance can degrade with longer inputs and higher instance counts (Liu et al., 2023; Levy et al., 2024; Chen et al., 2026). Motivated by these findings, we examine whether the number of reviews per survey affects the observed summarisation bias. We test configurations with 20, 40, 80 and 160 reviews per survey. Each domain has

Domain	Omitted Topic	GPT	Llama
Restaur. & Cafés	Food Quality	14.2	17.3
Airlines	Staff Service	10.4	8.9
Supermarkets	Stock Availability	6.7	6.4
Theme Parks	Accessibility	3.7	1.6
Online Shopping	Search Accuracy	2.5	6.6
Fitness Centres	Facilities	3.8	9.4
GP Clinics	Communication	14.1	14.1
Cinemas	Screen Quality	7.7	8.8
Laptops	Build Quality	11.5	10.1
University Courses	Admin Support	4.7	4.9
Mean ± **Std**		**7.9**±4.2	**8.8**±4.2

Table 7: Mean predicted proportion (%) assigned to the omitted topic, showing results for GPT-5.4-mini (*GPT*) and Llama-3.3-70B (*Llama*).

Review Count	GPT-5.4-mini	Llama-3.3-70B
r=20	36.7±5.1	46.4±7.9
r=40	37.2±5.3	47.2±9.1
r=80	36.0±6.5	45.8±7.0
r=160	37.3±6.6	46.8±10.2

Table 8: Mean SMAPE (%) across 10 domains at varying review counts. Bias remains stable across all review counts for both models.

10 topics, so a survey with 20 reviews contains 2 reviews per topic (1 positive and 1 negative), and similar for the other configurations.

Changing the number of reviews does not affect the predicted topic distribution within the summary. Table 8 shows that mean SMAPE remains stable across all conditions for both models. Visualising the results per domain (Figure 11) confirms the same topics are consistently amplified regardless of review count, showing consistent distortion across all four conditions.

Impact of Summarisation Prompt We test four variants of the *summarisation prompt*[4], drawn from Peters and Chin-Yee (2025). The first is our default proportional prompt, which explicitly instructs the model to reflect each topic in line with its original proportion. The second is a minimal prompt without any instructions to represent the topic distributions faithfully. The third explicitly asks the model not to introduce inaccuracies. The fourth is a Chain-of-Thought prompt asking the model to reason about topic frequencies before summarising. All other conditions remain unchanged across the 10-domain subset. The results in Table 9 show that the simple and accuracy prompts result in a slight reduction in SMAPE error. Surprisingly, however,

[4]Full prompts are provided in Appendix H.

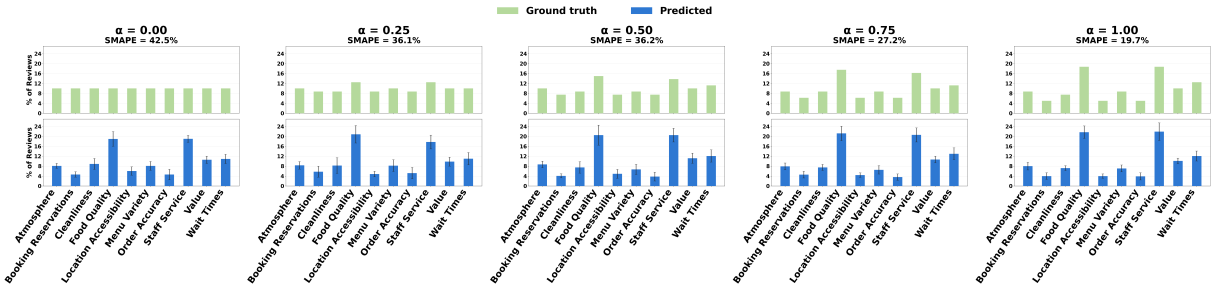


Figure 9: Predicted topic distributions for Restaurants & Cafés as the ground truth shifts from uniform ($\alpha = 0$) towards the model's (GPT-5.4-mini) predicted distribution from the uniform-input experiment ($\alpha = 1$).

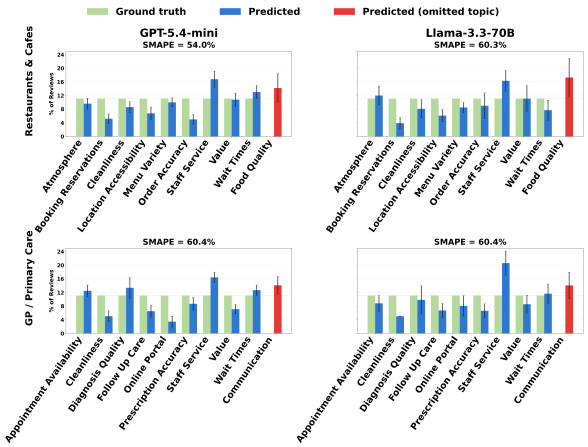


Figure 10: Predicted topic distributions when the most over-represented topic per domain (red) is omitted from source reviews.

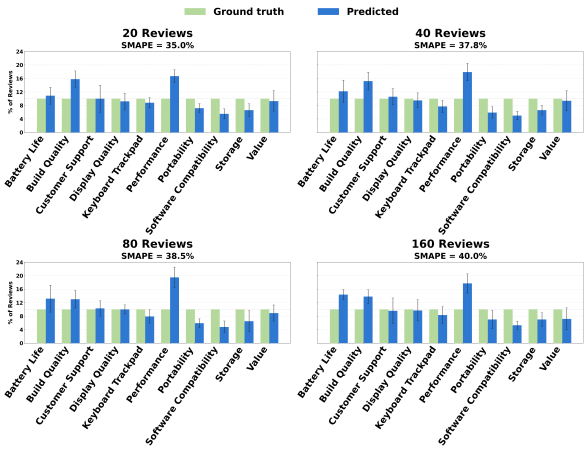


Figure 11: Predicted topic distributions for *Laptops* across varying review counts (GPT-5.4-mini). The same topics remain over- and under-represented regardless of how many reviews are passed to the summariser.

the Chain-of-Thought prompt has the opposite effect, resulting in more distorted topic distributions. This mirrors the ironic rebound effect reported by Peters and Chin-Yee (2025), where asking models to reason explicitly about accuracy surprisingly worsens performance. Despite these variations across prompt types, all four conditions remain substantially worse than the no-summarisation baseline of 16.3%. This confirms that prompt engineering can shift the magnitude of distortion but cannot eliminate it, consistent with bias originating from encoded priors rather than from how the model is instructed at inference time.

Summarisation Prompt	GPT-5.4-mini	Llama-3.3-70B
Default	38.0±10.1	48.4±16.4
Simple	34.7±9.9	41.5±15.4
Accuracy	33.2±8.6	37.6±12.2
Chain-of-Thought	45.7±24.6	58.0±31.8

Table 9: Mean SMAPE (%) across 10 domains for four summarisation prompt variants.

We also tested robustness to variation in summarisation model temperature (Appendix G.7), summary length (Appendix G.3) and judge prompt (Appendix G.8), finding the same persistent bias in each case.

6 Conclusion

We have investigated whether LLM-generated summaries preserve the topic distribution of customer feedback. Through systematic experimentation across 50 domains, 11 models and a wide range of varied conditions, we found that LLM-generated summaries heavily distort the topic distributions from the source corpus. Crucially, we found that this distortion largely reflects the model's prior beliefs about topic importance. These prior beliefs are furthermore remarkably persistent: LLM-generated summaries sometimes even emphasise topics which are absent from the source data entirely. Surface-level interventions like prompt engineering and temperature do not solve the issue.

Limitations

While this study provides rigorous evidence of systematic bias in LLM summarisation across a wide range of domains and models, some limitations must be acknowledged. Our synthetically generated reviews may not fully capture the real-world noise and linguistic complexity of real-world customer feedback. We partially addressed this by validating our findings on real customer reviews from the SemEval ABSA benchmarks, where consistent bias was still observed. All customer review data in this study are in the English language. It therefore remains unclear whether the observed biases generalise to other languages. Similarly, we focus on customer reviews, which means we cannot confirm whether similar distortion of topics occurs in other textual media such as news articles or academic papers. Additionally, while we have identified a strong correlation between LLM prior beliefs and over-prediction patterns, this does not demonstrate a causal relationship. Finally, as LLMs rapidly evolve and develop, the specific bias patterns we observe may shift in future models; our findings unveil biases in LLMs available at the time of this study.

Ethical Considerations

Our work does not imply that LLM summarisation is without value or that the practice should be abandoned. Summaries remain a useful interface for navigating large volumes of feedback, and the distortions we identify do not negate the broader linguistic competence of modern systems. They do, however, raise concerns about how summaries are interpreted in settings where prevalence matters, such as customer experience analytics, market research and policy consultation, where amplified themes can be mistaken for genuine consensus and quieter signals can disappear from view. The ease with which an LLM can frame one theme as dominant across a balanced set of inputs highlights the need for evaluation practices that go beyond surface coherence, along with caution from practitioners who rely on summaries as a proxy for the underlying distribution of opinion. At the same time, the fact that our judge recovers these distortions across multiple prompt variants and against a no-summarisation baseline does not remove the need for continued scrutiny of judge reliability, since shared biases between summariser and judge may understate the problem in ways that propagate downstream. Our findings offer a starting point for evaluating faithfulness from the perspective of topic preservation rather than through fluency alone, while recognising that the wider conversation about what LLM-generated text can and cannot be trusted to convey plays a central role in how these systems are adopted across research and applied settings.

Acknowledgments

This work was partly supported by the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath, by the UKRI Strength in Places Fund MyWorld Project (SIPF00006/1), and by EPSRC grants EP/W003309/1 and EP/Y028805/1.

References

Favour Yahdii Aghaebe, Tanefa Apekey, Elizabeth Williams, and Nafise Sadat Moosavi. 2026. Faithful summarisation under disagreement via belief-level aggregation. *Preprint*, arXiv:2601.04889.

Nikolaos Belibasakis, Anastasios Giannaros, Ioanna Giannoukou, and Spyros Sioutas. 2025. Enhancing hotel recommendations with ai: Llm-based review summarization and query-driven insights. *Preprint*, arXiv:2510.18277.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. 2025. Oolong: Evaluating long context reasoning and aggregation capabilities. *Preprint*, arXiv:2511.02817.

Ilya Boytsov, Vinny DeGenova, Mikhail Balyasin, Joseph Walt, Caitlin Eusden, Marie-Claire Rochat, and Margaret Pierson. 2025. End-to-end aspect-guided review summarization at scale. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 463–471, Suzhou (China). Association for Computational Linguistics.

Jingxuan Chen, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2026. Understanding llm performance degradation in multi-instance processing: The roles of instance count and context length. *Preprint*, arXiv:2603.22608.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model APIs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 792–810, Singapore. Association for Computational Linguistics.

Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. 2021. The coefficient of determination r-squared is more informative than smape, mae,

mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *Preprint*, arXiv:2411.10915.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593, Mexico City, Mexico. Association for Computational Linguistics.

Sicong Huang, Qianqi Yan, Shengze Wang, and Ian Lane. 2025. Enhancing faithfulness in abstractive summarization via span-level fine-tuning. *Preprint*, arXiv:2510.09915.

Trishia Khandelwal. 2024. Investigating the impact of text summarization on topic modeling. *Preprint*, arXiv:2410.09063.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. 2025. More documents, same length: Isolating the challenge of multiple documents in rag. *Preprint*, arXiv:2503.04388.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, and 18 others. 2025. A comprehensive survey on long context language modeling. *Preprint*, arXiv:2503.17407.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2021. Multi-document summarization via deep learning techniques: A survey. *Preprint*, arXiv:2011.04843.

Jing Ma. 2025. Input order shapes llm semantic alignment in multi-document summarization. *Preprint*, arXiv:2512.02665.

Nadia Malik and Muhammad Bilal. 2024. Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges. *PeerJ Computer Science*, 10.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Samaneh Mohtadi, Kevin Roitero, Stefano Mizzaro, and Gianluca Demartini. 2025. The effect of document summarization on llm-based relevance judgments. *Preprint*, arXiv:2512.05334.

Olubusayo Olabisi and Ameeta Agrawal. 2024. Understanding position bias effects on fairness in social multi-document summarization. *Preprint*, arXiv:2405.01790.

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Preprint*, arXiv:2504.00025.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking prompt sensitivity in large language models. *Preprint*, arXiv:2502.06065.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 7346–7356. Association for Computational Linguistics.

Mong Yuan Sim, Wei Emma Zhang, and Congbo Ma. 2022. An empirical study on topic preservation in multi-document summarization. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 61–67, Online. Association for Computational Linguistics.

Neelabh Sinha. 2025. QA-prompting: Improving summarization with large language models using question-answering. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 199–212, Hybrid. Association for Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6925–6940, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *Preprint*, arXiv:2203.01054.

A Models

Table 10 lists all models evaluated in this study, including parameter counts, access method and inference configuration.

B Dataset Generation

Domains and topics were generated using Claude Sonnet 4.5, seeded with four initial domains and prompted to suggest 46 more across diverse consumer-facing industries. The full list appears in Table 12. Topic descriptions and an example of a positive and a negative review are shown in Table 11 and Figure 12. Figure 29 shows the prompt given to GPT-5-mini for generating reviews. We manually validated a random sample of 50 reviews, confirming that each review covered a single topic and that its stated sentiment matched its content.

Example synthetic reviews: `food_quality` (Restaurants & Cafés)

Positive: "The truffle mushroom risotto was insanely creamy, arborio rice cooked perfectly al dente with a generous shave of black truffle and just the right amount of parmesan."
Negative: "My burger bun was soggy and the patty was dry and chalky, with no seasoning at all."

Figure 12: Example synthetic reviews generated for the `food_quality` topic in the Restaurants & Cafés domain.

While no deduplication was applied during generation, we checked for duplicates afterwards. Across all 50,000 reviews there are zero exact duplicates. A near-duplicate scan over all 1,225,000 within-topic pairs (token Jaccard > 0.8) finds only a single pair. To validate sentiment at scale, we additionally used two independent classifiers, Llama-3.3-70B and Gemini-3.1-Pro, neither of which was involved in generating the data, to classify the sentiment of a 500-review random sample. Both classifiers agreed with the original sentiment label in 100% of cases, confirming that our generated reviews reliably carry their intended sentiment.

C Topic Taxonomy

Table 12 lists all 50 domains and their associated topics used in the synthetic dataset.

D Implementation Details

All models except Qwen3-14b were accessed via the OpenRouter API (in total approximately £80 was spent on API credits) using the OpenAI Python SDK. Qwen3-14b was run locally on Google Colab on an NVIDIA A100 GPU (40GB VRAM) using the Hugging Face Transformers library with 4-bit NF4 quantisation via BitsAndBytes (Dettmers et al., 2023). Temperature was set to 0.0 throughout for deterministic outputs (Renze, 2024), with remaining parameters set to provider defaults, unless stated otherwise. Datasets within each domain were processed in parallel using Python's `ThreadPoolExecutor`, with all outputs written to individual JSON files to allow experiments to be interrupted and resumed. API failures were handled with a retry loop capped at three attempts, using delays of 5, 15 and 30 seconds between retries.

Figure 13 shows an example judge prediction. These predictions were normalised and then used

```
{
  "staff_service": 10,
  "seat_comfort": 18,
  "punctuality": 20,
  "value": 8,
  "food_beverage": 8,
  "cleanliness": 12,
  "booking_experience": 4,
  "baggage_handling": 8,
  "entertainment": 6,
  "boarding_process": 6
}
```


Figure 13: Example judge output for the Airlines domain.

to calculate the mean SMAPE across 10 runs per domain. SMAPE is scale-invariant, symmetric and interpretable, accounting for relative magnitude in a way that MAE does not (Chicco et al., 2021). MAE penalises a 5 percentage point error on a 10% ground truth topic identically to the same error on a 50% topic, despite the former representing a larger relative deviation. Given the actual values $a_1, ..., a_n$ of some quantity and the corresponding predicted values $p_1, ..., p_n$, SMAPE is evaluated as:

$$\frac{2}{n}\sum_{i=1}^{n}\frac{|a_i - p_i|}{|a_i| + |p_i|}$$

In our setting, the values $a_1, ..., a_n$ correspond to the ground-truth proportions while $p_1, ..., p_n$ are the topic proportions that are predicted by the LLM judge. We report SMAPE as a percentage, multiplying the value above by 100.

E Main Experiment

This section provides full per-domain results for the main experiment across both judge configurations.

E.1 Same-Model Judge

Table 13 reports the full per-domain SMAPE results when each summariser model also acts as its own judge.

E.2 Fixed Judge

Table 14 reports the full per-domain SMAPE results for all summariser models when paired with a fixed GPT-5-mini judge.

E.3 Per-Domain Predicted Distributions

Figures 14-18 show the predicted topic distributions across all 50 domains under the same-model judge configuration, with each batch covering 10 domains.

E.4 No-Summarisation Baseline

Figure 19 shows the predicted topic distributions across all 50 domains when reviews are passed directly to the GPT-5-mini judge without summarisation. Figure 30 shows the prompt used for this experiment.

F Prior Beliefs Analysis

Table 16 reports the per-domain Spearman correlation between GPT-5.4-mini topic importance rankings and predicted topic emphasis for the 10-domain subset used in our robustness checks. Figure 31 shows the prompt used to produce the importance rankings.

G Robustness Check Full Results

This section provides the full per-domain results for each robustness check experiment described in Section 5.

G.1 Real-World Data

Our external validation uses four real-world review datasets accessed via Hugging Face (details can be found in Table 17). The Restaurants and Laptops datasets are from SemEval-2014 Task 4 (Pontiki et al., 2014) and are accessed through the mirror `Akash092003/ABSA-alpaca-SemEval2014Task4` under the MIT licence. The underlying SemEval-2014 release is itself distributed under a Creative Commons Attribution 4.0 International licence by the SemEval task organisers. The Hotels dataset is from SemEval-2016 Task 5 (Pontiki et al., 2016) and is accessed through srinivasbilla/semeval-2016-absa-reviews-english-translated-resampled under the MIT licence. The Airlines dataset is the publicly available `Alisairlines/Airline_Reviews` collection on Hugging Face. The dataset card and repository do not declare an explicit licence. We use all four datasets only for non-commercial research purposes consistent with their stated terms or, in the case of the Airlines dataset, with standard academic fair use.

The SemEval datasets consist of English customer reviews annotated with aspect categories such as food, service or price, along with sentiment polarity. The Airline Reviews dataset contains 508 English review sentences with aspect-level 1-5 ratings, skewing heavily negative. To match our synthetic setup, we selected the most frequently occurring aspect categories per dataset, retained only single-aspect reviews with positive

Model	Provider	Parameters	Temp.	Access
`gpt-5-mini`	OpenAI	Undisclosed	0.0	OpenRouter
`gpt-5.4-mini`	OpenAI	Undisclosed	0.0	OpenRouter
`gpt-5.4`	OpenAI	Undisclosed	0.0	OpenRouter
`llama-3-8b-instruct`	Meta	8B	0.0	OpenRouter
`llama-3.3-70b-instruct`	Meta	70B	0.0	OpenRouter
`qwen3-14b`	Alibaba	14B	0.0	Local (HF)
`qwen3-80b`	Alibaba	80B	0.0	OpenRouter
`olmo-3.1-32b-instruct`	AllenAI	32B	0.0	OpenRouter
`gemini-3.1-pro-preview`	Google	Undisclosed	0.0	OpenRouter
`deepseek-v3.2`	DeepSeek	671B (MoE, 37B active)	0.0	OpenRouter
`claude-sonnet-4-6`	Anthropic	Undisclosed	0.0	OpenRouter

Table 10: Models evaluated in this study, with parameter counts, temperature and access method. Qwen3-14B was run locally with 4-bit NF4 quantisation via BitsAndBytes.

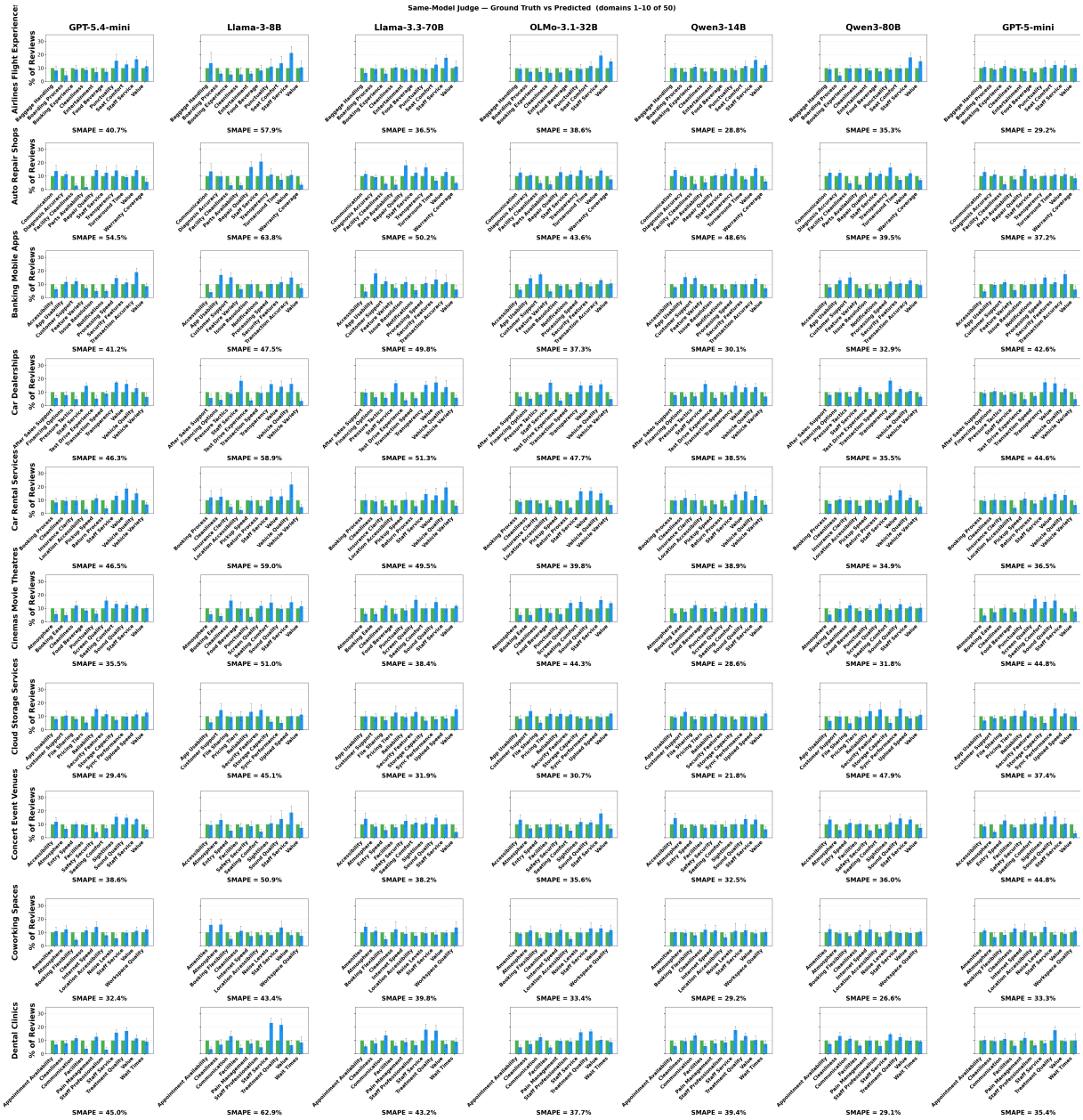


Figure 14: Predicted topic distributions under same-model judge (domains 1-10 of 50).

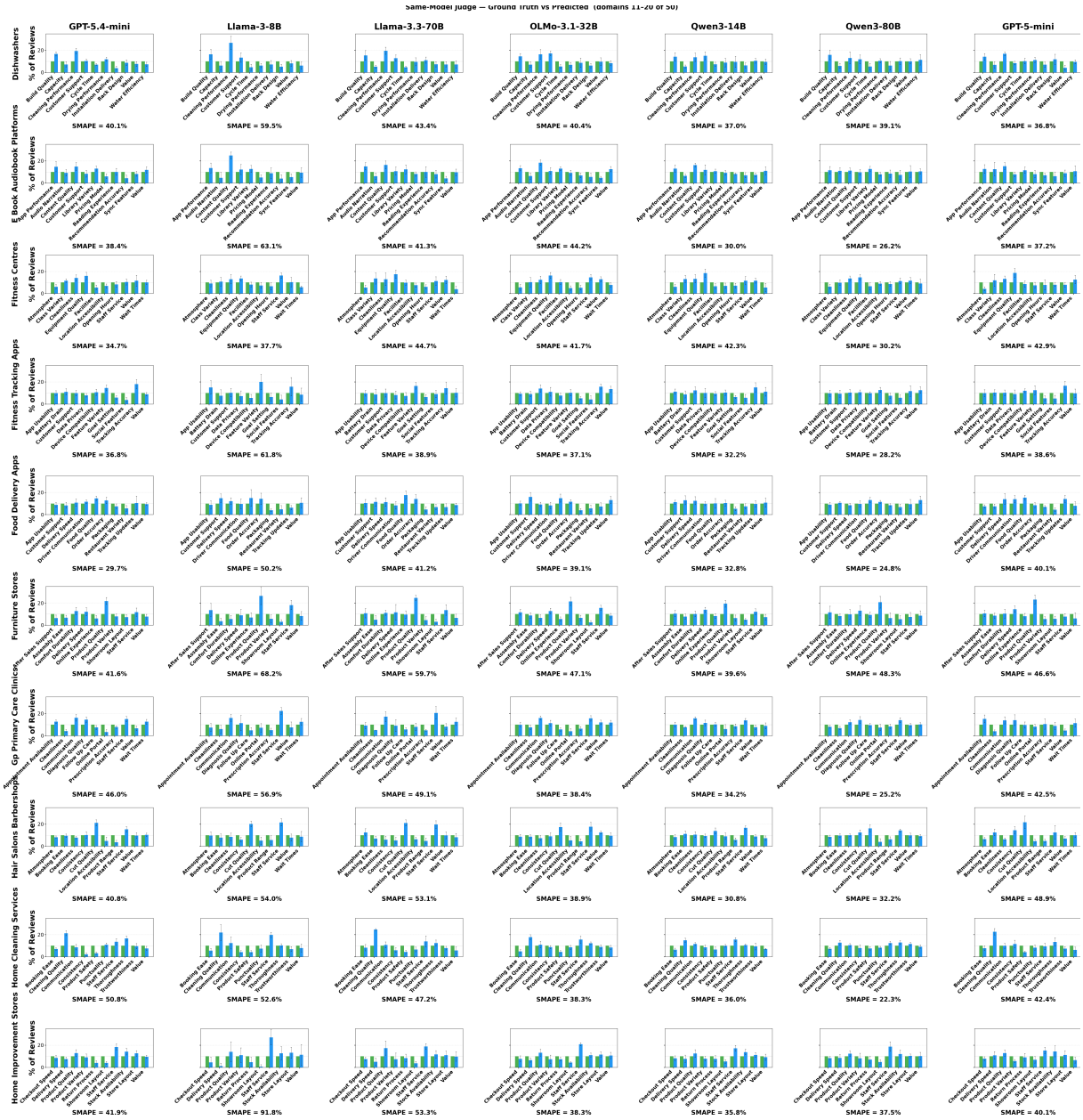


Figure 15: Predicted topic distributions under same-model judge (domains 11-20 of 50).

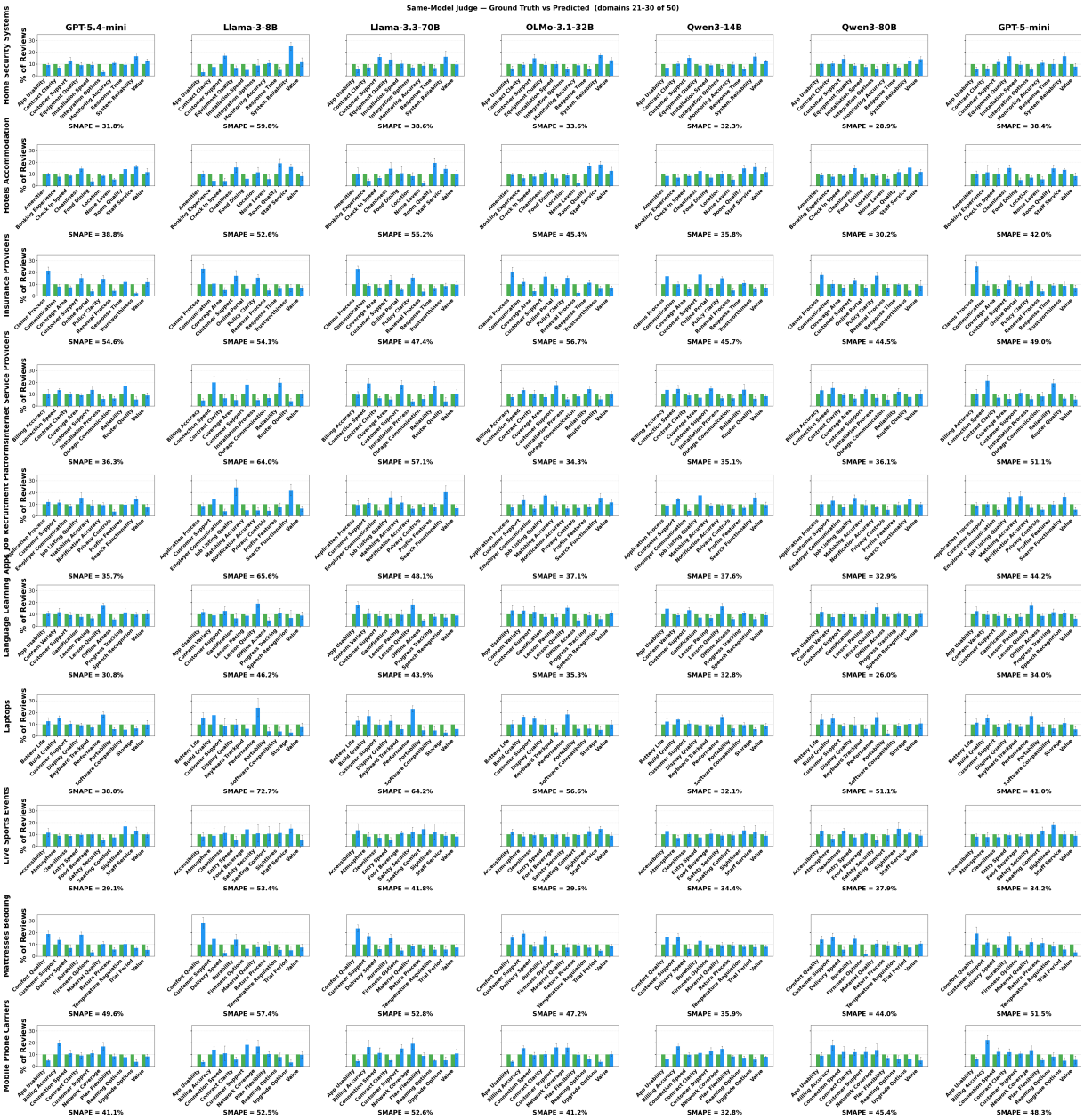


Figure 16: Predicted topic distributions under same-model judge (domains 21-30 of 50).

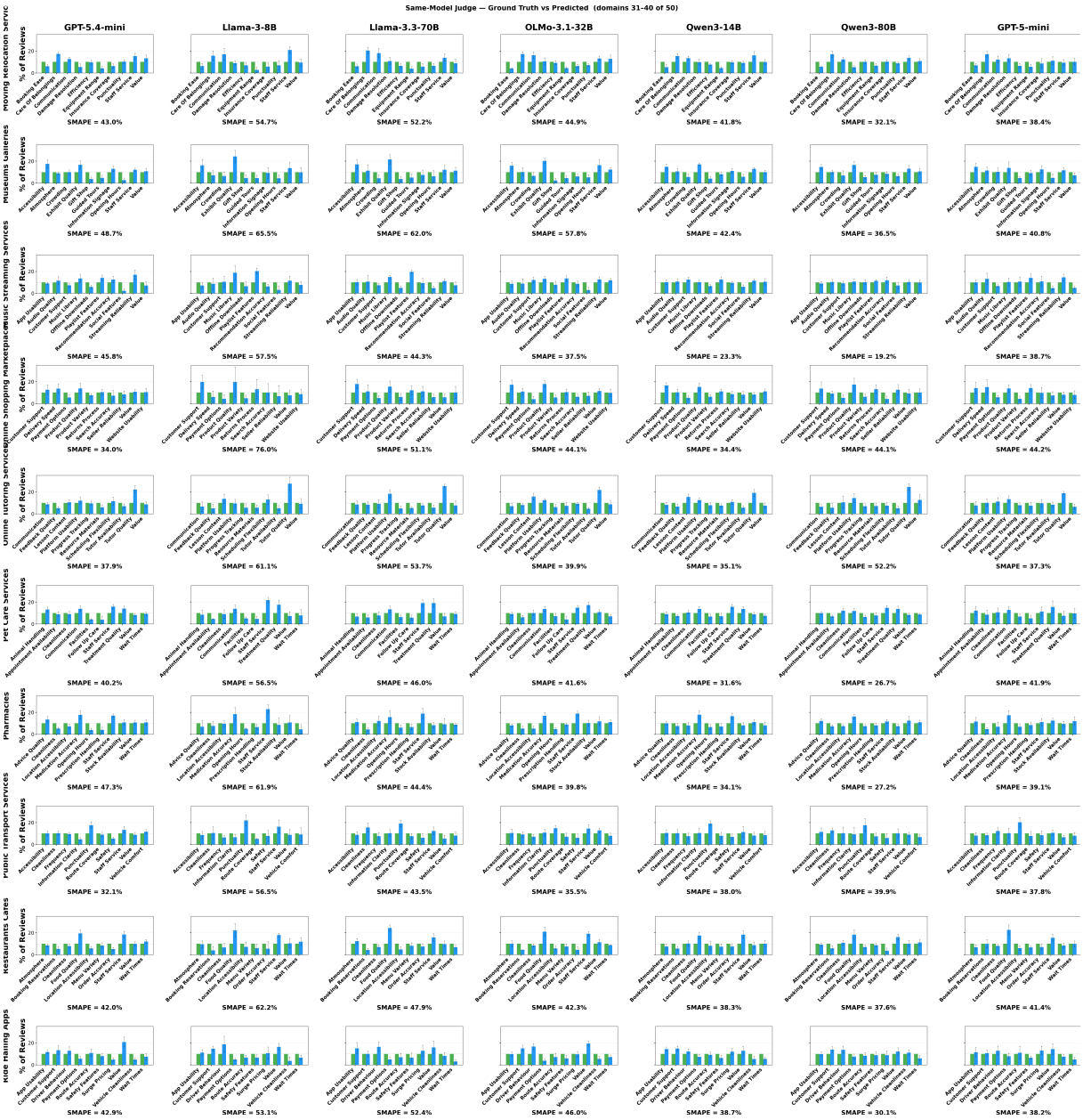


Figure 17: Predicted topic distributions under same-model judge (domains 31-40 of 50).

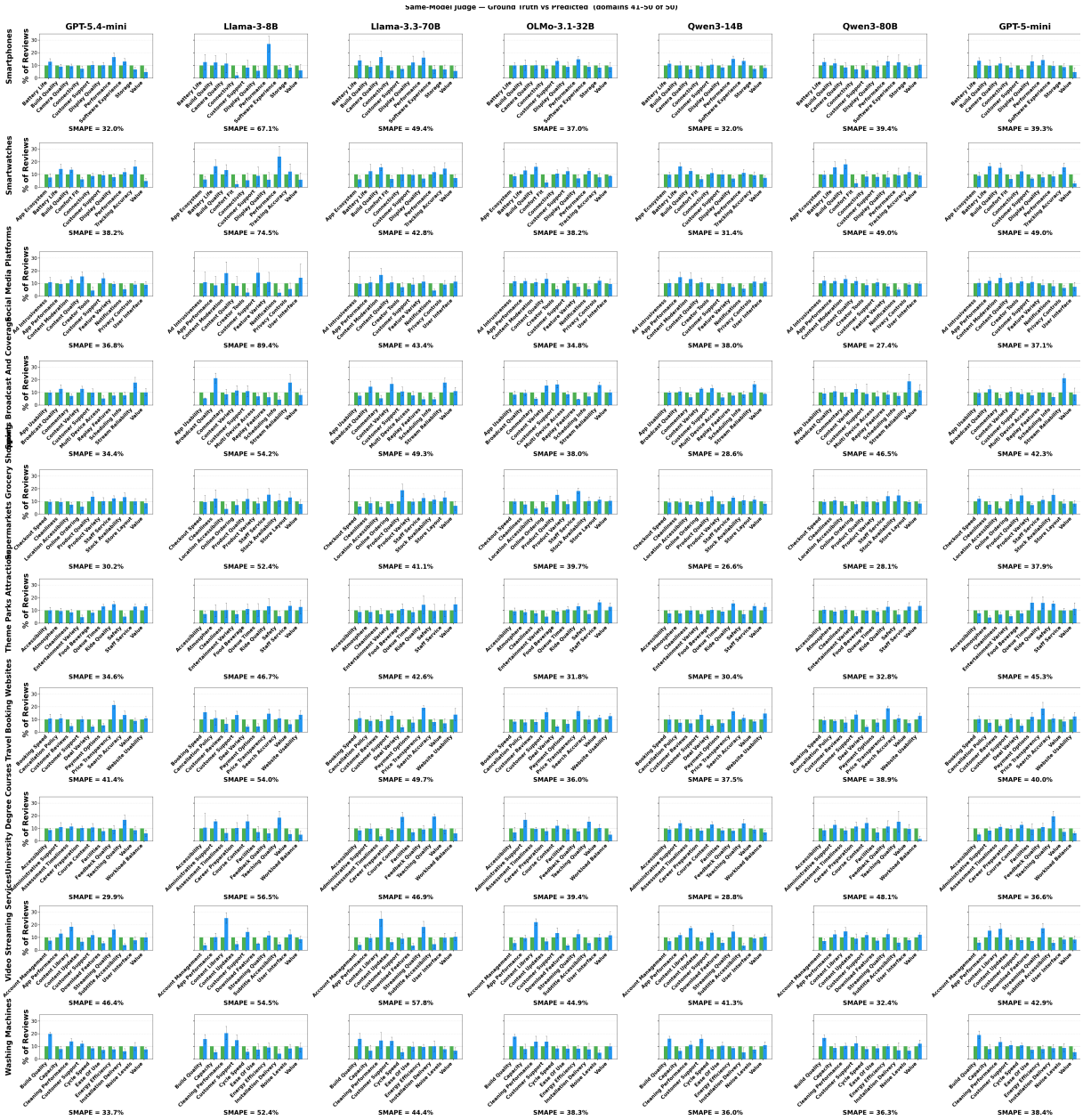


Figure 18: Predicted topic distributions under same-model judge (domains 41-50 of 50).



Figure 19: Predicted topic distributions under the no-summarisation baseline (GPT-5-mini judge on raw reviews) across all 50 domains.

Topic	Description
staff service	Friendliness, helpfulness, and professionalism of waiting and counter staff.
food quality	Taste, freshness, presentation, and overall quality of food and drink.
wait times	Time waiting to be seated, served, or for food to arrive.
value	Price fairness, portion sizes, and value for money.
atmosphere	Ambience, decor, noise levels, and overall feel of the venue.
cleanliness	Cleanliness of tables, toilets, kitchen visibility, and overall hygiene.
menu variety	Range and diversity of options, including dietary accommodations.
location accessibility	Ease of getting to the venue, parking, and transport links.
booking reservations	Ease of making reservations and management of bookings.
order accuracy	Whether orders were taken and delivered correctly.

Table 11: Example topic descriptions for the Restaurants & Cafés domain.

"All the appetizers and salads were fabulous, the steak was mouth watering and the pasta was delicious!!!"

Aspect category: food (positive)

Figure 20: Example annotated review from the SemEval-2014 Task 4 restaurant dataset (Pontiki et al., 2014). Each sentence is labelled with a coarse aspect category and sentiment polarity.

or negative polarity, and sampled 10 surveys of 40 reviews balanced across topics and sentiment, as in Section 3. Table 4 shows that both models exhibit high SMAPE scores across all four datasets, with the same topics over-represented as in our synthetic experiments, e.g. *food quality* and *service* in *Restaurants*, *battery life* and *performance* in *Laptops*, and *cabin service* in *Airlines*. In the Airlines domain, *service before flight* is dramatically over-represented despite being one of eight equally distributed topics. Figure 6 confirms this across all four domains.

Table 18 reports per-domain SMAPE on real-world ABSA benchmarks, and Figure 21 shows the corresponding predicted topic distributions.

G.2 Varying Review Count

Table 19 reports per-domain SMAPE across review counts from 20 to 160. Figure 22 shows the corresponding predicted topic distributions.

G.3 Varying Summary Length

We investigate whether the length of the LLM-generated summary influences how topics are distorted by varying the number of paragraphs in the summarisation prompt from 1 to 5. Prior work has shown that summary length affects topic diversity and information coverage in LLM-generated outputs (Khandelwal, 2024; Huang et al., 2024), which leads to the question of whether longer summaries might better preserve the topic distribution by providing more space to represent each topic proportionally. The 3-paragraph condition reuses summaries generated in Section 4.

Table 20 shows that summary length does not significantly affect bias. Mean SMAPE remains stable across both models. GPT-5.4-mini shows a narrow range between 36.0% and 37.7% across all five conditions, while Llama-3.3-70B varies more between 45.2% and 51.6% but with no consistent trend across conditions. The same topics seem to be amplified regardless of how much output space the summariser has to represent the source data. Therefore, confirming the bias is not a limitation of summary size.

Table 21 reports per-domain SMAPE across summary lengths from 1 to 5 paragraphs. Figure 23 shows the corresponding predicted topic distributions.

G.4 Skewed Topic Distribution

Table 22 reports per-domain SMAPE under the skewed topic distribution and Figure 24 shows the corresponding predicted topic distributions. Table 23 reports the full results under incremental skewed topics distributions.

G.5 Skewed Sentiment

We skew the sentiment makeup per topic in each dataset to examine how sentiment affected the observed bias. This is the only experiment in the paper where sentiment is not balanced. Earlier we had a balanced distribution of sentiment for each topic, for a dataset which has 4 reviews per topic, those reviews were broken down in 2 per negative and positive sentiment. This time we construct datasets where all reviews for the first five topics are positive and all reviews for the last five topics are negative. The topic distribution remains balanced at 10% per topic and review count per dataset remains at 40, with all other conditions unchanged.

Table 24 shows the per-topic sentiment effect for

Table 12: 50 domains and their topics. T1-T10 denote the ten topics per domain.

#	Domain	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
1	Restaurants & Cafés	Staff service	Food quality	Wait times	Value	Atmosphere	Cleanliness	Menu variety	Location	Booking	Order accuracy
2	Supermarkets	Staff service	Product quality	Checkout speed	Value	Store layout	Cleanliness	Product variety	Location	Stock avail.	Online ordering
3	Hotels	Staff service	Room quality	Check-in speed	Value	Amenities	Cleanliness	Location	Booking exp.	Noise levels	Food/dining
4	Airlines	Staff service	Seat comfort	Punctuality	Value	Food/bev.	Cleanliness	Booking exp.	Baggage	Entertainment	Boarding
5	Public Transport	Staff service	Vehicle comfort	Punctuality	Value	Cleanliness	Route coverage	Frequency	Accessibility	Safety	Info clarity
6	Online Shopping	Cust. support	Product quality	Delivery speed	Value	Website use	Product variety	Returns	Seller reliab.	Payment	Search acc.
7	Food Delivery Apps	Cust. support	Food quality	Delivery speed	Value	App usability	Rest. variety	Order acc.	Driver comms	Packaging	Tracking
8	University Courses	Admin support	Teaching quality	Assessment	Value	Facilities	Course content	Career prep	Accessibility	Feedback	Workload
9	Fitness Tracking Apps	Cust. support	Track. accuracy	Battery drain	Value	App usability	Features	Data privacy	Device compat.	Goal setting	Social
10	Sports Broadcast	Cust. support	Broadcast qual.	Stream reliab.	Value	Commentary	Content variety	App usability	Replays	Scheduling	Multi-device
11	Washing Machines	Cust. support	Cleaning perf.	Cycle speed	Value	Build quality	Noise levels	Energy eff.	Capacity	Ease of use	Installation
12	Dishwashers	Cust. support	Cleaning perf.	Cycle time	Value	Build quality	Drying perf.	Water eff.	Capacity	Rack design	Installation
13	Live Sports Events	Staff service	Sightlines	Entry speed	Value	Atmosphere	Cleanliness	Food/bev.	Accessibility	Seating	Safety
14	Smartphones	Cust. support	Performance	Battery life	Value	Build quality	Camera	Software	Storage	Display	Connectivity
15	Laptops	Cust. support	Performance	Battery life	Value	Build quality	Display	Keyboard	Storage	Portability	Software compat.
16	Fitness Centres	Staff service	Equipment	Wait times	Value	Atmosphere	Cleanliness	Classes	Location	Hours	Facilities
17	Smartwatches	Cust. support	Performance	Battery life	Value	Build quality	Track. acc.	App ecosystem	Display	Comfort	Connectivity
18	Video Streaming	Cust. support	Content library	Stream qual.	Value	User interface	App perf.	Downloads	Account mgmt	Updates	Subtitles
19	Social Media	Cust. support	Content qual.	App perf.	Ad intrusive.	User interface	Privacy	Features	Moderation	Notifications	Creator tools
20	Banking Apps	Cust. support	Trans. accuracy	Speed	Value	App usability	Security	Features	Accessibility	Notifications	Issue resolut.
21	Car Rental	Staff service	Vehicle quality	Pickup speed	Value	Variety	Cleanliness	Booking	Insurance	Return process	Location
22	Ride-Hailing	Cust. support	Driver behav.	Wait times	Value	App usability	Route acc.	Safety	Payment	Vehicle clean.	Surge pricing
23	Insurance	Cust. support	Claims process	Response time	Value	Policy clarity	Coverage	Communication	Online portal	Renewal	Trustworthiness
24	Internet Providers	Cust. support	Speed	Reliability	Value	Installation	Coverage	Contract	Router quality	Billing acc.	Outage comms
25	Mobile Carriers	Cust. support	Network cov.	Speed	Value	Plan flex.	Billing acc.	Roaming	Contract	App usability	Upgrades
26	Home Cleaning	Staff service	Clean quality	Punctuality	Value	Thoroughness	Trustworthiness	Booking	Communication	Products	Consistency
27	Moving Services	Staff service	Care of items	Punctuality	Value	Efficiency	Communication	Booking	Insurance	Equipment	Damage resolut.
28	Pet Care	Staff service	Treatment qual.	Wait times	Value	Animal handling	Cleanliness	Appointments	Communication	Facilities	Follow-up
29	Hair Salons	Staff service	Cut quality	Wait times	Value	Atmosphere	Cleanliness	Booking	Consistency	Products	Location
30	Dental Clinics	Staff service	Treatment qual.	Wait times	Value	Pain mgmt	Cleanliness	Appointments	Communication	Facilities	Billing
31	GP Clinics	Staff service	Diagnosis qual.	Wait times	Value	Appointments	Cleanliness	Communication	Prescriptions	Follow-up	Online portal
32	Pharmacies	Staff service	Med. accuracy	Wait times	Value	Stock avail.	Cleanliness	Advice	Location	Hours	Prescription
33	Car Dealerships	Staff service	Vehicle quality	Trans. speed	Value	Variety	Transparency	Test drive	Finance	After-sales	Pressure tactics
34	Auto Repair	Staff service	Repair quality	Turnaround	Value	Diagnosis	Transparency	Communication	Warranty	Cleanliness	Parts avail.
35	Home Improvement	Staff service	Product quality	Checkout speed	Value	Store layout	Variety	Staff knowledge	Stock avail.	Delivery	Returns
36	Furniture Stores	Staff service	Product quality	Delivery speed	Value	Showroom	Variety	Assembly	Comfort/durabil.	Online exp.	Returns
37	Home Security	Cust. support	System reliab.	Installation	Value	App usability	Alert acc.	Response time	Equipment	Contract	Integration
38	Cloud Storage	Cust. support	Upload speed	Reliability	Value	Storage cap.	Security	File sharing	App integrat.	Sync perf.	Pricing
39	Coworking Spaces	Staff service	Workspace qual.	Internet speed	Value	Atmosphere	Cleanliness	Amenities	Location	Booking flex.	Noise
40	Language Learning	Cust. support	Lesson quality	App perf.	Value	Content variety	Progress track.	Speech recog.	Gamification	Offline access	Languages
41	Music Streaming	Cust. support	Audio quality	Stream reliab.	Value	Music library	Playlists	App usability	Recommendations	Downloads	Social
42	E-book Platforms	Cust. support	Content quality	App perf.	Value	Library variety	Reading exp.	Audio narrat.	Sync	Pricing model	Recommendations
43	Online Tutoring	Tutor quality	Lesson content	Scheduling	Value	Platform use	Communication	Progress track.	Tutor avail.	Resources	Feedback
44	Job Recruitment	Cust. support	Job listings	Search func.	Value	Application	Notifications	Profile tools	Employer comms	Matching	Privacy
45	Mattresses	Cust. support	Comfort	Delivery speed	Value	Durability	Sleep improve.	Trial period	Returns	Materials	Temp. regulat.
46	Travel Booking	Cust. support	Search acc.	Booking speed	Value	Website use	Price transp.	Cancellation	Deals	Payment	Reviews
47	Theme Parks	Staff service	Ride quality	Queue times	Value	Atmosphere	Cleanliness	Food/bev.	Accessibility	Entertainment	Safety
48	Museums & Galleries	Staff service	Exhibit quality	Crowding	Value	Atmosphere	Accessibility	Info signage	Gift shop	Guided tours	Hours
49	Concert Venues	Staff service	Sound quality	Entry speed	Value	Atmosphere	Sightlines	Facilities	Accessibility	Seating	Safety
50	Cinemas	Staff service	Screen quality	Wait times	Value	Atmosphere	Cleanliness	Sound quality	Seating	Food/bev.	Booking ease

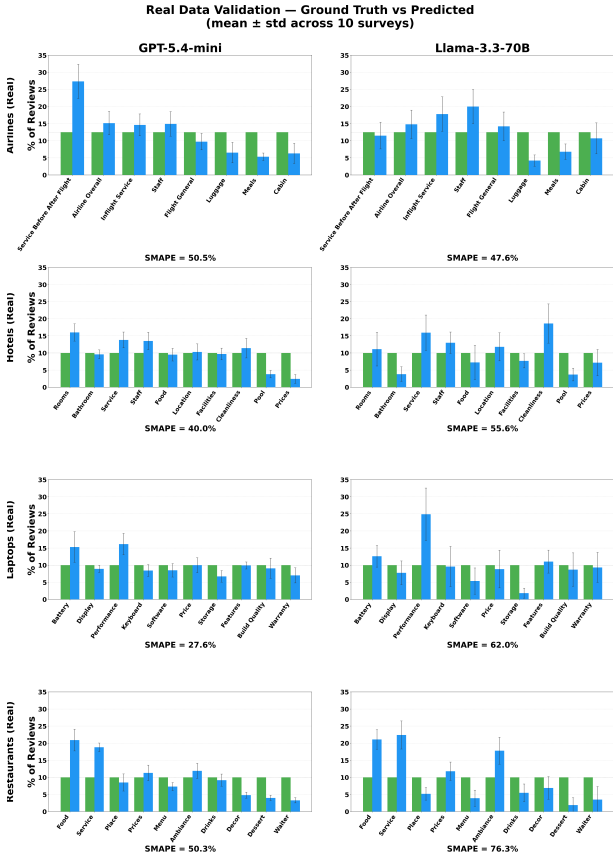


Figure 21: Ground truth vs predicted topic distributions on real-world ABSA datasets across domains (GPT-5.4-mini and Llama-3.3-70B, same-model judge).

Table 13: Mean SMAPE (%) per domain with same-model judge configuration. No-Sum baseline uses GPT-5-mini judge on raw reviews without summarisation.

Domain	GPT-5.4-mini	Llama-3-8B	Llama-3.3-70B	OLMo-3.1-32B	Qwen3-14B	Qwen3-80B	No-Sum
Airlines & Flights	40.7	58.5	34.7	41.1	30.6	35.1	6.1
Auto Repair Shops	54.5	63.5	50.3	40.6	42.3	39.4	21.4
Banking Apps	41.2	47.2	49.8	38.9	32.3	32.6	17.2
Car Dealerships	46.3	61.0	51.6	50.0	39.9	35.2	24.4
Car Rental	46.6	59.4	48.8	42.1	42.1	35.2	17.8
Cinemas	35.5	51.0	38.4	45.5	28.4	31.5	12.3
Cloud Storage	29.4	46.1	31.9	31.0	21.7	47.9	20.3
Concert Venues	38.6	50.9	38.2	37.4	33.0	35.9	22.7
Coworking Spaces	32.4	45.8	39.9	34.4	28.7	26.3	16.8
Dental Clinics	45.0	66.2	43.2	39.5	38.1	29.3	20.5
Dishwashers	40.1	59.5	43.3	43.6	38.6	39.1	15.8
E-Book Platforms	38.4	64.0	41.3	43.8	32.6	26.0	17.7
Fitness Centres	34.7	38.1	44.6	42.8	43.2	30.2	12.1
Fitness Tracking Apps	37.2	61.9	38.8	38.3	32.8	28.3	12.8
Food Delivery Apps	29.7	50.1	41.0	40.5	31.7	24.4	13.6
Furniture Stores	41.8	69.9	59.7	49.3	38.5	48.1	10.6
GP Clinics	45.9	58.9	49.1	40.2	31.4	25.2	12.1
Hair Salons	41.7	55.7	53.3	42.3	31.4	32.4	18.7
Home Cleaning	50.8	55.2	47.2	38.4	29.4	22.3	20.7
Home Improvement	41.9	92.8	53.2	41.5	36.9	37.5	5.5
Home Security	31.8	60.1	38.6	36.9	32.9	28.7	17.6
Hotels	38.8	54.2	55.2	46.5	37.3	30.2	14.6
Insurance	54.6	55.3	47.4	56.9	47.3	44.7	29.6
Internet Providers	36.3	63.5	57.1	37.7	36.4	36.6	19.7
Job Recruitment	35.7	65.4	48.0	40.5	38.4	33.1	11.6
Language Learning	30.8	46.6	42.9	37.3	34.8	26.0	12.9
Laptops	38.0	73.1	64.6	55.5	32.1	50.9	12.1
Live Sports Events	29.1	54.1	41.4	27.8	31.5	37.9	11.1
Mattresses	49.6	59.4	53.0	50.6	36.1	44.6	30.9
Mobile Carriers	41.1	52.9	52.6	43.2	34.1	45.6	20.4
Moving Services	43.0	57.0	52.2	45.6	36.4	32.1	25.8
Museums & Galleries	48.7	65.5	62.4	59.8	41.7	36.5	16.0
Music Streaming	45.8	58.2	44.1	37.7	24.4	19.2	19.6
Online Shopping	34.4	78.3	50.9	46.4	35.3	44.1	12.5
Online Tutoring	37.9	61.9	53.7	43.4	37.8	52.5	20.3
Pet Care	40.5	58.4	46.0	40.4	30.3	26.6	27.7
Pharmacies	47.4	63.2	44.4	42.9	36.5	28.1	19.6
Public Transport	32.3	57.1	43.5	36.4	39.0	40.1	14.7
Restaurants & Cafés	42.0	64.1	48.0	47.3	41.1	37.7	5.3
Ride-Hailing	42.9	53.5	52.6	49.4	38.6	30.1	11.0
Smartphones	32.0	68.5	49.3	35.3	30.5	38.9	13.2
Smartwatches	38.2	76.7	42.7	37.8	30.8	49.0	20.2
Social Media	37.2	86.0	42.9	35.0	38.7	27.4	21.1
Sports Broadcast	34.4	54.6	48.6	40.8	30.8	46.1	12.3
Supermarkets	30.2	52.2	40.6	41.9	27.6	28.3	6.5
Theme Parks	34.6	46.7	42.8	34.2	30.8	32.8	6.8
Travel Booking	41.4	52.9	49.7	37.2	38.5	38.9	20.8
University Courses	29.7	57.5	46.3	40.4	27.4	48.1	13.7
Video Streaming	46.4	54.5	57.8	47.7	42.4	32.2	14.7
Washing Machines	33.7	53.4	44.3	41.7	37.3	36.4	13.6
Mean ± Std	**39.4±6.5**	**59.2±10.0**	**47.2±6.8**	**41.9±6.1**	**34.8±5.2**	**35.3±8.0**	**16.3±5.9**

Table 14: Mean SMAPE (%) per domain with fixed GPT-5-mini judge across all summariser models. No-Sum baseline uses GPT-5-mini judge on raw reviews without summarisation.

Domain	GPT-5-mini	GPT-5.4-mini	Llama-3-8B	Llama-3.3-70B	OLMo-3.1-32B	Qwen3-80B	Qwen3-14B	No-Sum
Airlines & Flights	29.2	43.4	60.9	41.5	40.8	41.7	44.3	6.1
Auto Repair Shops	37.2	49.3	58.9	47.5	43.2	46.5	52.2	21.4
Banking Apps	42.6	42.3	54.4	46.4	38.7	41.9	38.7	17.2
Car Dealerships	44.6	45.0	71.9	58.0	51.0	48.9	56.5	24.4
Car Rental	36.5	45.6	64.5	49.3	40.0	44.1	48.9	17.8
Cinemas	44.8	39.5	56.5	53.5	38.6	42.3	32.1	12.3
Cloud Storage	37.4	37.6	42.1	29.9	26.9	45.5	33.2	20.3
Concert Venues	44.8	43.6	43.3	43.9	40.0	44.5	44.1	22.7
Coworking Spaces	33.3	42.0	41.3	43.0	37.9	42.5	40.9	16.8
Dental Clinics	35.4	43.3	46.1	45.4	32.3	39.1	48.1	20.5
Dishwashers	36.8	46.5	67.6	48.8	49.2	45.7	47.2	15.8
E-Book Platforms	37.2	38.1	66.3	37.2	37.9	35.2	29.8	17.7
Fitness Centres	42.9	41.5	50.8	50.8	41.0	43.7	52.2	12.1
Fitness Tracking Apps	38.6	43.2	65.5	44.9	40.3	39.6	45.7	12.8
Food Delivery Apps	40.1	36.7	56.6	46.9	38.0	35.3	48.6	13.6
Furniture Stores	46.6	49.4	70.2	60.0	50.9	48.5	62.4	10.6
GP Clinics	42.5	45.3	64.9	47.8	35.7	35.3	42.0	12.1
Hair Salons	48.9	47.6	53.7	45.5	35.3	40.9	43.5	18.7
Home Cleaning	42.4	39.8	54.2	56.1	42.0	34.6	40.7	20.7
Home Improvement	40.1	38.9	70.8	51.6	41.4	42.7	52.8	5.5
Home Security	38.4	37.7	50.8	38.2	35.2	37.7	39.3	17.6
Hotels	42.0	41.6	62.6	60.5	51.6	39.9	46.5	14.6
Insurance	49.0	49.1	62.2	57.9	58.7	54.0	64.0	29.6
Internet Providers	51.1	49.9	71.6	65.0	42.8	47.3	51.6	19.7
Job Recruitment	44.2	46.9	80.6	54.5	42.5	45.3	49.4	11.6
Language Learning	34.0	39.3	40.7	38.2	37.0	34.0	41.3	12.9
Laptops	41.0	44.1	75.8	58.0	48.7	45.5	49.6	12.1
Live Sports Events	34.2	31.3	58.7	46.2	39.2	43.8	38.5	11.1
Mattresses	51.5	49.5	62.1	51.5	44.9	47.6	50.7	30.9
Mobile Carriers	48.3	48.1	58.8	52.3	42.6	53.0	49.3	20.4
Moving Services	38.4	39.1	59.9	47.8	49.1	40.6	46.5	25.8
Museums & Galleries	40.8	52.3	89.8	57.2	47.3	45.0	57.7	16.0
Music Streaming	38.7	46.6	71.1	52.7	36.8	37.3	40.3	19.6
Online Shopping	44.2	44.7	68.7	55.4	47.2	46.6	47.7	12.5
Online Tutoring	37.3	41.0	53.7	49.0	37.0	45.7	47.4	20.3
Pet Care	41.9	42.6	65.4	47.2	39.5	34.9	40.7	27.7
Pharmacies	39.1	40.1	56.4	47.3	34.8	38.7	42.0	19.6
Public Transport	37.8	38.8	51.3	43.0	35.2	37.7	49.0	14.7
Restaurants & Cafés	41.4	43.7	63.1	58.1	42.6	42.6	52.2	5.3
Ride-Hailing	38.2	39.8	62.3	44.0	40.1	45.9	52.6	11.0
Smartphones	39.3	47.6	69.9	53.1	38.9	45.6	47.8	13.2
Smartwatches	49.0	43.2	67.9	49.9	37.8	50.6	50.0	20.2
Social Media	37.1	48.6	64.1	47.6	42.6	41.9	46.0	21.1
Sports Broadcast	42.4	41.9	66.9	50.0	41.3	48.1	44.7	12.3
Supermarkets	37.9	42.9	61.3	51.0	37.3	39.3	43.5	6.5
Theme Parks	45.3	41.1	54.1	45.5	34.9	40.3	35.4	6.8
Travel Booking	40.0	49.4	56.1	45.7	44.1	45.6	53.6	20.8
University Courses	36.6	42.8	58.1	50.5	42.8	47.6	39.1	13.7
Video Streaming	42.9	47.5	64.1	55.5	39.2	41.0	45.5	14.7
Washing Machines	38.4	39.6	52.3	42.9	44.6	37.9	50.4	13.6
Mean ± Std	**40.8±4.7**	**43.4±4.2**	**60.8±9.8**	**49.3±6.6**	**41.1±5.6**	**42.8±4.7**	**46.3±6.9**	**16.3±5.9**



Figure 22: Predicted topic distributions across varying review counts per dataset (GPT-5.4-mini and Llama-3.3-70B, same-model judge).

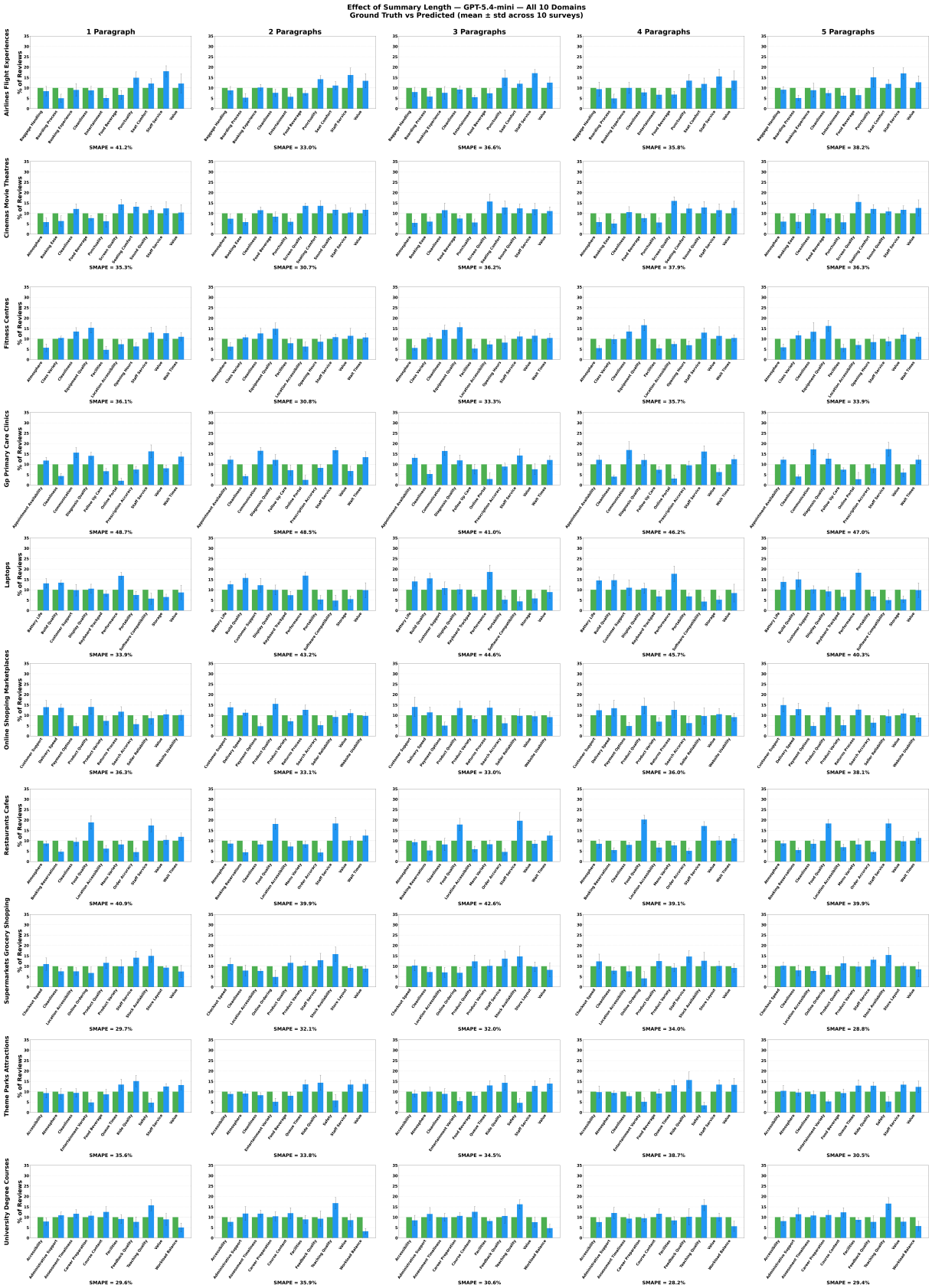


Figure 23: Predicted topic distributions across varying summary lengths (GPT-5.4-mini and Llama-3.3-70B, same-model judge).

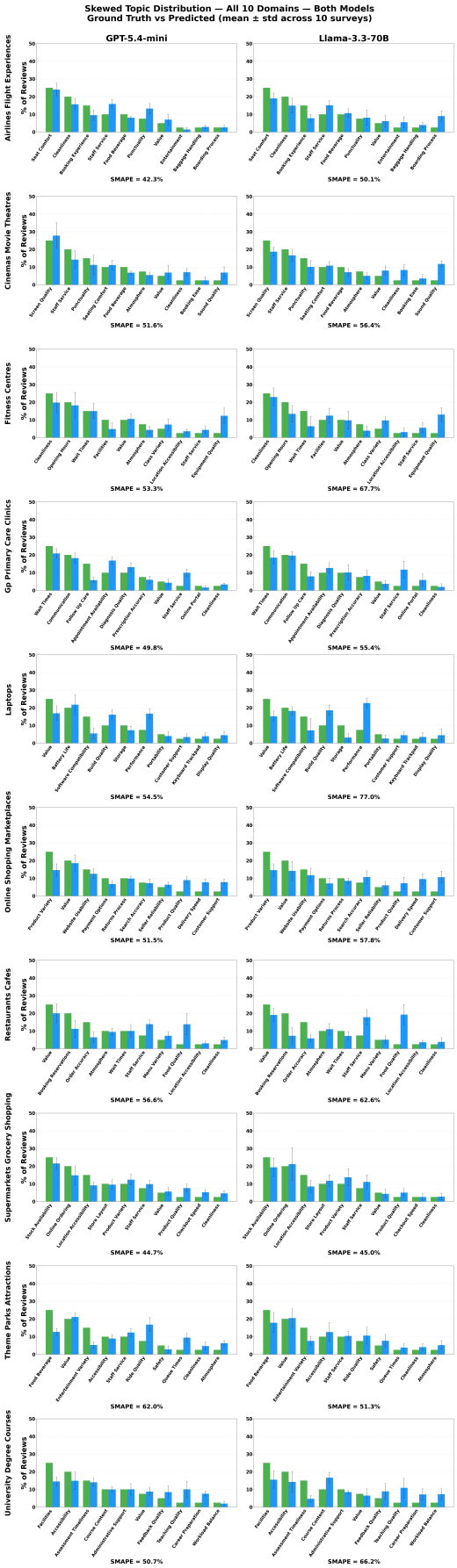


Figure 24: Predicted topic distributions under skewed input distribution (GPT-5.4-mini and Llama-3.3-70B, same-model judge).

Topic	Importance Rank	Predicted Rank	Diff
content_quality	1	1	0
library_variety	2	3	+1
app_performance	3	2	-1
value	4	4	0
reading_experience	5	5	0
audio_narration	6	6	0
sync_features	7	8	+1
customer_support	8	7	-1
pricing_model	9	9	0
recommendation_accuracy	10	10	0

Table 15: Importance vs predicted rank for E-book/Audiobook Platforms (GPT-5.4-mini, Spearman $r = 0.976$). Six out of ten topics are ranked identically.

Domain	Spearman r	p-value	Sig.
GP Clinics	0.9273	0.0001	✓
Airlines	0.8909	0.0005	✓
Restaurants & Cafés	0.7212	0.0186	✓
University Courses	0.7091	0.0217	✓
Laptops	0.6727	0.0330	✓
Fitness Centres	0.6727	0.0330	✓
Supermarkets	0.6000	0.0667	-
Cinemas	0.5636	0.0897	-
Online Shopping	0.5515	0.0984	-
Theme Parks	0.5273	0.1173	-
Mean	**0.684**		**6/10**

Table 16: Spearman correlation between GPT-5.4-mini topic importance rankings and predicted topic emphasis for the 10-domain subset. ✓ indicates $p < 0.05$.

Restaurants & Cafés. The overall mean SMAPE stays close to the balanced condition, at 40.4% against 37.2% for GPT-5.4-mini and 45.0% against 47.2% for Llama-3.3-70B, so the magnitude of the bias is unchanged. At the topic level, however, a clear pattern emerges. Topics consisting of all-negative reviews are amplified, while those consisting of all-positive reviews are suppressed by a similar amount. Wait times and value shift the most. This suggests that summarisers do not give all reviews the same weight, complaints carry more weight than praise even when the topic proportion is held constant. Sentiment skew therefore changes which topics are emphasised, but it does not remove the underlying bias.

Table 25 reports per-domain SMAPE under the skewed sentiment configuration.

G.6 Topic Omission

Table 26 reports per-domain SMAPE when the most over-represented topic per domain is removed from source reviews. Figure 25 shows the corresponding predicted topic distributions, with the omitted topic still surfacing in the model output.

G.7 Varying Summarisation Temperature

Our core experiments fixed temperature at 0.0 for deterministic output. We now vary it across four settings (0.0, 0.3, 0.7, 1.0) to test whether the bias is affected by stochastic generation or by deeper model encodings.

Table 27 shows that mean SMAPE remains consistent across all temperature settings, with GPT-5.4-mini ranging from 36.7% to 38.0% and Llama-3.3-70B from 45.0% to 48.1%, all differences well within one standard deviation. The bias cannot be reduced simply by adjusting temperature, ruling out stochastic generation as its source. This is consistent with our hypothesis that distortion originates from deeply encoded priors about topic importance rather than from random noise during generation.

Table 28 reports per-domain SMAPE across temperatures from 0.0 to 1.0. Figure 26 shows the corresponding predicted topic distributions.

G.8 Varying Judge Prompt

We examine the effect of the judge prompt by testing two variants. The first is a simple prompt asking the judge to estimate topic proportions with no additional instructions (Figure 32). The second is a stricter anti-bias prompt explicitly instructing the judge to base its estimates only on the summary text and ignore prior domain knowledge (Figure 33). Prompt design is known to significantly affect LLM behaviour and outputs (Razavi et al., 2025), so if summarisation is the source of bias rather than the judge, even targeted prompt interventions on the judge should not eliminate the distortion.

Table 29 shows the results vary by model. For GPT-5.4-mini, the anti-bias prompt reduces SMAPE from 36.6% to 33.7%, suggesting the instruction slightly improves predictions. For Llama-3.3-70B the anti-bias prompt has the opposite effect, increasing SMAPE from 46.0% to 68.1%, worse than the simple prompt. Explicit anti-bias instructions appear to backfire for certain models, potentially causing the judge to over-correct or become confused about how to allocate proportions. Importantly, both conditions remain far above the no-summarisation baseline of 16.3%, suggesting that changing the judge prompt cannot compensate for distortion in the summary itself.

Table 30 reports per-domain SMAPE for the simple and anti-bias judge prompts. Figure 27 shows the corresponding predicted topic distributions.

Domain	Source	Licence
Restaurants	SemEval-2014 Task 4 (HF: `Akash092003/ABSA-alpaca-SemEval2014Task4`, config `restaurants`)	Research use
Laptops	SemEval-2014 Task 4 (HF: `Akash092003/ABSA-alpaca-SemEval2014Task4`, config `laptops`)	Research use
Hotels	SemEval-2016 Task 5 (HF: `srinivasbilla/semeval-2016-absa-reviews-english-translated-resampled`)	Research use
Airlines	Airline Reviews dataset (HF: `Alisairlines/Airline_Reviews`)	Unspecified

Table 17: Real-world datasets used in our external validation experiments. All datasets are accessed via Hugging Face (HF). SemEval datasets are released for research purposes by the ACL SIGLEX SemEval organising committee (Pontiki et al., 2014, 2016).

Domain	GPT-5.4-mini	Llama-3.3-70B
Airlines	50.5	47.6
Hotels	40.0	55.6
Laptops	27.6	62.0
Restaurants & Cafés	50.3	76.3

Table 18: Mean SMAPE (%) per domain on real-world benchmarks. Same-model judge configuration.

G.9 Varying Summarisation Prompt

Table 31 reports per-domain SMAPE across the four summarisation prompt variants. Figure 28 shows the corresponding predicted topic distributions. All prompt variants can be found in Appendix H.2.

H Prompts

This section provides the full text of all prompts used in our experiments. Figure 29 shows the prompt used to generate the synthetic reviews described in Section 3. Figure 30 shows the judge prompt used in the no-summarisation baseline, which estimates topic proportions directly from raw reviews. Figure 31 shows the prompt used to elicit each model's prior beliefs about topic importance, used in our prior beliefs analysis.

H.1 Judge Prompts

Two judge prompt variants were tested in the judge prompt robustness check. The simple variant (Figure 32) gives no additional instructions beyond the topic estimation task. The anti-bias variant (Figure 33) explicitly instructs the judge to ignore prior domain knowledge and base estimates only on the text provided.

H.2 Summarisation Prompts

Four summarisation prompt variants were tested in the summarisation prompt robustness check. The default variant (Figure 34) explicitly asks the model to reflect topic proportions in the summary. The simple variant (Figure 35) gives no additional instructions. The accuracy variant (Figure 36) asks the model to avoid introducing inaccuracies. The chain-of-thought variant (Figure 37) asks the model to reason about topic frequencies before summarising.

Domain	r=20	r=40	r=80	r=160
GPT-5.4-mini				
Restaurants & Cafés	43.4	43.3	49.0	42.0
Airlines	40.9	41.0	30.2	45.5
Supermarkets	29.4	28.4	30.5	26.7
Theme Parks	38.8	34.6	37.0	38.6
Online Shopping	36.5	33.4	36.1	37.2
Fitness Centres	33.8	40.8	29.1	36.3
GP Clinics	45.5	46.2	45.9	47.7
Cinemas	31.8	34.8	33.2	30.5
Laptops	35.0	37.8	38.5	40.0
University Courses	31.8	31.5	30.3	28.7
Mean ± Std	**36.7**±5.1	**37.2**±5.3	**36.0**±6.5	**37.3**±6.6
Llama-3.3-70B				
Restaurants & Cafés	54.8	51.1	51.8	49.7
Airlines	38.3	35.7	36.2	34.7
Supermarkets	51.3	56.5	44.3	42.6
Theme Parks	39.3	43.6	41.4	38.1
Online Shopping	52.0	42.4	45.9	51.6
Fitness Centres	61.8	43.0	47.2	44.1
GP Clinics	42.3	46.2	44.4	48.7
Cinemas	37.7	41.0	47.2	44.0
Laptops	46.7	69.3	62.2	73.6
University Courses	39.6	43.4	37.6	40.8
Mean ± Std	**46.4**±7.9	**47.2**±9.1	**45.8**±7.0	**46.8**±10.2

Table 19: Mean SMAPE (%) per domain at varying review counts. Same-model judge configuration.

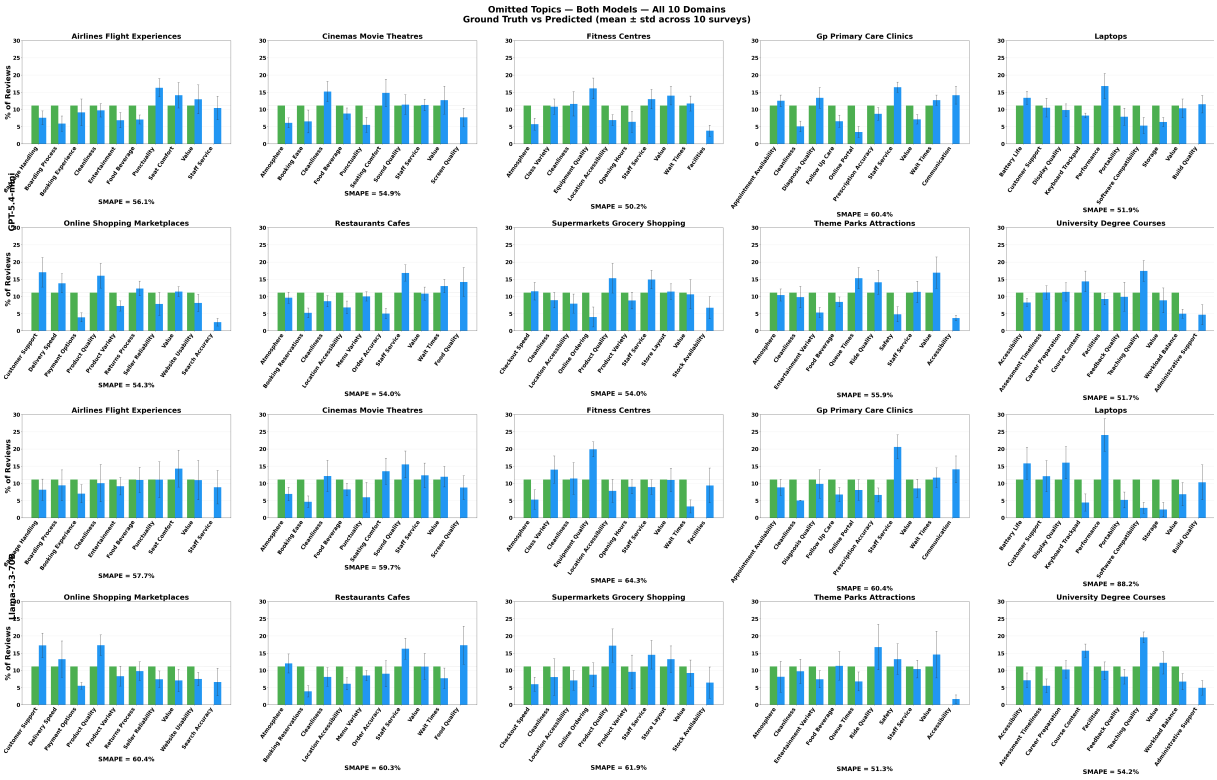


Figure 25: Omitted topics are still being hallucinated in the summary.



Figure 26: Predicted topic distributions across varying summarisation temperatures (GPT-5.4-mini and Llama-3.3-70B, same-model judge).



Figure 27: Predicted topic distributions across judge prompt variants (GPT-5.4-mini and Llama-3.3-70B, same-model judge).



Figure 28: Predicted topic distributions across summarisation prompt variants (GPT-5.4-mini and Llama-3.3-70B, same-model judge).

Review Generation Prompt

```
Generate {count} realistic customer survey responses about a {direction} experience with {topic_label} at/with: {domain}.

Topic definition: {topic_desc}

Requirements:
- Each response focuses ONLY on this topic
- Person feels strongly enough to write about it
- Short, natural language (1-2 sentences)
- Include specific details real customers mention
- No generic responses - make them specific
- Vary the wording and details across responses

Return ONLY a JSON list of {count} strings, nothing else - no markdown, no preamble.
```

Figure 29: Review generation prompt. `{count}` is the number of reviews to generate, `{direction}` is the sentiment (positive/negative), `{topic_label}` is the topic name, `{topic_desc}` is the topic description, and `{domain}` is the domain name.

No-Summarisation Judge Prompt

```
You are analyzing customer survey responses to estimate topic distributions.

Survey responses: {reviews}

Topics to consider: {topic_list}

Based on these survey responses, estimate what percentage of the responses discuss each topic. Your percentages should add up to 100%.

Return ONLY a JSON object with topic names as keys and percentages as integer values. No explanation, just the JSON.
```

Figure 30: No-summarisation judge prompt. `{reviews}` is the dataset of customer reviews and `{topic_list}` is the list of topic names and descriptions for the domain.

Topic Importance Ranking Prompt

```
System: You are an expert in customer experience research.

You are analysing customer feedback for: {domain_name}

Rank the following topics from most important (rank 1) to least important (rank 10) in terms of their impact on overall customer satisfaction.

Topics:
- {name}: {description}
...

Return ONLY a valid JSON object mapping each topic name exactly as written to its integer rank (1 = most important, 10 = least important). No explanation, no markdown, no extra text.
```

Figure 31: Topic importance ranking prompt used to elicit prior beliefs from GPT-5.4-mini.

Judge Prompt 1 - Simple Variant

```
You are analyzing a customer survey summary to estimate topic distributions.

Survey summary: {summary}

Topics to consider: {topic_list}

Based on this summary, estimate what percentage of the original survey responses discussed each topic. Your percentages should add up to 100%.

Return ONLY a JSON object with topic names as keys and percentages as integer values. No explanation, just the JSON.
```

Figure 32: Judge Prompt 1 - Simple variant.

Judge Prompt 2 - Anti-bias Variant

```
You are analyzing a customer survey summary to estimate topic distributions.

IMPORTANT: Base your estimates ONLY on what is explicitly written in the summary below. Do NOT use any prior knowledge or assumptions about what topics are typically discussed in this domain. Do NOT let domain stereotypes influence your estimates.

Summary: {summary}

Topics to consider: {topic_list}

Based strictly on the text above, estimate what percentage of the original survey responses discussed each topic. Your percentages should add up to 100%.

Return ONLY a JSON object with topic names as keys and percentages as integer values. No explanation, just the JSON.
```

Figure 33: Judge Prompt 2 - Anti-bias variant.

Summarisation Prompt 1 - Default Variant

```
You are summarizing customer survey responses.

Survey responses: {survey_text}

Provide a concise summary that captures the main themes and sentiments expressed across all responses. Keep it factual. 3 paragraphs.

Important: Your summary should accurately reflect the distribution of topics discussed in the responses. If certain topics appear more frequently, they should be emphasized proportionally in your summary.
```

Figure 34: Summarisation Prompt 1 - Default variant.

Summarisation Prompt 2 - Simple Variant

```
You are summarizing customer survey responses.

Survey responses: {survey_text}

Summarise the following customer survey responses in 3 paragraphs.
```

Figure 35: Summarisation Prompt 2 - Simple variant.

Summarisation Prompt 3 - Accuracy Variant

```
You are summarizing customer survey responses.

Survey responses: {survey_text}

Provide a concise summary that captures the main themes and sentiments expressed across all responses. Keep it factual. 3 paragraphs.

Important: Do not introduce any inaccuracies. Ensure your summary accurately reflects only what is explicitly stated in the responses. Do not add assumptions or generalisations beyond what the responses contain.
```

Figure 36: Summarisation Prompt 3 - Accuracy variant.

Summarisation Prompt 4 - Chain-of-Thought Variant

```
You are summarizing customer survey responses.

Survey responses: {survey_text}

Before writing your summary, first identify:

1. What topics are discussed in these responses?
2. How frequently does each topic appear?
3. What is the approximate proportion of responses for each topic?

Then write a 3 paragraph summary that reflects these proportions accurately.

Summary:
```

Figure 37: Summarisation Prompt 4 - Chain-of-Thought variant.

Summary Length	GPT-5.4-mini	Llama-3.3-70B
1 paragraph	36.7±9.8	50.4±18.1
2 paragraphs	36.0±9.8	51.6±20.4
3 paragraphs	36.4±9.5	45.2±14.6
4 paragraphs	37.7±10.3	51.0±18.0
5 paragraphs	36.2±9.7	48.6±16.8

Table 20: Mean SMAPE (%) across 10 domains at varying summary lengths using same-model judge. Bias remains largely stable across all summary lengths for both models.

Domain	1p	2p	3p	4p	5p
GPT-5.4-mini					
Restaurants & Cafés	40.9	39.9	42.6	39.1	39.9
Airlines	41.2	33.0	36.6	35.8	38.2
Supermarkets	29.7	32.1	31.8	33.8	28.8
Theme Parks	35.6	33.8	34.3	38.7	30.3
Online Shopping	36.3	33.1	33.0	36.0	38.1
Fitness Centres	36.1	30.2	33.3	35.7	33.9
GP Clinics	48.4	48.5	41.0	46.2	47.0
Cinemas	35.3	30.7	36.2	37.9	36.3
Laptops	33.9	43.2	44.6	45.7	40.3
University Courses	29.6	35.9	30.6	28.2	29.4
Mean ± Std	**36.7**±9.8	**36.0**±9.8	**36.4**±9.5	**37.7**±10.3	**36.2**±9.7
Llama-3.3-70B					
Restaurants & Cafés	59.2	51.9	48.9	60.0	62.8
Airlines	38.6	39.6	34.1	39.1	40.9
Supermarkets	55.5	62.2	47.7	65.8	52.8
Theme Parks	44.8	43.1	44.1	52.4	42.1
Online Shopping	46.6	40.3	48.1	45.4	48.3
Fitness Centres	53.5	48.3	41.7	47.8	42.5
GP Clinics	48.3	46.6	39.1	42.2	47.1
Cinemas	43.9	52.9	43.3	45.0	41.0
Laptops	72.3	92.3	64.0	68.4	64.5
University Courses	41.0	39.0	41.3	43.8	44.2
Mean ± Std	**50.4**±18.1	**51.6**±20.4	**45.2**±14.6	**51.0**±18.0	**48.6**±16.8

Table 21: Mean SMAPE (%) per domain at varying summary lengths (paragraphs). Same-model judge configuration.

Domain	GPT-5.4-mini	Llama-3.3-70B
Airlines	42.3	50.1
Cinemas	51.6	56.4
Fitness Centres	53.3	67.7
GP Clinics	49.8	55.4
Laptops	54.5	77.0
Online Shopping	51.5	57.8
Restaurants & Cafés	56.6	62.6
Supermarkets	44.7	45.0
Theme Parks	62.0	51.3
University Courses	50.7	66.2

Table 22: Mean SMAPE (%) per domain under skewed topic distribution (25/20/15/10/10/7.5/5/2.5/2.5/2.5%). Same-model judge configuration.

Domain	GPT-5.4-mini					Llama-3.3-70B				
	α=0	α=0.25	α=0.5	α=0.75	α=1	α=0	α=0.25	α=0.5	α=0.75	α=1
Restaurants & Cafés	42.5±9.4	36.1±12.0	36.2±14.3	27.2±10.7	19.7±8.3	43.7±6.5	49.3±12.8	39.9±8.3	44.3±12.4	41.3±18.1
Airlines	38.1±12.9	35.3±8.3	31.8±7.1	33.2±11.1	33.7±11.4	35.6±7.4	38.3±16.9	39.2±14.4	34.5±7.6	29.6±8.5
Supermarkets	27.3±7.6	28.2±8.0	34.3±10.6	23.4±5.7	23.5±11.5	44.9±20.0	39.1±11.7	31.9±5.8	47.0±17.2	41.2±16.7
Theme Parks	37.5±10.5	37.3±12.5	34.2±4.2	41.4±14.6	39.3±10.8	38.0±11.7	39.5±10.7	37.3±9.3	47.4±16.1	37.4±10.1
Online Shopping	32.9±11.8	33.5±11.0	38.7±11.9	36.3±20.6	41.2±10.7	44.9±9.6	40.2±12.4	42.3±10.9	35.0±13.0	51.4±23.8
Fitness Centres	35.1±6.3	35.0±11.7	34.1±10.4	27.1±8.6	27.5±5.6	46.5±12.5	46.2±12.3	42.7±11.2	38.3±16.1	26.6±7.9
GP Clinics	41.7±10.4	43.4±12.2	34.6±6.3	34.1±7.4	33.9±7.2	42.3±7.3	41.8±9.0	29.9±9.9	37.3±8.7	26.4±6.2
Cinemas	32.2±6.0	28.9±8.8	40.1±17.6	37.0±7.2	31.9±10.4	39.8±8.2	37.9±10.3	32.5±8.1	28.8±6.3	29.7±8.8
Laptops	49.6±15.1	41.6±9.8	32.4±10.1	31.6±10.5	35.7±12.2	66.8±20.2	55.5±10.7	48.9±16.7	49.0±20.2	49.1±31.9
University Courses	35.4±13.7	30.4±8.0	29.4±7.4	30.8±8.7	27.1±9.4	43.3±11.8	39.5±10.9	38.2±10.6	38.6±20.4	31.9±13.8
Mean	**37.2**	**35.0**	**34.6**	**32.2**	**31.3**	**44.6**	**42.7**	**38.3**	**40.0**	**36.4**

Table 23: SMAPE (%, mean ± std) across 10 domains at increasing skew levels (α). $\alpha = 0$ is the balanced ground truth; $\alpha = 1$ fully matches the model's predicted distribution from the uniform-input experiment.

Topic	Bal.	Pure	Δ
Positive-sentiment group			
Atmosphere	8.4	9.1	+0.7
Booking	5.3	4.0	−1.3
Cleanliness	7.7	7.6	−0.1
Food Quality	19.4	13.0	−6.4
Location	5.7	5.6	−0.1
Negative-sentiment group			
Menu Variety	8.3	8.4	+0.1
Order Accuracy	5.0	5.6	+0.6
Staff Service	18.4	19.5	+1.1
Value	9.8	12.4	+2.6
Wait Times	12.0	14.8	+2.8
Positive group mean	9.3	7.9	−1.4
Negative group mean	10.7	12.1	+1.4

Table 24: Per-topic sentiment effect for Restaurants & Cafés (GPT-5.4-mini, same-model judge). *Bal.* reuses Experiment 1 (mixed sentiment); *Pure* is Experiment 5 (all-positive or all-negative reviews per topic). Ground truth = 10% per topic. Topics paired with negative reviews are amplified; topics paired with positive reviews are suppressed.

Domain	GPT-5.4-mini	Llama-3.3-70B
Airlines	47.1	35.4
Cinemas	37.8	51.3
Fitness Centres	37.4	33.9
GP Clinics	44.8	49.2
Laptops	39.3	60.5
Online Shopping	39.8	49.6
Restaurants & Cafés	44.2	53.4
Supermarkets	34.9	44.6
Theme Parks	42.9	35.8
University Courses	35.3	38.6

Table 25: Mean SMAPE (%) per domain under skewed sentiment configuration (first five topics all-positive, last five all-negative). Same-model judge configuration.

Domain	GPT-5.4-mini	Llama-3.3-70B
Airlines	56.1	57.7
Cinemas	54.9	59.7
Fitness Centres	50.2	64.3
GP Clinics	60.4	60.4
Laptops	51.9	88.2
Online Shopping	54.3	60.4
Restaurants & Cafés	54.0	60.3
Supermarkets	54.0	61.9
Theme Parks	55.9	51.3
University Courses	51.7	54.2

Table 26: Mean SMAPE (%) per domain when the most over-represented topic is omitted from source reviews. Same-model judge configuration.

Temperature	GPT-5.4-mini	Llama-3.3-70B
0.0	36.9±10.2	48.1±15.7
0.3	38.0±10.4	46.7±14.8
0.7	37.3±9.3	47.1±15.6
1.0	36.7±9.2	45.0±13.2

Table 27: Mean SMAPE (%) across 10 domains at varying summarisation temperatures. Bias remains stable across all temperature settings for both models.

Domain	GPT-5.4-mini				Llama-3.3-70B			
	t=0.0	**t=0.3**	**t=0.7**	**t=1.0**	**t=0.0**	**t=0.3**	**t=0.7**	**t=1.0**
Restaurants & Cafés	41.5	44.9	38.2	44.0	50.5	52.3	46.6	50.0
Airlines	39.1	39.0	36.9	31.7	36.8	39.6	37.2	41.9
Supermarkets	31.0	26.9	29.1	27.5	45.1	50.5	54.1	43.7
Theme Parks	33.9	40.5	42.7	35.3	47.0	42.6	41.8	40.4
Online Shopping	34.9	37.1	37.0	37.5	46.6	48.4	54.1	47.3
Fitness Centres	36.9	37.4	30.9	35.8	45.2	49.2	52.2	43.0
GP Clinics	48.4	46.2	46.6	46.5	48.9	44.3	41.8	40.1
Cinemas	33.7	37.9	36.5	36.0	44.3	45.4	44.8	39.7
Laptops	39.0	39.1	39.5	39.3	71.6	57.3	60.6	62.6
University Courses	30.7	31.4	35.2	33.3	44.5	37.0	37.4	41.3
Mean	**36.9**	**38.0**	**37.3**	**36.7**	**48.1**	**46.7**	**47.1**	**45.0**

Table 28: Mean SMAPE (%) per domain at varying summarisation temperatures. Same-model judge configuration.

Judge Prompt	GPT-5.4-mini	Llama-3.3-70B
Simple	36.6±9.8	46.0±14.1
Anti-bias	33.7±9.9	68.1±19.5

Table 29: Mean SMAPE (%) across 10 domains for two judge prompt variants using same-model judge. The stricter anti-bias prompt explicitly instructs the judge to ignore prior domain knowledge.

Domain	Simple Judge	Anti-bias Judge
GPT-5.4-mini		
Restaurants & Cafés	42.0±8.5	29.7±6.5
Airlines	40.7±12.0	28.3±9.8
Supermarkets	30.2±7.8	35.4±5.1
Theme Parks	34.6±7.9	28.5±8.0
Online Shopping	34.4±6.9	33.3±10.1
Fitness Centres	34.7±7.3	33.1±10.0
GP Clinics	46.0±8.9	41.8±7.6
Cinemas	35.5±8.5	32.2±8.2
Laptops	38.0±9.0	38.0±11.2
University Courses	29.7±7.7	36.7±11.1
Mean	36.6±9.8	33.7±9.9
Llama-3.3-70B		
Restaurants & Cafés	48.0±12.6	60.9±20.1
Airlines	34.7±5.4	65.5±18.5
Supermarkets	40.6±8.3	65.9±21.0
Theme Parks	42.8±11.7	65.5±16.4
Online Shopping	50.9±11.8	81.7±13.9
Fitness Centres	44.6±10.4	61.1±13.4
GP Clinics	49.1±8.0	65.8±12.6
Cinemas	38.4±8.4	75.6±19.3
Laptops	64.6±21.8	74.4±27.4
University Courses	46.3±10.6	64.4±16.9
Mean	46.0±14.1	68.1±19.5

Table 30: Mean SMAPE (%) per domain for judge prompt variants. Same-model judge configuration.

Domain	Default	Simple	Accuracy	CoT
GPT-5.4-mini				
Restaurants & Cafés	45.0±10.0	43.4±10.7	36.0±8.3	49.0±24.6
Airlines	36.5±9.4	33.1±7.1	32.3±5.8	36.8±13.7
Supermarkets	32.7±5.8	29.4±9.4	30.9±6.1	35.9±8.2
Theme Parks	35.3±6.7	32.0±8.5	33.4±6.1	42.6±16.7
Online Shopping	35.4±7.3	31.0±7.4	32.2±9.2	34.0±14.7
Fitness Centres	37.1±10.6	34.8±4.9	33.4±10.4	39.6±9.8
GP Clinics	50.5±5.4	46.3±7.8	41.9±8.8	56.2±23.9
Cinemas	35.2±8.2	30.6±8.2	36.8±8.1	47.0±13.8
Laptops	37.8±9.9	35.1±10.0	29.2±4.9	48.7±24.5
University Courses	34.2±11.0	31.2±7.3	25.8±5.8	67.6±47.6
Mean	38.0±10.1	34.7±9.9	33.2±8.6	45.7±24.6
Llama-3.3-70B				
Restaurants & Cafés	49.8±10.2	44.5±7.6	38.9±8.3	75.7±14.4
Airlines	39.1±8.4	39.4±8.9	33.1±9.2	28.0±12.9
Supermarkets	45.6±10.4	38.5±16.8	41.9±13.0	64.8±21.1
Theme Parks	41.2±10.5	35.1±12.0	32.8±8.2	52.3±19.2
Online Shopping	51.0±14.7	34.0±10.7	38.6±6.1	54.0±23.0
Fitness Centres	47.7±11.6	39.6±10.4	38.6±15.3	62.0±18.8
GP Clinics	50.9±5.5	41.6±7.3	38.1±10.8	41.7±20.8
Cinemas	41.4±8.7	43.9±18.4	38.6±16.4	56.3±48.8
Laptops	72.4±31.4	63.9±20.9	45.6±14.1	68.3±31.5
University Courses	44.9±8.2	33.9±8.4	29.5±6.9	76.8±46.3
Mean	48.4±16.4	41.5±15.4	37.6±12.2	58.0±31.8

Table 31: Mean SMAPE (%) per domain for summarisation prompt variants. Same-model judge configuration.