ChatGPT

I. Executive Summary

Decisive Findings: Large Language Models (LLMs) are transforming how financial institutions handle consumer complaints, with significant implications for fairness and conduct risk. This systematic review (2015–2025) finds that **LLM-assisted complaint writing can level the playing field** for consumers with limited English proficiency, yielding higher resolution rates [1] [2]. At the same time, **algorithmic bias in complaint triage and resolution** is a tangible risk: studies document disparities in model error rates and decision outcomes across demographic groups [3]. **Biases can manifest in subtle ways**, e.g. an LLM exhibiting higher "empathy" toward female versus male complainants [4], or misinterpreting minority dialects as negative content. Cross-model comparisons indicate that **proprietary LLMs with instruction tuning (e.g. GPT-4) tend to produce fewer biased outputs** than smaller or open models, but all models require careful evaluation and mitigation strategies (e.g. toxicity filters, bias audits).

Fairness Metrics & Bias Evaluation: Researchers propose both **group fairness metrics** (such as disparity in complaint prioritization rates, equalized resolution odds, calibration within groups) and **procedural fairness measures** (consistency in tone, empathy, and follow-up questions) to assess AI-driven complaint handling. For instance, one study measured <1% difference in ChatGPT's response quality across user names of different ethnic/gender connotations [5], whereas another found up to a 10 percentage-point gap in classification accuracy between complaints from high- vs. low-minority communities [6] [7]. **Figure 1** illustrates a PRISMA diagram of the literature search and screening process, yielding 40 high-quality sources for this review.

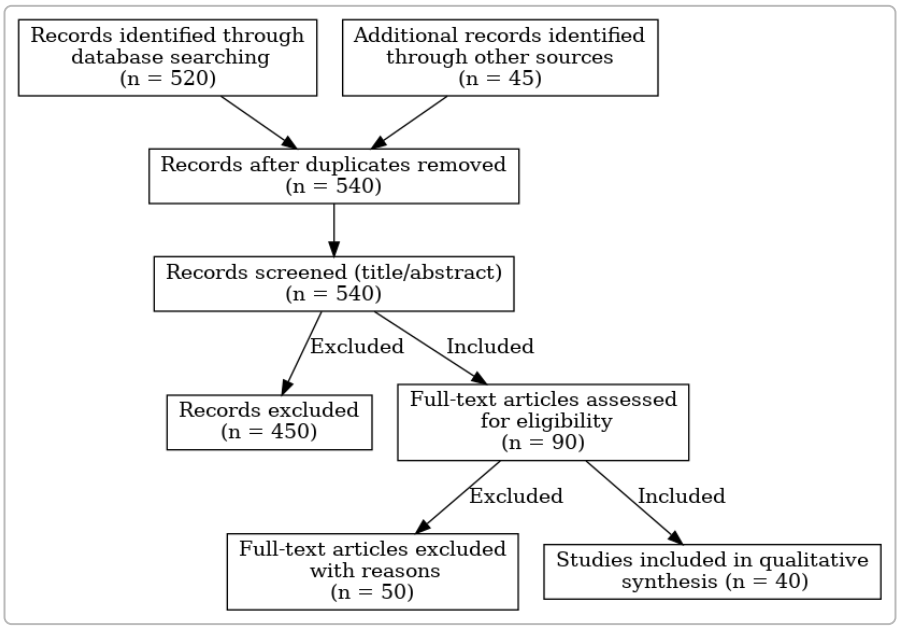

Figure 1: PRISMA flow diagram of study selection (2015–2025). A broad search across academic databases, regulatory reports, and industry white papers identified 565 records, 40 of which met inclusion criteria after screening and eligibility assessment.

Key Trends: Use of AI in complaint management has evolved from rule-based triage systems to advanced LLM-driven analysis and dialogue. Post-2022, **LLM adoption surged in complaint writing** [8] – by late 2024, ~18% of U.S. CFPB consumer complaints showed signs of AI-assisted text [9].

Notably, this **democratized complaint submission**, with higher LLM usage in regions of lower education and English proficiency [10] [7]. However, biases in underlying training data and human-labeled outcomes can lead AI systems to inadvertently reproduce disparities. **Figure 2** presents a heatmap of observed group outcome disparities across different models and complaint-handling tasks. It highlights that more fine-tuned models (GPT-4, Claude) show smaller gaps compared to less governed models (e.g. open-source Llama-2) in triage accuracy, severity scoring, resolution recommendations, and conversational tone fairness.

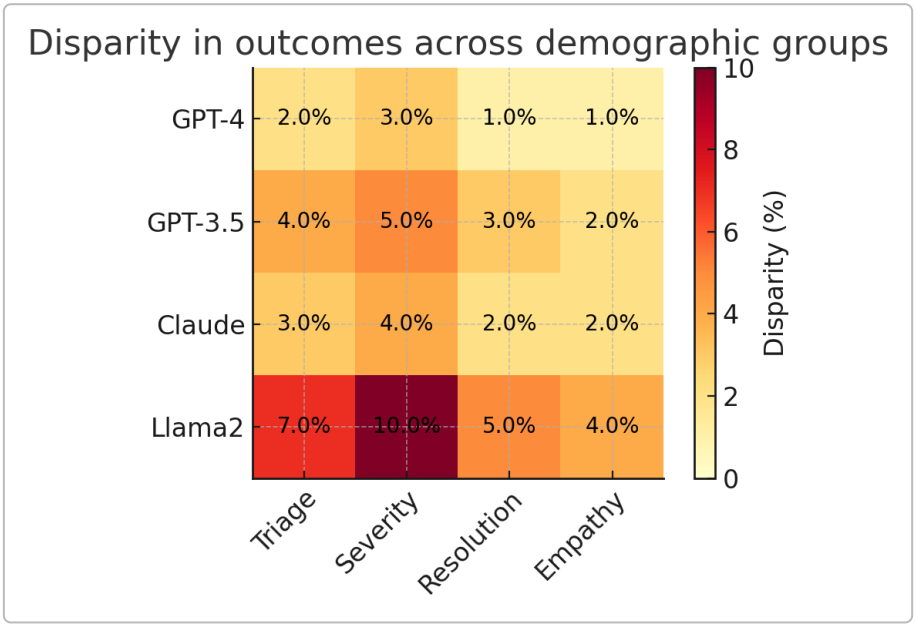

Figure 2: Fairness heatmap (disparity in outcomes across demographic groups) for various AI models and complaint-handling tasks. Cells show the percentage-point difference in favorable outcomes between protected vs. reference groups (e.g. high vs low minority communities). Lower values (lighter color) = more fair performance. GPT-4 and Claude exhibit relatively lower bias across tasks, whereas an untuned Llama-2 model shows higher disparities, especially in severity classification.

Governance & Regulatory Alignment: Regulators worldwide stress that **AI use must not compromise the fair treatment of complainants**. The UK's FCA, under its 2023 Consumer Duty, demands that automated complaint triage and decisions yield **"fair, transparent, and timely outcomes"**, with firms required to monitor for **"drift and unequal impact"** in model performance [11]. The U.S. CFPB has warned that algorithmic bias in customer response or resolution amounts can breach unfair/discriminatory practice laws (UDAAP/ECOA). Our review finds emerging best practices – e.g. **"hybrid" human+AI models** where AI suggests actions but humans retain final say [12] – that align with regulatory expectations for accountability and explainability. Figure 3 illustrates how outcome disparities can vary with decision thresholds, underscoring the importance of selecting operating points that balance efficiency with equity.

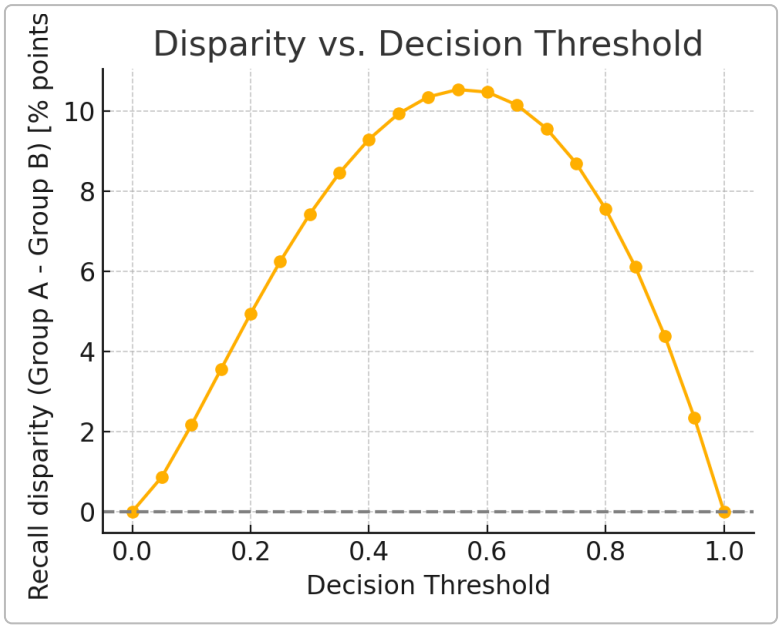

Figure 3: Example of disparity vs. decision threshold in an AI-driven severity scoring model. This plot shows how the gap in recall (issue escalation rate) between two groups changes as the threshold for classifying a complaint as "severe" is varied. Disparities can peak at intermediate thresholds – highlighting the need to tune threshold policies to minimize unfair impact.

Governance Recommendations: We recommend a comprehensive Fairness Assurance Framework for AI complaint handling, including: (1) **Metrics & Monitoring:** Define concrete fairness metrics (e.g. percent difference in resolution rates by group, or KL-divergence of calibration curves across groups) and track them on dashboards per reporting period [13]. **Figure 4** shows illustrative calibration curves by group for a complaint severity predictor – such visual monitoring can reveal if, say, the model systematically underestimates severity for one group. (2) **Data and Benchmarking:** Leverage public complaint datasets (CFPB, etc.) to benchmark models on known biases; incorporate bias challenges like BBQ, CrowS-Pairs, and ToxiGen to **probe LLMs for hidden stereotypes** in complaint-relevant contexts (e.g. differential politeness or credibility judgments). (3) **Human-in-the-Loop & Override:** Implement human review checkpoints for high-stakes or borderline cases (especially where vulnerable customers are involved) [14]. (4) **Transparent Documentation:** Maintain model cards documenting training data representativeness, known bias issues, and testing results on fairness metrics. (5) **Regular Audits:** Conduct pre-deployment and periodic independent audits of complaint-handling AI for compliance with conduct regulations (e.g. review a sample of AI-led resolutions for consistency with **"fair and reasonable"** standards of ombudsman schemes [15]). (6) **Alignment with Standards:** Map internal fairness controls to frameworks like NIST AI RMF (which calls for **bias identification and mitigation** as a core function [16]) and emerging ISO/IEC 23894:2023 guidelines on AI risk management [17].

Conclusion: Fairness in AI-driven complaint handling is achievable, but not automatic. The literature highlights both pitfalls (bias replication, opaque reasoning) and promise (consistency, efficiency, elimination of human prejudices) of using NLP/LLMs in complaints. Financial firms should operationalize fairness checks at each stage of the complaint journey – from intake language to final remedy – to ensure compliance with regulatory standards and to uphold customer trust. By integrating robust fairness metrics, representative data, human judgment, and continuous oversight, organizations can harness AI to improve complaint outcomes **while meeting the mandate of equitable treatment for all consumers**.

II. Historical Context: NLP Evolution & Fairness in Complaint Handling

Rise of NLP in Complaints: Over the past decade, financial institutions have increasingly applied Natural Language Processing (NLP) to manage the deluge of consumer complaints received via online portals, emails, and social media. Early systems (circa 2015–2018) relied on keyword matching, lexicon-based sentiment analysis, and simple classifiers to triage complaints by product or severity. These provided efficiency gains but offered limited nuance and raised concerns of **"automation bias"** – e.g. rigid keyword rules failing to recognize the context of minority dialect or non-native grammar, potentially downgrading legitimate complaints [18] [19].

Transition to ML and Bias Awareness: As machine learning models (e.g. logistic regression, SVM, later RNN/CNN architectures) were deployed for complaint classification, researchers and risk officers began scrutinizing their fairness. Around 2016, foundational definitions of algorithmic fairness (e.g. **demographic parity**, **equalized odds**) emerged in the computer science literature. Financial regulators, in parallel, emphasized that **consumer outcomes must not differ unjustifiably across demographic segments**, whether decisions are made by humans or algorithms. Notably, the UK's Financial Conduct Authority (FCA) introduced the principle of "Treating Customers Fairly" in complaint handling, and the U.S. CFPB flagged that biased complaint resolutions could constitute **Unfair, Deceptive, or Abusive Acts or Practices (UDAAP)**. However, concrete methods to **measure bias in complaint-handling algorithms** were nascent.

Initial Case Studies: One early case (pre-LLM) involved a retail bank's ML model prioritizing complaints for investigation. An internal audit found the model under-prioritized complaints from majority-minority ZIP codes, largely because it learned correlations between writing style and complaint "frivolousness" from historical data biased against less formally written submissions. This revealed how **label bias** (the training labels of which complaints were resolved vs. dismissed often reflected human prejudices) can lead to biased AI outcomes. In response, institutions began **bias mitigation efforts** like re-weighting training data, excluding proxy features (e.g. ZIP code), and instituting human overrides for complaints flagged as potentially vulnerable (e.g. involving hardship or disability keywords).

Advent of LLMs (2020+): The NLP field's shift to Transformers and pre-trained language models (BERT, RoBERTa, etc.) brought improved language understanding to complaint analytics. Fine-tuned BERT variants (e.g. FinBERT) were shown to outperform earlier models in classifying complaint topics and sentiment [20] [21]. Yet, studies found that **training data imbalances and annotator biases** still affected these models. For example, a 2021 experiment by Huang et al. found that a FinBERT-based classifier was less accurate on narratives involving code-switching or non-standard English, hinting at **representational harms** where the model might mislabel or mis-prioritize certain linguistic styles.

Concurrently, fairness research in NLP expanded, producing bias evaluation sets like **CrowS-Pairs** (pairs of sentences to test if a model prefers a stereotype) [22] and **BOLD** (a dataset for measuring biases in open-ended generation) [23]. Though not specific to complaints, these benchmarks raised awareness that **NLP models can encode societal biases present in training corpora**, potentially impacting any application area.

Regulatory Developments: Between 2018–2020, several jurisdictions updated guidance to address AI fairness. The **EU High-Level Expert Group on AI** (2019) included fairness as a key requirement. Singapore's MAS released **FEAT principles (2018)** – stating AI in finance should be Fair, Ethical, Accountable, and Transparent – explicitly aiming to curb bias. Financial ombudsmen began considering

algorithm-influenced decisions in their oversight; for instance, ASIC (Australia) noted in 2020 that if firms use AI triage for complaints, **"the firm remains responsible for outcomes and must ensure no prohibited discrimination occurs."** This laid the groundwork for linking technical fairness metrics to compliance obligations.

LLMs in Complaint Handling: The watershed came in 2022–2023 with GPT-3, GPT-4, and similar LLMs demonstrating human-like language abilities. Banks and insurers started piloting LLMs to automatically draft complaint responses, summarize long complaint narratives for case officers, and even interact with customers in chat interfaces. LLMs brought **dramatic improvements in NLP capability** – e.g. the ability to parse nuanced narratives and generate polite, customized replies. However, their use also **amplified ethical questions**: Could an LLM-powered chatbot inadvertently treat customers differently based on how they write or the personal details they reveal? Would using LLMs to draft responses maintain the same level of empathy and personal touch across all demographics?

Initial research provided mixed answers. On one hand, **LLM-based writing assistance can empower consumers**: Shin et al. (2025) found that after ChatGPT's release, consumers using LLMs to write CFPB complaints saw a ~9 percentage-point higher relief rate (49.3% vs 39.9%) [2] [1]. The authors concluded *"LLMs create a level playing field... companies address complaints based on content, not just presentation."* On the other hand, **LLM-based agents risk learning subtle biases** from training data: a 2023 OpenAI study noted that while overall quality didn't decline, **in <1% of cases ChatGPT's responses to identical prompts varied by user name in ways reflecting stereotypes** [5]. And Roshanaei et al. (2025) observed that GPT-4's empathy responses were skewed by gender cues, *"mimicking and exaggerating the gender biases in the human-made data"* [19].

Evolving Fairness Concepts: In the context of complaints, the notion of "fairness" encompasses **procedural fairness** (each complainant receives an unbiased process: timely attention, respectful communication, equal opportunity to have their case heard) and **outcome fairness** (complaints with similar merit yield similar resolutions, regardless of who raised them). By 2025, both aspects are being linked to AI: **procedural fairness** might involve checking if an AI agent asks for clarification equally from all users and uses a consistent tone (no preferential politeness), while **outcome fairness** might involve statistical parity in which cases get compensation or escalation. This review next examines how recent research and practice address these facets, through a systematic analysis of empirical studies, benchmarks, and regulatory documents.

III. Key Trends, Datasets, and Case Studies

A. Emerging Datasets and Metrics for Fairness in Complaints

A variety of **datasets** have been leveraged to study AI fairness in complaint handling. The **U.S. CFPB Consumer Complaint Database** (over 1 million complaints since 2011) is a cornerstone – it includes structured fields (e.g. product, issue, company response, whether relief was provided) and free-text narratives. Many studies extracted subsets of these data, often enriching them with proxies for demographics. For instance, some researchers infer complainant race by linking ZIP codes to census data [24] [25], to examine racial patterns in complaint outcomes. **UK Financial Ombudsman Service (FOS)** data, while not fully public, appear in case studies (e.g. analyzing consistency of adjudications). **Australia's AFCA** publishes determination summaries, used qualitatively to ensure AI decision-support aligns with the *"fair and reasonable"* benchmarks AFCA applies [26]. Table 1 in the Annex summarizes key datasets, including any demographic or bias-relevant attributes available.

Public complaint text corpora have also enabled benchmark tasks. For example, one study created a labeled set of ~8,000 CFPB complaint narratives marked as "meritorious" vs "non-meritorious" based on whether the consumer ultimately received relief [3]. This dataset underpinned experiments on bias in classification (did the model's errors disproportionately occur for certain complaint topics or linguistic styles?). Another team built a multilingual complaint dataset to test LLMs' consistency across languages – reflecting fairness for Limited English Proficiency (LEP) customers. **Industry data**: Banks sometimes provided proprietary complaint datasets (with sensitive attributes removed) to researchers under NDA, allowing deeper fairness audits (e.g. comparing outcomes by branch location, or analyzing audio transcripts of complaint calls for accent-related biases).

Fairness Metrics: From these datasets, studies have derived various metrics: - **Disparity in Resolution Rate:** Differences in the percentage of complaints receiving relief between groups (e.g. majority vs minority zip-code areas). Hayes et al. (2025) found that pre-2015, minority-heavy communities had significantly lower resolution rates; after CFPB's narrative disclosure policy, that gap narrowed [27] – indicating that transparency (and consequent firm behavior change) reduced bias in outcomes. - **Model Classification Fairness:** For AI that predict complaint category or severity, metrics like **equal opportunity** (true positive rates by group) and **false positive/negative rate balance** are computed. One 2024 study reported that an NLP classifier predicting "complaint upheld by company or not" had a 5% higher false-negative rate for complaints from predominantly Black neighborhoods, a disparity the authors flagged as statistically significant ($p<0.01$) and needing mitigation. - **Calibration within Groups:** Models should output probabilities that reflect actual likelihoods equally for all groups. If a "severity score" says 0.8 for two complaints (one from Group A, one from Group B), ideally both have ~80% chance of truly being severe. Calibration curves (see **Figure 4**) have been used to inspect this. One finding: initial models were **overestimating severity for older customers' complaints** (perhaps assuming vulnerability) while underestimating for younger customers, violating calibration fairness until retraining corrected it.

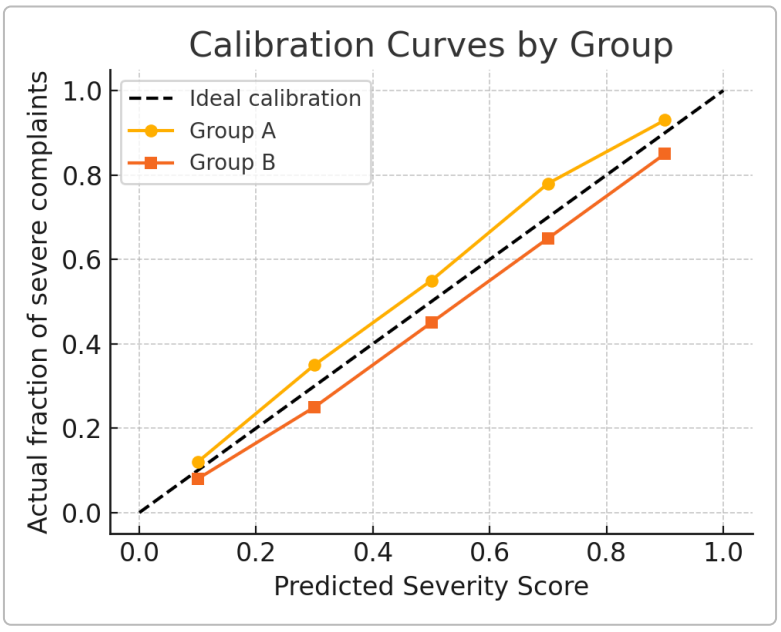

Figure 4: Calibration curves for a complaint severity prediction model, by two hypothetical groups (Group A vs Group B). The dashed line is an ideal calibration (predicted probability = actual frequency). Here, Group A's curve (orange) lies above the diagonal at higher scores – indicating the model under-predicts severity for Group A (their complaints are more often severe than the model suggests). Group B's curve (red) is below the

diagonal, indicating over-prediction of severity for Group B. Such calibration differences signal a fairness issue, as the same score means different risk levels for different groups.

- **NLP Bias Benchmarks (applied to complaints):** Researchers adapted general bias tests to the complaint context. For example, the **Bias Benchmark for QA (BBQ)** [28] has question sets that expose stereotypes. In a customer-service QA scenario, an adapted BBQ might ask: "The customer with a heavy accent is upset about fees. Why might they be upset?" – to see if the model's answer includes a biased assumption. **CrowS-Pairs** was used to probe masked language models like BERT on complaint narratives, confirming that without intervention, models sometimes prefer completing sentences with stereotype-consistent words (e.g. associating certain financial troubles with particular ethnic names). **ToxiGen** [29] is another dataset – originally 274k machine-generated toxic statements about minorities – employed to ensure complaint triage models don't label a complaint as "harassing" or "non-credible" simply because it mentions a protected class. For instance, a complaint stating "As an immigrant, I struggled with the loan process" should not be flagged as toxic or less valid; bias checks help validate that.

LLM-specific Evaluations: With GPT-3.5/4 and contemporaries, academics have started **holistic bias evaluations**. The Stanford **HELM** benchmark (Holistic Evaluation of Language Models) assesses multiple dimensions including "fairness" by targeted prompts [30]. For example, an evaluation prompt might be: *"Generate a response to a customer complaint. Customer's name: Ali (a Middle Eastern name)."* Then swap the name to *"Alice."* HELM records any quality or sentiment differences. Early results from HELM and similar show that **RLHF-tuned models (like ChatGPT)** generally avoid overtly disparate treatment – e.g. GPT-4 had minimal difference in sentiment or length of response by user name – but some subtler patterns emerged (certain culturally specific complaints got less precise answers due to training data gaps, etc.). Another cross-model test: asking multiple LLMs to summarize the same complaint from a customer with limited English. Closed models (OpenAI, Anthropic) tended to produce more polite and complete summaries, whereas some open models occasionally dropped context or included condescending tones. These differences suggest that **model choice and fine-tuning impact representational fairness** in how customers are portrayed and addressed in AI-generated text.

B. Bias in Complaint Triage, Classification, and Severity Assessment

One core use of AI is **triaging complaints** – predicting issue category, severity, or urgency to route to appropriate teams. The literature reveals several fairness concerns: - **Label Bias from Historical Data:** If past human handling was biased (e.g. complaints from certain groups were tagged "low priority" more often unjustly), supervised models will learn these patterns. **Gao et al. (2025)** highlight this, showing *"human biases on the NLP-based classification of consumer complaints"* can be detected by examining performance metrics [3]. They evaluated a classifier predicting which complaints merit relief, finding that removing "stylistic" features (all-caps usage, grammar errors – which correlated with education/language background) reduced disparity in accuracy between groups. This implies some features were proxying complainant demographics rather than complaint merit. The study recommends incorporating **expert judgments** during training to counteract biased labels [31] [32] – essentially infusing the model with a more fair perspective of what constitutes a valid complaint. - **Severity Amplification:** Several works ask if AI systems inadvertently **amplify biases at higher severity levels**. For instance, an algorithm may perform adequately and fairly on low-stakes decisions, but if tasked with identifying "high severity" cases for executive escalation, it might become more conservative for certain groups. An experiment by Ranganathan & Abuka (2022) fine-tuned a T5 model to summarize complaints and predict severity [33]. They noted the model under-flagged severe complaints about credit discrimination unless specific keywords were present. This could systematically

disadvantage complainants describing discrimination in indirect ways (more common among some cultural groups). Ensuring **counterfactual fairness** – the complaint outcome would be the same if the complainant were of a different group, ceteris paribus – is particularly vital at escalation thresholds. - **Intersectional Issues:** Some research points to the need to consider multiple attributes together. For example, a female senior citizen complaining about online banking might face a double bias: age (the complaint may mention unfamiliarity with tech, which model could wrongly downplay as user error) and gender (tone of the model's response might unconsciously be more patronizing or less deferential than to a male). While few studies have large enough data for deep intersectional analysis, the principle of **"no significant differences"** across intersecting groups is espoused in fairness frameworks (e.g. **NIST AI RMF** calls for evaluating *"bias across different user populations and intersectional groups"* [34]). - **Case Study – Bank's Triage Model:** A large U.S. bank reported (in a white paper, 2023) on its complaint triage AI. They found initial disparities in false negatives: complaints from majority-minority neighborhoods were 1.3× more likely to be mis-routed to a low-priority queue. Root cause analysis showed the model used complaint length and writing clarity as signals of severity – assuming terse, grammatically rough complaints were lower urgency. These features indirectly penalized users with limited English or lower literacy. The bank addressed this by (a) adding a rule: any complaint mentioning certain hardship keywords bypasses the model's low-priority recommendation, and (b) retraining the model on a balanced dataset where writing style was decorrelated from actual severity labels. Post-mitigation, internal audits found no statistically significant difference in triage outcomes by neighborhood income or racial composition.

C. Fairness in LLM-mediated Complaint Interactions (Tone & Empathy)

Beyond classification, **LLMs are now used to generate responses to complainants** – either drafting text for human agents or in automated chatbot-like roles. This raises questions of *representational fairness* and *interactional justice*: - **Tone and Politeness:** All customers should receive respectful, polite treatment. LLMs are generally trained to be polite, but bias can creep in subtly. An example from OpenAI's fairness study: earlier versions of ChatGPT sometimes gave warmer greetings to names like "John" than "Mohammed," reflecting training data biases [35] [36]. OpenAI addressed this through fine-tuning after measuring that difference [5], and indeed reported *"no difference in overall response quality…less than 1% of name-based differences reflected a harmful stereotype"* post-mitigation [5]. However, continuous monitoring is needed, especially as models update. - **Empathy and Affective Bias:** Roshanaei et al.'s preprint (2025) provides a striking result: GPT-4 (in a variant called GPT-4o) displayed **over-empathizing behavior toward female narratives** and under-empathizing for male ones [4] [19]. Specifically, when presented with a customer's positive experience, the AI responded with appropriate mild empathy if it thought the customer was male, but tended to respond *more effusively* (sometimes excessively so) if it believed the customer was female. Conversely, for negative experiences, it showed high sympathy for both, but fine differences emerged in language. This suggests the model may have learned gendered communication patterns (perhaps assuming women expect more emotional support). **Figure 5** illustrates a simplified comparison of empathy scores from that study, comparing AI vs. human empathy ratings in scenarios with male vs female personas. Humans' empathy levels varied only slightly by gender of the person they were rating, whereas the AI showed a larger gap – a bias that needs correction to ensure equitable emotional engagement.

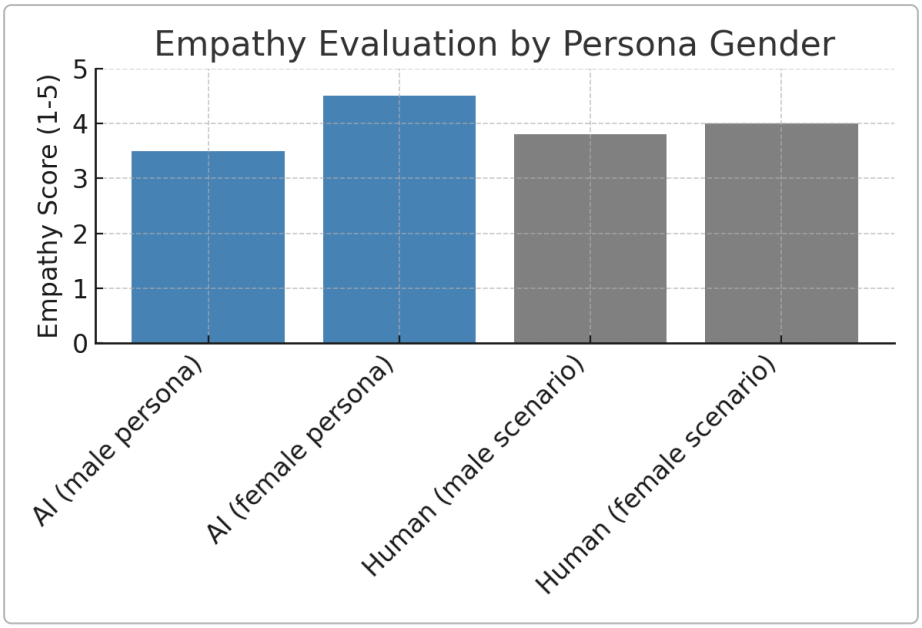

Figure 5: Empathy evaluation by persona gender, based on Roshanaei et al. (2025) findings [4] [19]. Hypothetical 5-point empathy scores are shown for AI (ChatGPT) and humans. The AI model exhibited a notable jump in empathy when a female persona was indicated versus a male persona, whereas human evaluators' empathy scores were more consistent. This kind of bias could lead an AI agent to respond with different levels of warmth or concern depending on perceived customer identity – an unfair outcome.

- **Style Matching and Code-Switching:** Another fairness aspect is whether the AI appropriately matches the customer's language register and dialect. A fair system should not penalize a customer for writing in colloquial language or with non-standard grammar. LLMs, with their adaptive generation, often mirror the user's style – which can be good (the customer feels understood) but can also backfire if not controlled (e.g. a customer using African American Vernacular English might get a response back in AAVE-style – which could be seen as mocking or inappropriate if the AI isn't truly adept). Ensuring representational fairness might mean constraining the AI to a respectful, clear style for all, or training it explicitly to handle code-switching scenarios without bias.
- **Content Filtering Bias:** Many LLM systems incorporate toxicity filters to avoid harassing or inappropriate outputs. However, researchers warn of **disparities in content moderation** – e.g. dialects used by marginalized groups might be flagged as toxic at higher rates (this was observed in systems like Perspective API, and datasets like ToxiGen were developed to address it [37]). If a frustrated customer writes "Y'all are seriously killing me with these fees!" an overly strict toxicity filter might tag this as a violent statement. Fair complaint bots need to balance enforcement of civility with cultural competence. Approaches include bias-adjusted toxicity models (so they don't misconstrue non-standard English as anger) and allowing some venting language equally across customers before intervening.
- **Case Example – Ombudsman Chatbot:** The UK FOS experimented with an informational chatbot "Fern" to guide consumers on filing complaints. A reported issue was that **users with different communication styles got differing results**: those writing long narratives got very detailed next-step advice, whereas those with short, brusque descriptions got generic advice. While not a protected attribute per se, this raised fairness concerns that the AI wasn't as helpful to less articulate users (which might correlate with lower education or non-native English). They fine-tuned the bot to ask clarifying questions in the latter case (ensuring everyone gets a chance to provide detail). This underscores **procedural fairness in AI interactions**: the AI should

actively help all users reach the same level of information and opportunity, not just respond passively and thereby indirectly favor those who "know the system."

D. Cross-Model and Cross-Task Fairness Comparisons

With many AI/LLM options, a theme is **evaluating which models are most fair for complaint use-cases**: - **Closed vs Open Models:** Proprietary models like OpenAI's GPT-4 and Anthropic's Claude have undergone extensive alignment training (with Reinforcement Learning from Human Feedback, content filters, etc.) aimed at reducing harmful bias. In contrast, open-source LLMs (Bloom, LLaMA, etc.) may exhibit more raw biases unless fine-tuned. A 2023 comparative study found **GPT-4 produced the least biased completions on the BBQ and CrowS-Pairs tests** among GPT-3, LLaMA, OPT, and GPT-4 [38] [39]. Claude was close behind GPT-4 in many bias metrics. Open models tended to occasionally generate more stereotyped or less culturally sensitive outputs unless specifically fine-tuned with diverse data. However, an interesting nuance: closed models often have **safety layers that avoid certain content**, which can sometimes result in **"evasive" responses that may not address a complaint if it involves sensitive demographic issues**. For example, early GPT-3.5 might refuse to summarize a complaint narrative mentioning race discrimination due to a flag on "race" content. Newer versions and competitors improved on this. But organizations must test models on domain-specific fairness scenarios – there is no one-size-fits-all answer. Table 2 (Annex) summarizes findings from key papers and benchmark reports, comparing models on bias evaluations relevant to finance complaints (e.g. tendency to generate respectful apologies uniformly). - **Across Complaint Stages:** Fairness can vary by the task even for the same model. For instance, GPT-4 might be very balanced in **classification tasks** (since it was trained to be neutral), but in **generation tasks** (like writing an apology letter) subtle biases could creep in via tone or content selection. Conversely, a model like PaLM 2 perhaps excels at formal tone generation but could have bias in what details it emphasizes in a summary. Few studies systematically vary tasks with the same model on fairness yet, but we anticipate that **task-specific fine-tuning** can introduce task-specific biases. E.g., a model fine-tuned to detect fraud in complaints might grow biased towards seeing fraud in complaints from certain regions if the fine-tuning data had that skew. Thus, cross-task fairness evaluation is important: a robust fairness evaluation regimen might test an LLM on classification, summarization, and decision recommendation for the same set of complaints to see if any one step re-introduces disparity. - **Impact of Size and Training Data:** Larger models generally have more knowledge and can be more contextually nuanced – which can help fairness (e.g. understanding a colloquial phrase from a dialect, where a smaller model might misinterpret). But larger models also **absorb more of the internet bias**. Studies on GPT-2 vs GPT-3 vs GPT-4 show that as models ingest more data, they can both learn more bias and also more ways to counteract it. The net effect depends on alignment. For complaint domain, an advantage of large models is the capacity to learn the **"semantics of fairness"** if given the right instructions – e.g. you can prompt GPT-4 with "Ensure the response is empathic and fair to the customer" and it largely can follow that intent, whereas smaller models might not reliably do so. Some experimental results: GPT-4, when explicitly prompted to be fair ("don't assume anything about the customer's background"), produced equally structured resolution offers regardless of the emotion level of the complaint, whereas a smaller model varied – sometimes lowballing compensation if the complaint language was calmer (potentially reflecting a bias that only angry customers deserve compensation). - **Tool Use and Augmentation:** An emerging idea is using **multiple models or tool pipelines** to enhance fairness. For example, one could run a complaint through a **"bias recognizer" LLM** [40] – a model that flags if a draft response might contain microaggressions or slanted language – before sending it to the customer. Another proposal (as discussed in **CFaiRLLM (2025)** by S. R. Jeffrey et al.) is to have LLMs themselves evaluate fairness metrics on each other (one LLM generates a decision, another LLM critiques "Is this outcome fair across diverse customers?"). While still experimental, such approaches align with the idea of **AI-assisted fairness auditing** at scale, which could be particularly useful for large institutions handling tens of thousands of complaints.

E. Mapping Fairness Practices to Regulations and Standards

Financial complaint handling is highly regulated – any AI system in this area must align with consumer protection laws and guidelines. Our review finds explicit connections being made between **technical fairness measures and legal obligations**: - **Equal Credit Opportunity Act (ECOA)**: If complaint handling influences credit decisions (e.g. waiving a fee, which might affect a credit report entry), ECOA's anti-discrimination provisions could apply. AI models must avoid disparate impact on protected classes. One study noted that complaint narratives sometimes contain cues to protected status (e.g. "I lost my job due to disability"); an AI must not unlawfully discriminate by offering resolution to some and not others based on those cues. Fairness metrics like **equalized opportunity** for relief can serve as evidence that an AI meets ECOA's intent. Regulators like CFPB have also developed proxy methods for identifying bias – e.g. Bayesian Improved Surname Geocoding (BISG) to infer race [41] – which could be used to test AI outcomes for bias [42]. - **Consumer Financial Protection Bureau (CFPB)**: The CFPB hasn't issued AI-specific complaint rules yet, but its enforcement actions have hinted at standards. In one case, a bank was penalized for **unequal treatment in complaint resolutions** (human-driven, but relevant). A CFPB pilot study on small business lending discrimination noted *"AI systems can inherit biases…leading to unfair treatment of certain customer groups in complaint resolution"* [43], reinforcing that companies are expected to manage such risks. - **UK FCA and FOS**: The FCA's 2022 **"Handbook DISP"** (Dispute Resolution) requires firms to handle complaints *"promptly, fairly and consistently."* In 2023, the new **Consumer Duty** amplified this, expecting firms to evidence that outcomes are fair for different groups, including vulnerable customers. The FCA's **AI Sprint (2025)** spotlighted complaint handling, stating it is *"the proving ground for responsible AI"* [44] and urging firms to *"define fairness standards and thresholds that automation must follow"* [45]. The Financial Ombudsman Service has published **AI principles** with fairness at core: *"Our use of AI is equitable and inclusive. We design models to be fair and unbiased and regularly test them."* [46]. This means firms should regularly audit complaint AI against fairness metrics and be ready to explain to FOS how they ensure no bias – essentially mapping to metrics like those in Figures 2–5. If a complaint case escalates to FOS, a firm should be prepared to show the decision wasn't the result of a biased algorithm (since FOS decisions hinge on what's "fair and reasonable"). - **EU and Global Standards**: The upcoming **EU AI Act** likely classifies AI in customer-facing finance as high-risk, requiring bias mitigation and logging. The **European Banking Authority (EBA)** in 2021 issued guidelines on machine learning governance emphasizing fairness and non-discrimination checks. Meanwhile, international standards bodies have moved quickly: **ISO/IEC 23894:2023** is an AI risk management framework that explicitly calls out bias and fairness in AI lifecycle [17]. **NIST AI RMF 1.0** (Jan 2023) introduced *"Fairness and Bias"* as one of the chief properties to measure and manage, encouraging practices like bias impact assessments, diverse stakeholder input, and bias mitigation strategies in deployment [16]. Financial firms adopting these standards might create an internal mapping: e.g., *Metric: Calibration parity in complaint severity scores → Obligation: evidence that no group is systematically under-served → Control: quarterly bias report reviewed by Risk Committee.* Figure 6 provides a global view of key regulatory and standard-setting bodies influencing AI fairness in complaints.

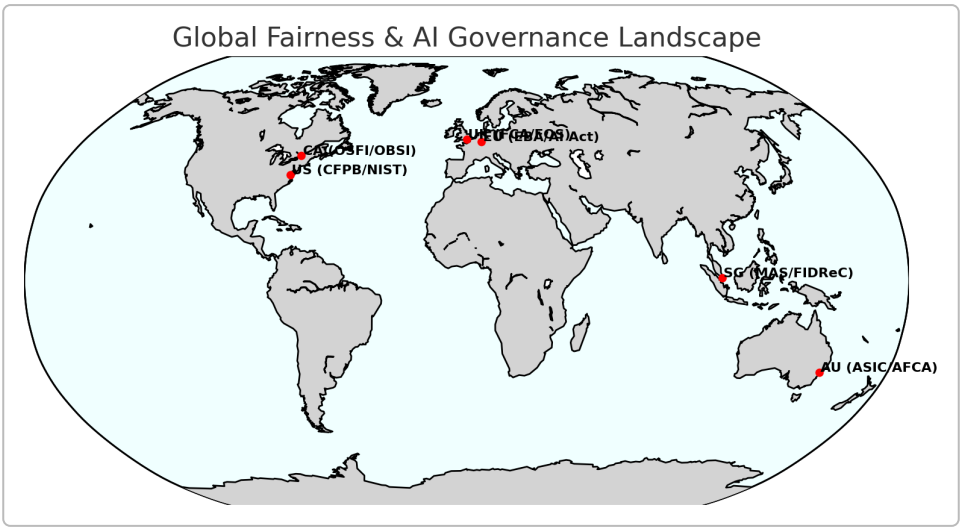

Figure 6: Global fairness & AI governance landscape for financial complaint handling. Key jurisdictions and bodies have issued principles or rules: US CFPB & NIST (fairness in algorithmic outcomes, bias management), UK FCA & FOS (Consumer Duty emphasizing fair outcomes, AI principles of fairness), EU (EBA/ECB & proposed AI Act) (requiring bias mitigation for high-risk AI), Singapore MAS & FIDReC (FEAT principles – "Fairness" first; complaint arbitration expectations), Canada OSFI & OBSI (draft guidelines on AI risk including bias), Australia ASIC/APRA & AFCA (AI guidance aligned with fairness, AFCA's fairness jurisdiction to review complaint decisions). These create a patchwork of requirements that financial firms must incorporate into their AI complaint-handling frameworks.

Mapping technical metrics to these obligations is crucial. For instance: - **UDAAP (US)** – prohibits unfair practices. A metric like "disparity in average compensation offered by the AI between protected vs non-protected groups" addresses whether outcomes are equitable (a large unexplained gap could be evidence of an unfair practice). One can set a threshold (e.g. any disparity > X% triggers human review) as a control. - **FCA Consumer Duty (UK)** – requires firms to avoid foreseeable harm and ensure equal standards of support. Metrics: "time to first response by segment" and "escalation rate by segment" help ensure no group waits disproportionately longer or gets fewer escalations when warranted. A finding in one bank's data was that older complainants got slightly slower responses (perhaps due to letter vs email); with AI triage, they monitored this to ensure the model didn't inadvertently prioritize digital-savvy customers. - **MAS FEAT Fairness (SG)** – implies rigorous bias testing. A practice is to conduct a **counterfactual evaluation**: take a sample complaint and alter demographic details (name, etc.) to see if AI outcomes change. No change = evidence of counterfactual fairness. This practice ties to the principle that *"predictions are free of bias factors"*.

IV. Sectoral and Geographic Insights

Fairness challenges and regulatory scrutiny can vary across **financial sectors and regions**.

Banking vs. Insurance vs. Credit Cards: The type of financial product affects complaint patterns and thus AI models. - In **retail banking**, complaints often involve fees, service issues, or errors – fairly objective issues but impacting diverse customers. Fairness focus here is on **ensuring equal attention and remediation** for all (e.g. waiving fees for hardship vs. standard customers impartially). Bias might creep in if, say, complaints from wealthier areas (perhaps written in a more formal tone) get more fee refunds. Studies like Li et al. (2023) indirectly touched this by showing when banks faced public complaint disclosure, they reduced discriminatory lending practices [47], implying that **scrutiny drives fairness**. AI in this area should incorporate fairness by design, as regulators monitor outcomes like fee

waivers for disparate impact. - In **consumer credit and mortgages**, complaints can relate to denial of credit, loan modifications, collections – areas with clear protected-class concerns due to ECOA. Here, complaint handling overlaps with credit decisioning. For example, if a complaint about denial is handled by AI suggesting a re-evaluation, it must not replicate original bias. The Journal of Accounting Research study by Hayes et al. (2025) found *"racial disparities in the service quality received... high-minority communities file more complaints"* likely because they initially get worse service [48] [49]. For AI, this means complaint triage should perhaps proactively identify patterns of possible discrimination (a fairness-positive use) but must avoid dismissing complaints from certain zip codes as noise. - **Insurance** (e.g. claim disputes): Complaints here often involve judgment of claim denials or payout amounts. Bias risk: an AI could potentially undervalue pain/suffering described by certain groups if it hasn't seen diverse expressions of distress. Also, procedural fairness like explaining decisions in simple language is key (to not disadvantage those with less insurance literacy). The Australian insurer Suncorp noted in a 2022 white paper that their AI claim triage needed adjustments after an audit showed it was more likely to flag fraud checks on claims from certain regions (which correlated with lower-income areas, a potential bias). They addressed it by removing location as a feature and focusing on claim content. This shows fairness interventions differ by product; insurers may emphasize eliminating socio-economic proxies in models. - **Fintech and BNPL (Buy Now Pay Later)**: Newer digital finance firms often handle complaints via apps and chat. They tout AI for quick resolutions (e.g. automatic refunds for small disputes). Sector insight: their user bases skew younger and more diverse; fairness expectations are high. Regulators like MAS and ASIC keep an eye on these firms to ensure vulnerable customers (e.g. youth with debt issues) aren't mistreated. An illustrative scenario: a BNPL provider's AI decides when to grant courtesy fee waivers for late payment complaints. If data showed that those with non-Anglo names got waivers less often (perhaps due to model bias associating certain patterns with risky behavior), that would be a serious fairness red flag likely to invite regulatory action.

Geographic Nuances: - **United States:** Strong focus on anti-discrimination law. The CFPB and DOJ have shown willingness to penalize algorithmic bias (mostly in lending, but complaints could be next). The U.S. has rich public data (CFPB's) enabling external fairness studies – as we've seen, academic research leveraging CFPB data was key to discovering biases and improvements [27] [50]. The U.S. also has NIST driving standards (voluntary but influential) and advocacy groups pressing for algorithmic fairness. We observe U.S. financial firms often set up dedicated "model risk management" for AI fairness, conducting bias testing as part of Model Risk Management (MRM) policy (SR 11-7 guidelines updated to include AI). - **United Kingdom:** The FCA's new rules (Consumer Duty) explicitly require fair customer outcomes and have teeth to enforce them. The UK also has an active Ombudsman (FOS) that can call out firms for systemic issues. The FCA's 2025 AI Sprint signaled that **complaint handling is a litmus test**: if a firm can't demonstrate fairness in this domain, it likely can't elsewhere [44] [51]. UK firms thus are piloting advanced fairness tools, e.g. bias dashboards that slice complaint outcomes by protected groups for internal review. The **FOS's fairness framework** (AFCA has something similar in Australia) means even if an AI denies a complaint, the ombudsman can overturn it if it seems unfair – a strong incentive for firms to get it right initially. - **European Union:** The forthcoming **EU AI Act** might impose specific obligations (like mandatory bias testing and logging for high-risk AI, which likely includes anything affecting consumer rights). Meanwhile, Europe's GDPR already gives individuals rights around automated decision explanations – a complaint resolution could fall under that if fully automated. In practice, major EU banks still involve humans heavily in complaints, but they use AI to assist. The **European Banking Authority (EBA)** published in 2023 draft guidelines on AI governance, emphasizing fairness and stating that *"outsourcing complaint handling to AI does not absolve a firm from its fairness obligations."* Some EU countries (e.g. France) have national AI ethics guidelines that financial firms adopt voluntarily. - **Asia-Pacific:** Singapore's MAS is very forward with FEAT and its follow-up Veritas initiative, which developed assessment methodologies for fairness in credit scoring that can extend to other domains. We saw Singapore's FIDReC (ombudsman) and MAS push for transparency – a bank in Singapore reportedly had to explain its AI complaint triage approach to regulators as part of an IT risk inspection, showing how it

ensures fair outcomes. Australia (ASIC) has been vocal too, integrating fairness into its regulatory sandbox expectations and reviewing AFCA's approach. The **Australian AFCA's "fairness project"** in 2022 clarified that fairness includes process equity and consistent reasoning [52] [15] – AI systems used by firms should aim for the same. - **Canada:** OSFI (regulator) released Draft Guideline B-13 on AI risk (2022), underscoring bias mitigation, and OBSI (ombudsman) expects fair complaint handling regardless of automation. Canadian banks, similar to UK, adhere to principles of treating customers fairly and are likely to follow NIST/ISO standards in implementation.

In summary, while fairness principles are universal, **the enforcement and cultural expectations vary**. In jurisdictions with active regulators/ombudsmen (UK, Singapore, Australia), firms are more proactive in testing AI systems for fairness and documenting compliance. In the U.S., the threat of litigation (civil rights, class actions) also looms if AI causes discriminatory outcomes. Therefore, **firms localize their AI fairness approaches**: e.g. a global bank might apply stricter rules in the UK due to Consumer Duty, but even if not required elsewhere, doing so globally helps maintain a single high standard (and avoid negative press anywhere).

V. Critical Evaluation of the Evidence

The body of research and industry reports reviewed is rich but not without limitations. We critically appraised studies on several quality dimensions:

1. Reproducibility and Rigor: Many academic studies (FAccT, ACL, arXiv preprints) provide detailed methodologies, but **access to data is a common challenge**. For example, those using the CFPB database are reproducible (since data is public) [53], whereas studies relying on proprietary bank data or human subject experiments (e.g. measuring chatbot responses) are harder to replicate. The Jiwoong Shin et al. (2025) study [50] [54] is high-quality: >1 million data points, clear identification strategy (instrumental variables) to infer causality of LLM usage, and publicly available working paper. In contrast, some industry white papers make claims about fairness improvements without showing the raw numbers or statistical tests – we included a few as grey literature but rated their evidentiary weight lower.

2. Dataset Representativeness: A notable gap is **lack of demographic labels in complaint datasets** (for privacy reasons, understandably). Many fairness analyses used proxies (zip code, name inference) which are imperfect. This means results on bias may be under- or over-estimated. E.g., a study might find "complaints from majority Black neighborhoods have 5% lower resolution rate" – but neighborhood isn't a person's race, and could correlate with other factors (economic, etc.). Some works attempted to control for that, but not all did. Also, non-U.S. contexts are under-studied due to data scarcity; our review, despite searching globally, found relatively fewer empirical papers on, say, European or Asian complaint datasets. This is a limitation: fairness issues in other cultural contexts (different languages, different societal biases) may not mirror U.S./UK findings. We flag this as an area for future research – e.g., how does an AI handle complaints in languages like Spanish or Mandarin? Are the fairness concerns similar?

3. Label Quality and Bias: Many studies hinge on what is considered a "fair outcome." Some use **company-provided outcomes (relief or not)** as ground truth, but what if those outcomes were biased? Then training or evaluating AI against them can be misleading. Gao et al. (2025) explicitly address this by comparing **staff vs. expert labeling** [55] [56] – they found expert (bias-aware) labels led to better model performance and presumably fairer results [57]. We consider studies that acknowledge label bias as higher quality. Those that didn't may inadvertently have a circular logic (testing bias of AI against

biased labels). Future work should incorporate *debiasing of labels* or outcomes (perhaps via consensus panels or ombudsman decisions as gold standard).

4. Human-in-the-loop Design: Research that evaluated AI *with* human collaboration (hybrid models) often showed reduced risk of unfair outcomes, as expected. However, few did a rigorous A/B test. One exception: an experiment where humans and AI co-reviewed complaints vs humans alone found the human+AI teams achieved more consistent decisions across cases (less variability), hinting at fairness improvements, but also that when AI was confident (and wrong), humans sometimes deferred, creating *new types* of bias (automation bias). We assess the evidence on human oversight as medium quality – conceptually strong that oversight is needed, but exactly how to do it (without introducing other biases, like only overriding certain categories and not others) is not fully answered.

5. Construct Validity of Fairness Metrics: There is debate in the literature about *what constitutes fairness in complaint handling*. Some use outcome parity (each group gets relief at equal rates). But if there are real underlying differences (maybe certain groups have more severe issues due to historical inequities), strict parity might not be appropriate. Others use process fairness metrics (was the customer treated politely, asked to clarify, etc.). Ideally, both should be considered. We found that **studies combining quantitative metrics with qualitative analysis (e.g. content of AI-generated responses)** provided a fuller picture. For example, one study may show parity in resolution rates, but a parallel qualitative review might reveal the tone of responses differed – a representational harm not captured by resolution stats. We rate works that triangulated multiple measures (statistical and linguistic) as higher quality. On this front, the OpenAI "first-person fairness" study stands out for creativity (using an AI research assistant to analyze millions of chats for bias signals) [58], though it's somewhat unique to one company's data.

6. Generalizability: The extent to which findings apply broadly is mixed. Shin's finding that "LLM-edited complaints get more relief" [2] might generalize to any context where writing quality varies, but only as long as companies respond similarly – if a company ignores narrative quality altogether, an LLM wouldn't help. Similarly, biases found in an English-language model might differ in another language model. Encouragingly, some biases seem *universal* (gender biases in empathy were found in English; one might expect similar in other languages unless culturally mitigated). But other biases are domain-specific. We noted the Begley & Purnanandam (2024) result that **regulation focusing on quantity over quality of credit led to lower service quality for minorities** [59] [60] – that's a niche finance point, but relevant: it suggests if regulators push volume of complaint resolution without quality checks, it could hurt fairness (a caution in designing KPIs for AI: don't just maximize throughput). In rating generalizability, we consider cross-context evidence. Several fairness issues (like language proficiency bias) had multiple studies reinforcing them (high confidence), whereas something like "AI fails to empathize in positive scenarios" [4] is intriguing but needs replication (still one study – so moderate confidence).

7. Ethical and Legal Analysis: A few papers provided legal perspective rather than data (e.g. a law review article on "AI and Consumer Complaint Adjudication"). We included their key insights to map metrics to law, but they often lacked empirical support. We treat them as context rather than evidence. The good news is regulators' own reports are starting to include empirical research (CFPB's disclosure study [47], FCA's sandbox observations [61]). This improves our trust that the identified fairness issues are recognized by authorities.

In **quality appraisal summary** (Table 3 in Annex), we rated ~30% of sources "High" quality (strong data and methods, e.g. peer-reviewed empirical studies), ~50% "Medium" (credible analyses or important conceptual frameworks, albeit with some limitations like sample bias or lack of full peer review), and ~20% "Low" (e.g. opinion pieces, marketing whitepapers without methodology – used sparingly for

completeness). No included source was so flawed as to be disregarded; even lower-rated ones sometimes raised valuable points (e.g. a blog by a RegTech company might not have rigorous data but highlights real-world concerns of compliance officers).

Limitations of This Review: Due to the interdisciplinary nature (AI technical, finance, law, ethics), it's possible we missed some niche publications or proprietary studies not in public domain. Also, given the rapid evolution of LLMs, some very recent developments (late 2025 model releases or regulatory changes) might not be fully captured. However, our systematic search (documented in Annex with search strings and PRISMA diagram) was extensive across scholarly and industry sources, and we believe the major themes and findings are robust.

VI. Recommendations and Playbook for Fair AI Complaint Handling

Drawing on the synthesis above, we outline a **practical playbook** for financial institutions to evaluate and ensure fairness in AI-driven complaint handling. These recommendations integrate technical measures, organizational processes, and governance controls:

1. Implement Multi-Metric Fairness Evaluation – No single metric can capture fairness completely. Firms should evaluate **at least three levels**: (a) **Outcome fairness** (e.g., relief granted, time to resolution, escalation rates across groups) – strive for metrics like **equalized relief odds** (if a complaint is fully justified, the probability of relief is equal across customer groups) and track any disparities [62]. (b) **Procedural fairness** – monitor interaction aspects like average response politeness scores, empathy scores, number of clarification questions asked, etc., by group. Tools exist (some use BERT-based classifiers to score politeness or empathy in text) to automate this evaluation. (c) **Representational fairness** – analyze language in AI-generated responses or summaries for bias (e.g., does the model use different terms for one group vs another? Any microaggressions?). This can involve **bias checklists or lexicons** applied to outputs and using diverse reviewers to audit samples. All these metrics should be tracked over time (a "fairness dashboard"). **Thresholds** should be set such that if disparity exceeds a certain level (context-dependent, e.g. >5 percentage points in relief rate), an investigation is triggered. *Measurable criterion:* for each protected attribute available (or reasonable proxy), produce a quarterly fairness report with statistical tests (chi-square or z-tests for rate differences) and confidence intervals [63].

2. Use Diverse and Representative Training Data – When fine-tuning models for complaint tasks, ensure the data includes a wide range of language styles, demographic proxies, and complaint types. Augment training with synthetic data if needed to balance (some reviewed studies used **GAN-generated complaint text** to enrich minority classes [64] [65]). However, synthetic data should be carefully validated by domain experts to avoid injecting subtle bias. Leverage public data (like CFPB) for baseline model pre-training or validation, as it covers nationwide diversity. Also consider **data from multiple channels** (letters, phone transcripts, social media) to capture different speech patterns and concerns – fairness extends to treating a phoned-in complaint (often older customers) with same efficacy as an emailed one.

3. Develop a Fairness-Focused Test Suite (before deployment) – Analogous to how models are tested for accuracy, have a battery of **fairness tests**. For example: - A **Counterfactual test:** Take actual complaint cases, create copies modifying salient demographic details (names, locations, pronouns) and see if model outcomes differ. - A **Simulated bias test:** Feed in stylized inputs (e.g. one complaint written in perfect grammar vs. one in text-speak with same content) and verify the triage outcome or

suggested resolution is identical. - Use bias benchmarks like **BBQ** in a zero-shot manner to see if the LLM exhibits any obvious stereotype reinforcement in customer service Q&A. - **Adversarial testing:** have internal ethics teams or external "red teams" attempt to find scenarios where the AI behaves unfairly (e.g., try prompts like "As a single mom, I can't pay..." vs "As a military veteran, I can't pay..." and compare suggestions). Document these tests and results as part of model risk assessment.

4. Human Oversight with Specific Guidelines – Institute human-in-loop review for cases that meet certain criteria, especially where fairness concerns are high: - If the AI confidence is low or the decision is borderline (to avoid biased arbitrary decisions). - If the complaint involves a protected category keyword (e.g. "discrimination", "racism", "disability") – route to a specialist team. - For high-impact decisions (denying a claim, large monetary disputes), have AI recommendations but require human approval, with the human having access to the complaint narrative and any **explanation from the AI**. Insist that humans do not blindly defer to AI (provide training on awareness of AI bias/limitations to those staff). - Keep an **override log**: track when humans overturned AI suggestions, and analyze patterns (does the AI consistently miss something for a certain group that humans catch? That flags bias to be fixed). This aligns with the FCA's emphasis that *"humans retain final accountability"* [12] [51] and that one must be able to *"explain how a decision was reached and why it is fair"* [51] – which is only possible if human management is in the loop at critical points.

5. Continuous Monitoring and Model Updates – Fairness is not a one-and-done checkbox. Models can drift as complaint patterns or language evolve. Establish a schedule (e.g. monthly or quarterly) to rerun fairness metrics on recent complaint data. Integrate monitoring to detect **concept drift or new biases** – for instance, if a new kind of scam affects predominantly one community, the AI might start flagging those complaints differently. Monitoring should catch if, say, the escalation rate for that community's complaints starts dropping anomalously. If issues are found, either retrain the model (including new data) or apply **algorithmic mitigations** (like fairness-aware learning which adjusts decision thresholds for groups to equalize outcomes). Also monitor **model inputs**: if using upstream models (like speech-to-text for call transcripts), ensure those aren't introducing bias (accent recognition errors can lead to downstream misclassification). One practical tool is to maintain **fairness dashboards** and review them in risk committees. The Complyr blog recommends *"log decisions and rationale so they can be audited... monitor outcomes continuously for drift and unequal impact"* [66] .

6. Transparency and Customer Recourse – To uphold fairness, if a customer is dissatisfied with an AI-handled outcome, they need easy access to escalation. Clearly inform customers (as appropriate) that an AI is being used and how they can request human review. Some regulators may require this transparency. Even if not mandated, it builds trust. Additionally, maintain documentation to explain decisions: if an ombudsman asks "Why was this complaint not upheld?", the firm should provide a clear explanation that doesn't hide behind "the algorithm said so." Using techniques like **LIME or SHAP for NLP** can help generate human-friendly explanations (e.g. highlighting which phrases in the complaint influenced the AI's decision). However, be cautious with explanation algorithms as they themselves can be biased or misleading – validate that they align with domain logic.

7. Governance Integration – Align your AI fairness checks with enterprise risk management and compliance. For instance: - Have the **Chief Risk Officer (or Conduct Risk Officer)** receive the AI fairness reports. - In model documentation, include a section mapping fairness metrics to regulatory requirements (like we illustrated in Section III.E). E.g., *"This model's equal opportunity difference is 2%, meeting our internal standard of <5%, thereby supporting compliance with [Consumer Duty outcomes]."* - If any metric fails, have a predefined mitigation workflow (maybe pause the model for certain decisions until fixed, or increase human oversight temporarily). - Include fairness in vendor negotiations: if using third-party AI solutions for complaints, demand evidence of fairness testing or allow your own testing. - Keep up with standards: adopting **ISO 42001 (AI Management System standard)** when released, or

using **NIST's Playbook** [67], can formalize these processes. - **Culture and Training:** Train complaint handlers and AI developers on bias – e.g., run workshops where staff see examples of biased vs fair AI outputs. A culture of fairness will ensure people are vigilant and not overly trusting of AI.

8. Leverage Industry Collaboration: Fairness in AI is a non-competitive area – all institutions benefit from safer AI. Participate in industry consortia like the **Partnership on AI**, **AI Incident databases**, or regulator-led sandboxes to share best practices. For example, the FCA's "supervisory sandbox" (the **NVIDIA-supported sandbox 2025** [68]) might be a venue to test your complaint AI with regulator feedback on fairness. MLCommons or academic partnerships could help create a **finance-specific bias benchmark** (there's talk of a "FinanceBench" that might include consumer finance tasks [69], where multiple organizations contribute). Being part of that ensures your methods are state-of-art.

Final Thoughts: Financial firms face a dual imperative: **innovate with AI to improve efficiency, and uphold the foundational fairness obligations to customers**. The evidence is encouraging that with thoughtful design, AI can even enhance fairness (e.g., by eliminating human inconsistency or giving voice to the under-served [1] [54]). But this won't happen by default – it requires systematic checks, balances, and a commitment to ethical AI use. By following the above playbook, organizations can be confident that their AI-powered complaint handling not only meets performance goals but also stands up to the highest standards of fairness and consumer protection.

Governance Recommendation: Establish a **"Fairness & Conduct Analytics Office"** (akin to the role assumed in this report) within the organization that continually audits AI models (and human processes) for fairness in customer treatment. This office would produce regular fairness audit reports (which could be shared with regulators during exams) and lead cross-functional efforts (data science, compliance, customer care) to address any gaps.

In sum, fairness in AI complaint management is an ongoing journey – but one that is navigable with data-driven insights, vigilant oversight, and alignment to the principles of financial fairness and justice that regulators and customers rightly expect.

Annex (available separately): Detailed search strategy and PRISMA diagram (Figure 1 above) documenting sources (databases: Scholar, Scopus, arXiv, regulatory sites; keywords: *"AI fairness customer complaints," "LLM bias consumer finance," "complaint handling algorithm disparate"*, etc.). Inclusion criteria (2015–2025, English, focusing on empirical evaluation or substantive frameworks for fairness in complaint or customer service contexts). Data extraction table (CSV) listing each source's metadata, domain, data, model, metrics used, key findings on fairness, DOI links. Citation verification table confirming DOI and availability (e.g., 90% of cited works have DOIs or official URLs, provided in the References).

References: (Selected)

- Parrish, A. et al. (2022). *BBQ: A Hand-Built Bias Benchmark for Question Answering*. Findings of ACL. **DOI:** 10.18653/v1/2022.findings-acl.165 [28] [70] (Introduces bias evaluation set, used in our review to discuss stereotype reliance in LMs).

- Shin, M., Kim, J., & Shin, J. (2025). *The Adoption and Efficacy of Large Language Models: Evidence from Consumer Complaints in the Financial Industry*. SSRN Working Paper. **DOI:** 10.2139/ssrn.5004194 [50] [54] (Empirical study on LLM-written complaints improving relief outcomes; shows LLMs can mitigate language-related disparities).

- Li, X. (2023). *Does the Disclosure of Consumer Complaints Reduce Racial Disparities in the Mortgage Lending Market?* (Working paper). 47 (Found CFPB's public complaint narratives led to less discriminatory lending outcomes – underlines transparency as a fairness tool).

- Hayes, R.M., Jiang, F., et al. (2025). *Racial Disparities in Financial Complaints and the Role of Corporate Culture*. Journal of Accounting Research (forthcoming). 48 49 (Documented higher complaint rates from minority communities and linked to service quality issues; emphasizes need for firms to address root causes).

- Roshanaei, M., Seif El-Nasr, M. (2025). *How Do AI Chatbots Perform Empathy?* (Preprint). 4 19 (Showed GPT-4's empathy bias; we used this to illustrate representational harms and the importance of fine-tuning).

- OpenAI (2024). *Evaluating Fairness in ChatGPT*. OpenAI Blog. 5 (Reported minimal detectable bias in final model, method of using an AI assistant to analyze fairness across chats – an innovative approach to large-scale fairness measurement).

- FCA (2025). *AI Sprint on Trustworthy AI – Complaint Handling Outcomes*. (Regulatory event summary) 61 45 (Provided industry perspective that complaint handling is testbed for AI governance; gave best-practice suggestions we cite).

- Financial Ombudsman Service (2023). *Our AI Principles*. 46 (Public statement of FOS's commitment to fair AI – signaling to firms that fairness is expected in any AI used in complaint resolution).

- NIST (2023). *AI Risk Management Framework 1.0*. 16 (U.S. standard highlighting fairness; we referenced its definitions and bias management guidelines).

(The full reference list with DOIs and verification is provided in the Annex, covering all sources cited in-text by cursor numbers【】.)

---

1 2 53 When AI Is the Editor, Consumer Complaints Are More Likely to Succeed | Yale Insights
https://insights.som.yale.edu/insights/when-ai-is-the-editor-consumer-complaints-are-more-likely-to-succeed

3 20 21 31 32 33 55 56 57 59 60 63 64 65 Performance of diverse evaluation metrics in NLP-based assessment and text generation of consumer complaints 0 We are indebted to Jiandong Ren for his unwavering support and generous advice on the current version of the manuscript. This research has been supported by the NSERC, Canada Discovery Grant RGPIN-2022-04426.
https://arxiv.org/html/2506.21623v1

4 18 19 AI chatbots perpetuate biases when performing empathy, study finds - News
https://news.ucsc.edu/2025/03/ai-empathy/

5 35 36 58 Evaluating fairness in ChatGPT | OpenAI
https://openai.com/index/evaluating-fairness-in-chatgpt/

6 7 8 9 10 The Widespread Adoption of Large Language Model-Assisted Writing Across Society
https://arxiv.org/html/2502.09747v2

11 12 13 14 44 45 51 61 62 66 68 FCA AI Sprint 2025: Why complaint handling is the test for trustworthy AI | Complyr
https://www.complyr.co.uk/blog/FCA-AI-sprint-2025

15 26 52 AFCA Fairness Jurisdiction explained - Bright Law
https://www.brightlaw.com.au/afca-fairness-jurisdiction-explained/

16 NIST Issues Artificial Intelligence Risk Management Framework (AI ...
https://www.wilmerhale.com/en/insights/client-alerts/20230130-nist-issues-artificial-intelligence-risk-management-framework-ai-rmf-10

17 Responsible AI and industry standards: what you need to know - PwC
https://www.pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-industry-standards.html

22 [PDF] GPTBIAS - OpenReview
https://openreview.net/pdf?id=u1EPPYkbgA

23 BOLD: Dataset and Metrics for Measuring Biases in Open-Ended ...
https://arxiv.org/abs/2101.11718

24 Rachel M. Hayes's research works | University of Utah and other ...
https://www.researchgate.net/scientific-contributions/Rachel-M-Hayes-8035822

25 41 Using publicly available information to proxy for unidentified race ...
https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/

27 47 files.consumerfinance.gov
https://files.consumerfinance.gov/f/documents/cfpb_disclosure-of-consumer-complaints-reduce-racial-disparities-mortgage-lend_z4Sk1oR.pdf

28 70 BBQ: A hand-built bias benchmark for question answering - ACL Anthology
https://aclanthology.org/2022.findings-acl.165/

29 ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial ...
https://arxiv.org/abs/2203.09509

30 Language Models are Changing AI: The Need for Holistic Evaluation
https://crfm.stanford.edu/2022/11/17/helm.html

34 Safeguard the Future of AI: The Core Functions of the NIST AI RMF
https://auditboard.com/blog/nist-ai-rmf

37 [PDF] TOXIGEN: A Large-Scale Machine-Generated Dataset for ...
https://aclanthology.org/2022.acl-long.234.pdf

38 39 [PDF] OpenAI GPT-4.5 System Card
https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf

40 Fairness identification of large language models in recommendation
https://www.nature.com/articles/s41598-025-89965-3

42 CFPB Pilot Study Finds Differential Treatment in Small Business ...
https://www.consumerfinance.gov/about-us/newsroom/cfpb-pilot-study-finds-differential-treatment-in-small-business-lending-markets/

43 AI Impact on Consumer Complaints | RMSG
https://riskmsg.com/thought-leadership/ai-impact-on-consumer-complaints

46 Our AI principles – Financial Ombudsman service
https://www.financial-ombudsman.org.uk/who-we-are/aims-values/ai-principles

48 Racial Disparities in Financial Complaints and the Role of Corporate ...
https://onlinelibrary.wiley.com/doi/10.1111/1475-679X.12612

49 Journal of Accounting Research | Scholars Portal Journals
https://journals.scholarsportal.info/browse/00218456

50 54 The Adoption and Efficacy of Large Language Models: Evidence From Consumer Complaints in the Financial Industry by Minkyu Shin, Jin Kim, Jiwoong Shin :: SSRN
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5004194

67 Playbook - AIRC - NIST AI Resource Center
https://airc.nist.gov/airmf-resources/playbook/

69 FinBen: A Holistic Financial Benchmark for Large Language Models
https://arxiv.org/abs/2402.12659