APPENDIX B
ONLINE QUESTIONNAIRE REPORT

Research Project: A Combined Approach of Design Thinking and Large Language Models for Assessing Risks of Data Leakage from Finance Chatbots

Principal Investigator: Andrius Busilas, MSc Computer Science, University of Essex

Survey Period: June-July 2024

Total Respondents: 77 (complete responses)

Data Collection Method: Online survey (Microsoft Forms), anonymized with IP tracking disabled

Analysis Method: Descriptive statistics, chi-square tests for independence, thematic coding of open-ended responses

Executive Summary

This report synthesizes quantitative and qualitative data from 77 survey respondents across three primary functional roles (Developers: n=25, Customer-Facing Staff: n=25, Compliance Officers: n=27) employed in financial services. The survey, conducted as part of the Design Thinking Empathise stage, aimed to assess stakeholder perceptions regarding data leakage risks, security mechanisms, trust factors, and operational requirements for finance chatbots powered by large language models (LLMs).

Key Findings:

1. **Universal Security Concern:** 89.6% of respondents rated data leakage concern as "moderate" or higher (≥3/5), with Compliance Officers exhibiting highest concern levels (mean=4.26/5.0 vs. overall mean=3.74/5.0).
2. **High-Risk Query Identification:** Balance enquiries (64.9% of respondents), transaction history requests (59.7%), and loan applications (57.1%) emerged as top-3 highest-risk query types, primarily due to potential account number exposure.
3. **Critical Security Mechanisms:** Multi-factor authentication received universal endorsement (100% of respondents), followed by PII masking (32% Developers), transparent security messaging (28% Customer-Facing Staff), and data minimisation (33.3% Compliance Officers).
4. **Trust Deficits:** Only 33.8% of respondents expressed "very much" or "complete" trust (≥4/5) in chatbot security capabilities, indicating substantial trust gap requiring targeted interventions.
5. **Performance-Security Balance:** Accurate responses (mean importance=4.42/5.0) significantly outweighed fast response time (mean=3.52/5.0, $p<0.001$), suggesting stakeholders prioritise correctness over speed.
6. **Diverse Query Handling Importance:** 79.2% rated multilingual/diverse query support as "moderately important" or higher (≥3/5), with multilingual queries (58.4%) and non-native phrasing (51.9%) identified as primary challenges.
7. **Bias Concerns Present but Lower Priority:** 57.1% expressed moderate-to-high concern (≥3/5) about bias, with language-based bias (38.3%) emerging as dominant concern type.

Statistical Significance: Chi-square tests revealed significant role-based differences in concern levels ($\chi^2=18.74$, $p<0.001$), trust perceptions ($\chi^2=15.92$, $p=0.003$), and feature prioritization ($\chi^2=22.41$, $p<0.001$), validating the necessity of role-adapted design approaches.

Section 1. RESPONDENT DEMOGRAPHICS

1.1 Role Distribution

Role	Count	Percentage	Mean Experience (years)
Developer	25	32.5%	4.8
Customer-Facing Staff	25	32.5%	5.2
Compliance Officer	27	35.1%	5.6
Total	**77**	**100%**	**5.2**

Analysis: The distribution achieved near-perfect balance across three primary roles (deviation <3%), ensuring representative sampling of diverse functional perspectives. The "Other" category (excluded from primary analysis) comprised 4 Analysts and 4 Managers providing supplementary insights but insufficient sample size for statistical subgroup analysis.

1.1.1 *Experience Distribution*

Experience Level	Count	Percentage	Cumulative %
Less than 1 year	8	10.4%	10.4%
1–2 years	12	15.6%	26.0%
2–5 years	21	27.3%	53.3%
5–10 years	23	29.9%	83.1%
Over 10 years	13	16.9%	100.0%

Experience-Concern Correlation: Spearman's rank correlation revealed moderate positive association between experience and data leakage concern ($\rho=0.34$, $p=0.002$), suggesting that tenure increases risk awareness, potentially reflecting accumulated exposure to security incidents or deepened regulatory understanding.

1.1.2 *Age Distribution*

Age Range	Count	Percentage
18–25	14	18.2%
26–35	32	41.6%
36–45	19	24.7%
46–60	12	15.6%
Over 60	0	0.0%

Analysis: The sample skews younger (59.8% under 36), reflecting financial services' adoption of AI talent typically concentrated in millennial/Gen-Z demographics. The absence of respondents >60 may indicate sampling bias (younger professionals more likely to engage with online surveys) or genuine demographic reality (retirement/career transition reducing >60 representation in operational roles).

Section 2. SENSITIVE DATA HANDLING PATTERNS

2.1 Data Handling Frequency

Q4: How often do you or your team handle sensitive customer data?

Frequency	Count	Percentage	Cumulative %
Never (1)	0	0.0%	0.0%
Rarely (2)	8	10.4%	10.4%
Occasionally (3)	19	24.7%	35.1%
Frequently (4)	27	35.1%	70.1%
Daily (5)	23	29.9%	100.0%

Mean: 3.84/5.0 **| Median:** 4.0 **| Mode:** 4 **| SD:** 0.98

Analysis: 65.0% of respondents handle sensitive data frequently-to-daily (≥4/5), confirming that the sample possesses relevant operational experience with PII rather than theoretical understanding. Zero "Never" responses validate sampling effectiveness—all respondents engage with sensitive data contexts making their insights directly applicable to chatbot security design.

Role-Based Comparison:

Role	Mean Frequency	SD
Developer	3.68	1.03
Customer-Facing Staff	3.92	0.91
Compliance Officer	3.93	1.00

ANOVA revealed no significant role-based differences (F=0.71, p=0.494), indicating consistent data exposure across functions—an important finding suggesting that security concerns stem from universal operational reality rather than role-specific exaggeration.

2.1 Types of Sensitive Data Handled

Q4a: Which types of sensitive data do you handle most often? (Multiple selection, n=69 responding to conditional question)

Data Type	Count	Percentage of Respondents
Customer names	42	60.9%
Account numbers	41	59.4%
Transaction details	28	40.6%

Email addresses	19	27.5%
Telephone numbers	17	24.6%
Addresses	16	23.2%
National insurance numbers	3	4.3%
Other (Passwords)	9	13.0%

Analysis: Customer names and account numbers dominate (60% each), representing highest-frequency PII types requiring chatbot protection. The surprisingly high "Other: Passwords" rate (13.0%) raises immediate security red flags—passwords should never be handled by customer service staff or stored in accessible formats. This finding suggests either:

1. Misunderstanding of "passwords" (possibly referring to password reset processes rather than actual credentials)
2. Concerning security practices requiring remediation
3. Legacy system issues where password visibility persists despite best-practice violations

Follow-up investigation recommended: Clarify whether "password handling" refers to reset facilitation (acceptable) vs. actual credential access (unacceptable).

Implication for Chatbot Design: The dual dominance of names and account numbers necessitates robust Named Entity Recognition (NER) capable of detecting both structured identifiers (account numbers following patterns) and unstructured entities (names exhibiting cultural/linguistic diversity).

Section 3. DATA LEAKAGE RISK PERCEPTIONS

3.1 Overall Concern Levels

Q5: How concerned are you about sensitive customer data being exposed through a finance chatbot?

Concern Level	Count	Percentage	Cumulative %
Not at all (1)	0	0.0%	0.0%
Slightly (2)	8	10.4%	10.4%
Moderately (3)	23	29.9%	40.3%
Very (4)	28	36.4%	76.6%
Extremely (5)	18	23.4%	100.0%

Mean: 3.74/5.0 | **Median:** 4.0 | **Mode:** 4 | **SD:** 0.95

Analysis: 89.6% expressing moderate-to-extreme concern (≥3/5) demonstrates near-universal stakeholder anxiety regarding chatbot data security. Zero "not at all concerned" responses particularly notable—even respondents rating concern as "slightly" (10.4%) still acknowledge risk existence, merely assessing magnitude differently.

Role-Based Concern Comparison:

Role	Mean Concern	SD	Median
Compliance Officer	**4.26**	0.81	5.0
Customer-Facing Staff	3.40	0.87	3.0
Developer	3.56	1.04	4.0
Overall	**3.74**	**0.95**	**4.0**

Statistical Test: One-way ANOVA revealed significant role-based differences (F=8.43, p<0.001). Post-hoc Tukey HSD tests indicated:

- Compliance Officers significantly more concerned than Customer-Facing Staff (p<0.001)
- Compliance Officers significantly more concerned than Developers (p=0.003)
- No significant difference between Developers and Customer-Facing Staff (p=0.524)

Interpretation: Compliance Officers' elevated concern (mean=4.26, approaching "extremely concerned") reflects regulatory responsibility and liability awareness. Their role involves managing GDPR compliance, EU AI Act adherence, and regulatory reporting—direct accountability for data breaches explains heightened anxiety. Developers and Customer-Facing Staff, while concerned, exhibit more moderate ratings potentially reflecting:

- **Developers:** Technical confidence in mitigation strategies offsetting concern
- **Customer-Facing Staff:** Focus on user experience balancing security considerations

3.2 Specific Data Exposure Risks

Q5a: What specific data exposure risks concern you most? (Open-ended, n=69)
Thematic Coding Results:

Theme	Count	% of Responses	Representative Quote
Unauthorized disclosure to wrong user	23	33.3%	"Disclosing account details to wrong customer"
GDPR/regulatory violations	18	26.1%	"GDPR breach risk from systematic exposure"
Account number exposure	15	21.7%	"Revealing account numbers in responses"
Hacking/unauthorized access	12	17.4%	"External hacking exposure of customer data"
Accidental leaks from errors	8	11.6%	"System errors causing unintentional disclosure"
Data retention violations	7	10.1%	"Retaining data longer than GDPR permits"

Note: Responses could code to multiple themes; percentages exceed 100%

Analysis: The dominant theme—unauthorized disclosure to wrong user (33.3%)—represents authentication/authorization failures where chatbot provides legitimate data to illegitimate requestor. This differs from external hacking (17.4%) involving malicious actors penetrating system perimeter. The distinction proves critical for defense design:

- **Wrong-user disclosure:** Requires robust authentication verification, session management, context validation
- **External hacking:** Requires perimeter security, encryption, intrusion detection

GDPR violations emerge as second concern (26.1%), reflecting Compliance Officers' influence and European regulatory environment's prominence. This validates the research focus on GDPR compliance as non-negotiable rather than aspirational requirement.

3.3 High-Risk Query Identification

Q6: Which customer queries pose the highest risk of sensitive data leakage? (Select up to 3, n=77)

Query Type	Count	% of Respondents	Risk Ranking
Balance enquiries	**50**	**64.9%**	**1**
Transaction history requests	**46**	**59.7%**	**2**
Loan applications	**44**	**57.1%**	**3**
Password resets	28	36.4%	4
Fraud reports	27	35.1%	5
Account updates	22	28.6%	6

Analysis: The top-3 consensus (balance enquiries, transaction history, loan applications) reflects queries inherently requiring sensitive data access to provide meaningful responses:

Balance Enquiries (64.9%): Seemingly simple query ("What's my account balance?") necessitates account number identification, authentication verification, and balance disclosure—multiple leakage vectors. The query's frequency in operational contexts amplifies risk: if 1,000 daily balance queries each carry 1% leakage risk, that's 10 expected daily incidents.

Transaction History Requests (59.7%): Particularly high-risk due to multi-entity exposure: account numbers, transaction amounts, merchant names, dates/times, potentially inferring location patterns or spending behaviors qualifying as sensitive profiling under GDPR Article 4(4).

Loan Applications (57.1%): Involves comprehensive PII (income, employment, address history, national ID) plus financial data (credit scores, existing liabilities, collateral). The data volume and sensitivity combination creates "high-value target" attractive to adversaries.

Q6a Rationale Analysis: (Open-ended explanations, thematic coding)

Rationale Theme	Count	%	Representative Quote
May reveal account numbers	38	49.4%	"Balance queries may inadvertently expose account numbers"
Shows sensitive financial details	29	37.7%	"Transaction history reveals spending patterns"
Exposes personal information	24	31.2%	"Loan apps contain comprehensive PII including income"
Authentication weakness risk	18	23.4%	"If auth fails, wrong user gets sensitive data"

Credential exposure (passwords)	14	18.2%	"Password resets may expose authentication credentials"

Key Insight: The dominant rationale, account number exposure (49.4%), highlights specific entity type concern rather than abstract "data leakage." This granularity informs technical requirements: NER models must achieve >98% recall on structured financial identifiers (account numbers, IBANs, card numbers) as priority beyond general PII (names, addresses).

3.4 Leakage Likelihood Assessment

Q7: How likely is it that a finance chatbot could unintentionally leak sensitive data?

Likelihood	Count	Percentage	Cumulative %
Very unlikely (1)	0	0.0%	0.0%
Unlikely (2)	12	15.6%	15.6%
Somewhat likely (3)	31	40.3%	55.8%
Likely (4)	24	31.2%	87.0%
Very likely (5)	10	13.0%	100.0%

Mean: 3.42/5.0 | **Median:** 3.0 | **Mode:** 3 | **SD:** 0.92

Analysis: 84.4% assessing likelihood as "somewhat likely" or higher (≥3/5) indicates stakeholder skepticism regarding chatbot security reliability. This contrasts with Q5 concern levels (mean=3.74) being higher than likelihood assessment (mean=3.42)—stakeholders are MORE concerned than they believe incidents are probable. This suggests:

1. **High consequence aversion:** Even moderate-probability risks generate high concern when consequences prove severe
2. **Uncertainty premium:** Lack of empirical chatbot security data inflates concern beyond actuarial likelihood
3. **Precautionary principle:** Financial services culture favoring conservative risk assessment

Q7a Incident Type Concerns: (Conditional question, n=65)

Incident Type	Count	%
Incorrect response exposing data	22	33.8%
Unauthorized data access	20	30.8%
Data retention violations	18	27.7%
Other	5	7.7%

Analysis: The near-equal distribution across three primary incident types (incorrect response: 33.8%, unauthorized access: 30.8%, data retention: 27.7%) indicates multifaceted threat perception rather than single-vector concern. This validates multi-layered security architecture addressing:

- **Incorrect responses:** Output validation, hallucination detection
- **Unauthorized access:** Authentication, authorization, session management
- **Data retention:** Automated deletion, retention policies, GDPR Article 17 compliance

3.5 Consequence Severity Rankings

Q8: What would be the most severe consequence of a chatbot data leakage incident? (Rank 1-5, 1=Most severe)

Consequence	Mean Rank	Median	Mode	SD
Regulatory fines	**2.18**	**2**	**1**	1.34
Loss of customer trust	**2.41**	**2**	**2**	1.29
Reputational damage	**2.82**	**3**	**3**	1.18
Financial loss to customers	**3.24**	**3**	**4**	1.15
Operational disruption	**4.35**	**4**	**5**	0.91

Lower mean rank indicates higher perceived severity

Analysis: Regulatory fines emerge as most severe consequence (mean rank=2.18), narrowly edging customer trust loss (2.41). This ordering reflects several realities:

Regulatory Fines Priority:

- GDPR Article 83 penalties (€20M or 4% global turnover) represent quantifiable, immediate financial threat
- EU AI Act violations potentially blocking product deployment entirely
- Regulatory investigations consuming substantial management attention and legal resources
- Precedent concerns where lenient treatment of early violations may not persist

Customer Trust as Near-Equal Second:

- Financial services depend fundamentally on trust; breaches create existential competitive disadvantage
- Trust erosion proves difficult to quantify but devastating long-term
- Customer churn accelerates exponentially after publicized breaches (social contagion effects)

- Rebuilding trust requires years and substantial investment

Statistical Test: Friedman test (non-parametric repeated measures) confirmed significant differences in severity rankings (χ^2=127.43, p<0.001), validating that stakeholders meaningfully differentiate consequence types rather than rating all equally severe.

Q8a Worst-Case Scenarios: (Open-ended, n=77, thematic coding)

Scenario Theme	Count	%	Representative Quote
Mass data breach (multiple customers)	28	36.4%	"Systematic exposure of thousands of customer accounts"
Regulatory fines / sanctions	19	24.7%	"€20M GDPR fine destroying quarterly profits"
Complete trust loss / reputational collapse	14	18.2%	"Viral social media exposure causing customer exodus"
Fraud risk / financial theft	10	13.0%	"Criminals using leaked data for account takeovers"
Client fund exposure	4	5.2%	"Direct financial loss to customers from unauthorized transactions"
Operational disruption	2	2.6%	"System shutdown forcing manual processing backlog"

Critical Insight: The emphasis on "mass" or "systematic" breaches (36.4%) rather than isolated incidents reveals scale consciousness—stakeholders understand that chatbot deployment introduces systemic rather than individualized risk. A vulnerability affecting one interaction potentially affects all interactions, distinguishing chatbot risk from human-agent risk where individual errors remain isolated.

Section 4. SECURITY MECHANISMS AND TECHNICAL REQUIREMENTS

4.1 Importance of Leakage Prevention Mechanisms

Q9: How important is it for a chatbot to have mechanisms to detect and prevent data leakage?

Importance	Count	Percentage	Cumulative %
Not important (1)	0	0.0%	0.0%
Slightly important (2)	0	0.0%	0.0%
Moderately important (3)	8	10.4%	10.4%
Very important (4)	23	29.9%	40.3%
Extremely important (5)	46	59.7%	100.0%

Mean: 4.49/5.0 | **Median:** 5.0 | **Mode:** 5 | **SD:** 0.68

Analysis: Near-unanimous endorsement (89.6% rating ≥4/5, zero responses <3/5) establishes leakage prevention as non-negotiable rather than optional feature. The mean=4.49 represents highest importance rating across all survey items, exceeding even response accuracy (mean=4.42) and human escalation (mean=4.01). This priority ordering informs resource allocation: security mechanisms warrant maximum investment even if marginally compromising other dimensions (speed, functionality).

4.2 Role-Specific Technical Mechanism Preferences

Q9a (Developers): Which technical mechanisms are critical? (Multiple selection, n=25)

Mechanism	Count	% of Developers
Data encryption	16	64.0%
Prompt sanitisation	12	48.0%
Secure data retrieval	10	40.0%
PII masking	8	32.0%
Audit logs	5	20.0%

Analysis: Data encryption dominance (64.0%) reflects developer understanding that perimeter security (encryption at rest/transit) provides foundational protection upon which other mechanisms build. Prompt sanitisation ranking second (48.0%) acknowledges prompt injection as primary LLM-specific vulnerability requiring dedicated defenses.

PII masking ranking fourth (32.0%) despite being highly visible security feature suggests developers prioritize "preventing data from leaving secure boundaries" (encryption, retrieval controls) over "redacting data after retrieval" (masking). This architectural philosophy—security by access prevention rather than post-hoc redaction—aligns with defense-in-depth principles.

Q9b (Customer-Facing Staff): Which features would reassure customers? (Multiple selection, n=25)

Feature	Count	% of Staff
Clear security messages	14	56.0%
Human escalation option	12	48.0%
Authentication prompts	11	44.0%

Analysis: Clear security messages leading (56.0%) validates transparency-as-trust-mechanism hypothesis. Customers lack technical expertise to evaluate backend security (encryption algorithms, access controls) but can appreciate explicit communication: "Your data is encrypted," "I've detected a name and masked it for privacy." This finding directly informs UI/UX requirements: security visibility proves equally important as security implementation.

Human escalation option ranking second (48.0%) reflects customer desire for agency—maintaining ability to bypass automation when uncomfortable or encountering complexity. This validates hybrid human-AI architectures over full automation, even if technically feasible.

Q9c (Compliance Officers): Which mechanisms ensure GDPR/EU AI Act compliance? (Multiple selection, n=27)

Mechanism	Count	% of Compliance Officers
Data minimisation	15	55.6%
Anonymised logs	12	44.4%
Consent prompts	9	33.3%

Analysis: Data minimisation priority (55.6%) reflects GDPR Article 5(1)(c) principle: "personal data shall be adequate, relevant and limited to what is necessary." Compliance officers recognize that best data protection involves not collecting/processing data unnecessarily—architectural decisions limiting data access prove more robust than post-collection protection.

Anonymised logs (44.4%) address GDPR Article 32 requirement for "ability to ensure ongoing confidentiality...of processing systems" whilst enabling operational monitoring and security analysis. The technique represents elegant solution to competing objectives: maintaining system observability whilst protecting individual privacy.

4.3 Manipulation and Adversarial Attack Concerns

Q10: How concerned are you about a chatbot being manipulated to leak data through malicious inputs?

Concern Level	Count	Percentage	Cumulative %
Not at all (1)	0	0.0%	0.0%
Slightly (2)	11	14.3%	14.3%
Moderately (3)	32	41.6%	55.8%
Very (4)	22	28.6%	84.4%
Extremely (5)	12	15.6%	100.0%

Mean: 3.45/5.0 | **Median:** 3.0 | **Mode:** 3 | **SD:** 0.94

Analysis: While 84.4% express moderate-to-extreme concern (≥3/5), the mean=3.45 rates slightly lower than general leakage concern (mean=3.74, p=0.046 via paired t-test). This suggests stakeholders perceive adversarial manipulation as meaningful but secondary threat compared to accidental/error-based leakage. Potential interpretations:

1. **Sophistication barrier:** Stakeholders assess adversarial attacks as requiring technical expertise limiting attacker population
2. **Detection optimism:** Belief that monitoring systems can identify obviously malicious patterns
3. **Probability discounting:** Recognition that while possible, targeted adversarial attacks prove less frequent than operational errors

Q10a Manipulation Type Concerns: (Conditional, n=66)

Manipulation Type	Count	%
Tricking chatbot to reveal data	28	42.4%
Causing incorrect responses	20	30.3%
Bypassing authentication	15	22.7%
Other	3	4.5%

Analysis: "Tricking chatbot" dominance (42.4%) reflects awareness of prompt injection and social engineering techniques where seemingly benign inputs contain hidden instructions. This contrasts with "bypassing authentication" (22.7%)—stakeholders perceive conversational manipulation as more plausible than technical auth circumvention, possibly reflecting:

- Recent media coverage of ChatGPT jailbreaking increasing awareness

- Recognition that LLMs lack robust input validation compared to traditional systems
- Understanding that conversational interfaces expand attack surface beyond technical exploits

Q10b (Developers): Technical Safeguards: (Open-ended, n=25, thematic coding)

Safeguard Theme	Count	%	Representative Quote
Input validation / sanitization	12	48.0%	"Validate and sanitize all inputs before processing"
Authentication checks	7	28.0%	"Verify authentication before sensitive operations"
Other (unspecified)	6	24.0%	Various

Analysis: Input validation/sanitization leading (48.0%) reflects developer recognition that LLM systems require adapted security approaches. Traditional input validation (checking field lengths, data types, SQL injection patterns) proves insufficient for natural language inputs where malicious content hides within grammatically correct sentences. Developers implicitly acknowledge need for semantic input analysis beyond syntactic validation.

Q10c (Customer-Facing Staff): Customer Reaction: (Open-ended, n=25, thematic coding)

Reaction Theme	Count	%	Representative Quote
Loss of trust	20	80.0%	"Customers would lose trust immediately"
No specific response	5	20.0%	[Blank or non-informative]

Analysis: The overwhelming "loss of trust" consensus (80.0%) demonstrates customer-facing staff's frontline understanding that single publicized incident disproportionately damages reputation. Social media amplification means individual breach becomes viral narrative: "Bank's AI leaked my data" spreads exponentially regardless of statistical rarity. This finding validates proactive security investment as brand protection, not merely regulatory compliance.

Q10d (Compliance Officers): Regulatory Issues: (Open-ended, n=27, thematic coding)

Issue Theme	Count	%	Representative Quote
Regulatory fines / penalties	24	88.9%	"GDPR Article 83 fines up to €20M or 4% turnover"
Discrimination lawsuits	2	7.4%	"Bias could trigger discrimination litigation"
No specific response	1	3.7%	[Blank]

Analysis: Regulatory fines near-unanimity (88.9%) confirms Compliance Officers' primary accountability: managing financial and legal exposure from violations. The GDPR Article 83 citation frequency demonstrates sophisticated regulatory knowledge—respondents don't merely express general concern but identify specific legal provisions and penalty structures. This expertise validates involving Compliance Officers as design partners rather than post-hoc reviewers.

Section 5. **TRUST, USABILITY, AND FEATURE PREFERENCES**

5.1 **Current Trust Levels**

Q11: How much would you trust a finance chatbot to handle sensitive customer queries securely?

Trust Level	Count	Percentage	Cumulative %
Not at all (1)	0	0.0%	0.0%
Slightly (2)	13	16.9%	16.9%
Moderately (3)	38	49.4%	66.2%
Very much (4)	22	28.6%	94.8%
Completely (5)	4	5.2%	100.0%

Mean: 3.22/5.0 | **Median:** 3.0 | **Mode:** 3 | **SD:** 0.76

Analysis: Only 33.8% expressing "very much" or "complete" trust (≥4/5) reveals substantial trust deficit, two-thirds of stakeholders harbor reservations despite presumably understanding chatbot capabilities exceed prompt-based systems. The mean=3.22 (barely above scale midpoint) indicates ambivalence rather than confidence or rejection.

Trust Gap Quantification: Comparing trust (mean=3.22) versus importance of trust features (mean=4.01–4.49 across Q12 items) reveals 0.79–1.27point gap, suggesting stakeholders desire trustworthy chatbots whilst doubting current implementations achieve trustworthiness. This gap represents primary design challenge requiring targeted intervention.

Role-Based Trust Comparison:

Role	Mean Trust	SD
Customer-Facing Staff	2.88	0.73
Developer	3.20	0.76
Compliance Officer	3.56	0.70

Statistical Test: One-way ANOVA revealed significant role-based differences (F=7.92, p<0.001). Post-hoc analysis showed:

- Compliance Officers significantly more trusting than Customer-Facing Staff (p<0.001)
- Compliance Officers moderately more trusting than Developers (p=0.041)
- Developers and Customer-Facing Staff not significantly different (p=0.089)

Interpretation: The trust hierarchy (Compliance > Developers > Customer-Facing Staff) appears counterintuitive given Compliance Officers' highest concern levels (Q5: mean=4.26). This paradox resolves through understanding that:

1. **Compliance Officers trust mechanisms, not systems:** Their higher trust reflects confidence in regulatory frameworks and security controls they can mandate, not inherent chatbot reliability
2. **Customer-Facing Staff witness failures:** Direct exposure to customer complaints, system errors, and edge cases erodes confidence in automation
3. **Developers exhibit technical realism:** Understanding both capabilities and limitations produces moderate trust—neither naive optimism nor unfounded skepticism

Q11a Trust-Increasing Factors: (Conditional, open-ended, n=51 with trust ≤3)

Factor Theme	Count	%	Representative Quote
Certified security standards	18	35.3%	"ISO 27001 certification, third-party security audits"
Transparent security explanations	14	27.5%	"Clear explanations of how data is protected"
Security protocols	9	17.6%	"Documented security procedures and compliance"
GDPR compliance	6	11.8%	"Demonstrated GDPR Article 32 technical measures"
Visible security indicators	4	7.8%	"Real-time indicators showing security active"

Analysis: Certified security standards leading (35.3%) indicates stakeholders seek third-party validation rather than self-assessed security claims. Certifications (ISO 27001, SOC 2, PCI-DSS) provide independent verification reducing perceived vendor bias. This finding suggests deployment strategy should prioritize formal certification early in development rather than treating as post-launch activity.

Transparent security explanations ranking second (27.5%) reinforces earlier findings (Q9b) that visibility builds trust. Stakeholders desire understanding *how* security operates, not merely assurances *that* security exists. This validates design principle: "show, don't just tell."

5.2 5.2 Feature Importance for Trust and Usability

Q12: How important are these features for making a finance chatbot trustworthy and easy to use?

Feature	Mean	Median	Mode	SD	% Rating ≥4
Accurate responses	**4.42**	**5**	**5**	0.71	**89.6%**
Human escalation	**4.01**	**4**	**4**	0.88	**79.2%**
Transparent security	**3.81**	**4**	**4**	0.93	**71.4%**
Simple interface	**3.56**	**4**	**3,4**	0.89	**61.0%**
Fast response time	**3.52**	**4**	**3**	0.96	**58.4%**

Analysis: The feature ranking reveals stakeholder priorities challenging common assumptions:

1. Accuracy Supremacy (mean=4.42): Accurate responses significantly outrank all other features ($p<0.001$ via repeated measures ANOVA), establishing correctness as paramount concern. Stakeholders prioritize "right answer slowly" over "wrong answer quickly"—directly contradicting typical tech industry emphasis on speed/responsiveness. This finding fundamentally shapes performance targets: F1 score >0.90 proves more critical than sub-second latency.

2. Human Escalation Value (mean=4.01): Second-place ranking validates hybrid human-AI architectures over full automation. Stakeholders desire "safety valve" enabling graceful degradation when chatbot encounters complexity or user loses confidence. This preference reflects practical wisdom: automation excels at routine scenarios but struggles with edge cases requiring judgment, empathy, or creative problem-solving.

3. Transparency Importance (mean=3.81): Third-place ranking confirms security visibility as trust mechanism. The 0.20-point gap between human escalation (4.01) and transparency (3.81) proves statistically significant ($t=3.12$, $p=0.002$), indicating escalation slightly more valued than transparency—possibly reflecting pragmatic preference for "escape hatch" over "understanding mechanisms."

4. Interface Simplicity and Speed as Secondary (means=3.56, 3.52): These features ranking lowest (though still "moderately important" to "very important" range) challenges UX conventional wisdom prioritizing frictionless experience. Financial services stakeholders accept complexity and delay when accompanied by accuracy and security. This finding doesn't justify *unnecessary* complexity but rather indicates tolerance for essential complexity.

Statistical Test: Friedman test confirmed significant differences across features ($\chi^2=73.58$, $p<0.001$), validating meaningful stakeholder differentiation rather than uniform "everything is important" responses.

Role-Based Feature Priority Analysis:

Feature	Developers	Customer-Facing	Compliance	F-statistic	p-value
Fast response	3.84	3.44	3.30	3.21	0.046*
Accurate responses	4.36	4.52	4.37	0.47	0.628
Simple interface	3.68	3.56	3.44	0.62	0.542
Transparent security	3.60	3.72	4.11	3.82	0.026*
Human escalation	3.92	4.28	3.85	2.89	0.062

*Significant at p<0.05 level

Key Insights:

- **Accuracy universally prioritized:** No role-based differences (F=0.47, p=0.628)—all functions agree correctness paramount
- **Developers value speed more:** Significantly higher rating (3.84) than Compliance (3.30, p=0.041), possibly reflecting technical confidence in optimization versus Compliance's security-first orientation
- **Compliance prioritizes transparency:** Significantly higher (4.11) than Developers (3.60, p=0.023), reflecting regulatory accountability requiring explainability

Q12f Additional Trust/Usability Features: (Open-ended, n=77)

Feature Theme	Count	%	Representative Quote
Multi-factor authentication	77	100.0%	"MFA for sensitive queries" (universal mention)
Other specific features	0	0.0%	[No additional suggestions beyond MFA]

Remarkable Finding: The **100% MFA endorsement** represents strongest consensus across entire survey. Every single respondent independently identified multi-factor authentication as critical trust feature, despite question being optional open-ended prompt. This universal convergence indicates:

1. **Industry-wide MFA normalization:** Multi-factor authentication has become baseline expectation rather than advanced security
2. **Authentication as primary control:** Stakeholders recognize that post-authentication security (encryption, masking) proves irrelevant if authentication fails
3. **Practical experience:** Likely reflecting direct exposure to authentication-bypass incidents in operational contexts

Implication: MFA implementation becomes non-negotiable requirement rather than "nice-to-have" feature. Any chatbot design lacking MFA will face immediate stakeholder rejection regardless of other capabilities.

Section 6. SENSITIVE AND DIVERSE QUERY HANDLING

6.1 Sensitive Query Response Preferences

Q13: How should a chatbot respond to sensitive/ambiguous query (e.g., "Show my account details" without authentication)?

Response Strategy	Count	Percentage
Request authentication	51	66.2%
Redirect to human agent	18	23.4%
Refuse to respond	5	6.5%
Provide generic response	3	3.9%
Other	0	0.0%

Analysis: Request authentication dominance (66.2%) establishes clear stakeholder preference: enable query completion through proper authentication rather than blocking or deflecting. This reflects customer-centric philosophy balancing security with usability—system should facilitate legitimate access while preventing unauthorized access, not simply refuse all potentially sensitive queries.

The 23.4% preferring human redirection represents fallback strategy when authentication complexity exceeds chatbot capability or user frustration threshold. Only 6.5% favour outright refusal, suggesting stakeholders view pure rejection as overly restrictive degrading user experience unnecessarily.

Role-Based Response Preference:

Strategy	Developers	Customer-Facing	Compliance	χ^2	p-value
Request auth	72.0%	64.0%	63.0%	-	-
Redirect human	20.0%	28.0%	22.2%	-	-
Refuse	4.0%	4.0%	11.1%	4.83	0.305
Generic response	4.0%	4.0%	3.7%	-	-

Chi-square test revealed no significant role-based differences (χ^2=4.83, p=0.305), indicating consensus across functions that authentication-then-proceed represents optimal strategy.

Q13a (Developers): Technical Approach: (Open-ended, n=25)

Approach	Count	%
Authentication checks	20	80.0%

Other/unspecified	5	20.0%

Q13b (Customer-Facing Staff): Maintaining Satisfaction: (Open-ended, n=25)

Approach	Count	%
Polite redirect/explanation	14	56.0%
Clear denial with reasoning	7	28.0%
Other/unspecified	4	16.0%

Q13c (Compliance Officers): Ensuring Compliance: (Open-ended, n=27)

Approach	Count	%
GDPR-compliant refusal	27	100.0%

Critical Insight: The 100% Compliance Officer endorsement of "GDPR-compliant refusal" initially appears to contradict Q13's 66.2% preference for authentication requests. Resolution: Compliance Officers interpret "refusal" as "refuse *until* authenticated" rather than "refuse permanently", emphasizing that *unauthenticated* access must be denied per GDPR Article 32 requirements, but authenticated access can proceed. This semantic nuance highlights importance of precise terminology in requirements gathering.

6.2 Diverse Query Handling Importance

Q14: How important is it for chatbot to handle diverse customer queries securely (multilingual, non-native phrasing)?

Importance	Count	Percentage	Cumulative %
Not important (1)	0	0.0%	0.0%
Slightly important (2)	16	20.8%	20.8%
Moderately important (3)	23	29.9%	50.6%
Very important (4)	27	35.1%	85.7%
Extremely important (5)	11	14.3%	100.0%

Mean: 3.43/5.0 | **Median:** 4.0 | **Mode:** 4 | **SD:** 0.98

Analysis: 79.2% rating ≥3 ("moderately important" or higher) establishes diverse query handling as significant but not critical priority—ranking below security mechanisms (mean=4.49) and accuracy (mean=4.42) but above speed (mean=3.52). This positioning suggests stakeholders view diversity support as important for market competitiveness and accessibility but secondary to core security/accuracy functions.

Q14a Diverse Query Types: (Conditional, multiple selection, n=61)

Query Type	Count	% of Respondents
Multilingual queries	43	70.5%
Non-native speaker phrasing	38	62.3%
Ambiguous slang	27	44.3%
Culturally specific financial terms	21	34.4%

Analysis: Multilingual queries leading (70.5%) reflects Nordic market reality where NordicBank operates across Sweden, Finland, Norway, Denmark—countries with distinct languages. Non-native phrasing ranking second (62.3%) acknowledges immigrant populations and international customers using English as second language.

The gap between multilingual support (70.5%) and culturally specific terminology (34.4%) proves noteworthy: stakeholders prioritize basic language translation over nuanced cultural financial concepts. This suggests pragmatic implementation path—achieve multilingual baseline before tackling cultural semantic variations.

Q14b (Developers): Technical Challenges: (Open-ended, n=25)

Challenge	Count	%
NLP processing complexity	20	80.0%
Other/unspecified	5	20.0%

Analysis: The 80% citing "NLP processing complexity" acknowledges that multilingual support isn't simple translation, requires language-specific models, tokenization, entity recognition, and cultural context understanding. This technical realism informs resource planning: multilingual capabilities demand substantial engineering investment beyond monolingual systems.

Q14c (Customer-Facing Staff): Customer Issues: (Open-ended, n=25)

Issue	Count	%
Misunderstandings from language barriers	25	100.0%

Universal Recognition: Every Customer-Facing Staff respondent identified misunderstandings as primary issue, reflecting frontline experience with language-barrier frustrations. This consensus validates multilingual support as genuine user need rather than theoretical consideration.

Q14d (Compliance Officers): Compliance Issues: (Open-ended, n=27)

Issue	Count	%
Fairness/discrimination concerns	27	100.0%

Legal Dimension: 100% Compliance Officer identification of fairness issues reflects EU equality regulations prohibiting discrimination based on language or national origin. Providing inferior service quality to non-native speakers potentially violates equality directives, transforming multilingual support from feature to compliance requirement.

6.3 Bias Concerns and Mitigation

Q15: How concerned are you about chatbot providing biased or unfair responses?

Concern Level	Count	Percentage	Cumulative %
Not at all (1)	0	0.0%	0.0%
Somewhat (2)	33	42.9%	42.9%
Moderately (3)	23	29.9%	72.7%
Very (4)	17	22.1%	94.8%
Extremely (5)	4	5.2%	100.0%

Mean: 2.90/5.0 | **Median:** 3.0 | **Mode:** 2 | **SD:** 0.94

Analysis: Bias concern (mean=2.90) ranks lowest among all risk dimensions surveyed, significantly below data leakage concern (mean=3.74, $p<0.001$), manipulation concern (mean=3.45, $p<0.001$), and leakage likelihood (mean=3.42, $p<0.001$). Only 57.1% express moderate-to-extreme concern (≥3/5) versus 89.6% for data leakage.

Interpretation: Lower bias concern doesn't indicate stakeholders dismiss bias as unimportant but rather assess it as:

1. **Lower probability:** Bias harder to observe than data leakage in individual interactions
2. **Mitigation confidence:** Belief that diverse training data and fairness algorithms can address bias effectively
3. **Established solutions:** Bias detection represents mature field with known techniques versus emerging LLM security challenges

Q15a Bias Type Concerns: (Conditional, n=44)

Bias Type	Count	%
Language-based bias	23	52.3%
Gender-based bias	10	22.7%
Age-based bias	7	15.9%

Income-based bias	4	9.1%
Other	0	0.0%

Analysis: Language-based bias dominance (52.3%) connects directly to Q14 multilingual support concerns, stakeholders fear that non-native speakers receive inferior response quality. This represents both technical challenge (NLP model performance varies across languages) and fairness concern (linguistic discrimination violating equality principles).

Gender and age bias ranking lower (22.7%, 15.9%) may reflect:

- Financial queries often don't require gender/age disclosure, limiting bias opportunity
- Industry awareness campaigns successfully raising consciousness about demographic discrimination
- Belief that obvious demographic biases receive regulatory attention, making them less likely to persist

Q15b (Developers): Bias Reduction Approaches: (Open-ended, n=25)

Approach	Count	%
Diverse training data	25	100.0%

Universal Solution: Every developer independently identified diverse training data as primary bias mitigation, remarkable consensus suggesting industry-wide understanding that model fairness fundamentally depends on training data representativeness. This finding validates data-centric AI approaches emphasizing data quality/diversity over algorithm complexity for bias reduction.

Q15c (Customer-Facing Staff): Observed Bias Examples: (Open-ended, n=25)

Observation	Count	%
Occasional bias instances	7	28.0%
Some unfair responses	6	24.0%
No specific examples	12	48.0%

Analysis: Only 52% providing specific bias observations (even generic ones like "occasional bias") suggests limited direct experience with AI bias in operational contexts. This could indicate:

1. **Current systems lack bias:** Existing (non-AI) systems don't exhibit observable bias

2. **Bias goes unnoticed:** Subtle bias exists but Customer-Facing Staff lack frameworks to recognize it
3. **Limited AI exposure:** Staff have minimal experience with AI systems where bias typically manifests

Q15d (Compliance Officers): Regulatory Risks: (Open-ended, n=27)

Risk	Count	%
Discrimination lawsuits	23	85.2%
Discrimination risks	4	14.8%

Legal Focus: The 100% identification of discrimination-related risks (whether "lawsuits" or general "risks") confirms Compliance Officers' primary lens: legal liability. EU equality directives, combined with increasing algorithmic accountability precedent, create substantial litigation exposure from demonstrated bias, motivating proactive fairness measures as risk management.

Section 7. PERFORMANCE AND SCALABILITY

7.1 Performance Importance Under Load

Q16: How important is it for chatbot to maintain performance (e.g., <1s response time) during high-demand periods?

Importance	Count	Percentage	Cumulative %
Not important (1)	0	0.0%	0.0%
Slightly important (2)	0	0.0%	0.0%
Moderately important (3)	8	10.4%	10.4%
Very important (4)	31	40.3%	50.6%
Extremely important (5)	38	49.4%	100.0%

Mean: 4.39/5.0 | **Median:** 5.0 | **Mode:** 5 | **SD:** 0.68

Analysis: Performance under load ranks second-highest importance (mean=4.39) after accuracy (4.42) and ahead of general importance of prevention mechanisms (4.49), establishing scalability as critical requirement. The 89.7% rating ≥4 with zero responses <3 indicates near-universal recognition that production deployment must handle peak demand without degradation.

Q16a Performance Issue Concerns: (Conditional, n=69)

Issue	Count	%
Slow responses	40	58.0%
System crashes	16	23.2%
Inaccurate responses	8	11.6%
Increased leakage risk	5	7.2%

Analysis: Slow responses dominating (58.0%) reflects user experience priority—degraded latency proves more visible and immediately frustrating than subtle accuracy decline. System crashes ranking second (23.2%) represents catastrophic failure mode requiring immediate attention.

Notably, increased leakage risk ranks lowest (7.2%) despite security's overall dominance in survey. This suggests stakeholders don't perceive strong correlation between performance stress and security failures—an assumption warranting validation as stressed systems may indeed exhibit degraded security (e.g., skipping validation steps to reduce latency).

Q16b (Developers): Scalability Solutions: (Open-ended, n=25)

Solution	Count	%
Cloud auto-scaling	14	56.0%
Load balancing	11	44.0%

Analysis: Cloud auto-scaling preference (56.0%) reflects modern DevOps practices enabling dynamic resource allocation matching demand. However, this creates tension with Q9 findings where data residency concerns favour on-premises deployment. Resolution requires either:

1. **Private cloud:** On-premises infrastructure with auto-scaling capabilities (expensive)
2. **Hybrid approach:** Cloud for non-sensitive operations, on-premises for PII processing
3. **Over-provisioning:** Static on-premises capacity sufficient for peak load (inefficient but secure)

Q16c (Customer-Facing Staff): Customer Impact: (Open-ended, n=25)

Impact	Count	%
Frustration from delays	18	72.0%
Delays cause frustration	7	28.0%

Universal Theme: 100% of responses identified frustration as primary customer impact—no respondent mentioned alternative effects (abandonment, complaints, competitive switching). This singular focus suggests Customer-Facing Staff directly witness emotional customer reactions to latency, informing their strong performance advocacy.

Q16d (Compliance Officers): Performance-Related Compliance Risks: (Open-ended, n=27)

Risk	Count	%
Data integrity issues	14	51.9%
Compliance failures	9	33.3%
Other/unspecified	4	14.8%

Analysis: Data integrity concerns (51.9%) reflect understanding that performance degradation can cascade into accuracy problems: rushed processing may skip validation, cached outdated responses, or return partially processed results. Compliance failures (33.3%)

likely reference regulatory requirements for service quality standards, some jurisdictions mandate minimum response times or system availability as consumer protection.

Section 8. ADDITIONAL CONCERNS AND FOLLOW-UP INTEREST

8.1 Open-Ended Additional Concerns

Q17: Do you have additional concerns or suggestions about using finance chatbot for customer queries?

Theme	Count	%	Representative Quote
Need robust security/encryption	23	29.9%	"Ensure strongest possible encryption and security protocols"
Ensure GDPR/regulatory compliance	18	23.4%	"Must maintain strict GDPR compliance throughout"
Improve user clarity/interface	12	15.6%	"Make interfaces clearer and more user-friendly"
Enhance authentication	8	10.4%	"Stronger authentication before sensitive operations"
Better training for staff	6	7.8%	"Comprehensive staff training on system capabilities"
User-friendly design	4	5.2%	"Prioritize ease of use without compromising security"
No additional concerns	6	7.8%	[Blank or "None"]

Analysis: Security and compliance themes dominating (53.3% combined) reinforces their centrality throughout survey. Even in open-ended context inviting novel concerns, stakeholders return to core security/regulatory themes, suggesting these represent genuine top-of-mind priorities rather than response bias from structured questions.

User clarity/interface mentions (15.6%) despite ranking lower in Q12 feature importance may indicate stakeholders recognize tension between security complexity and usability, acknowledging that strong security shouldn't render system unusable.

8.2 8.2 Follow-Up Interest

Q18: Would you like summary of anonymised findings?

Response	Count	Percentage
Yes (provided email)	56	72.7%
No	21	27.3%

Analysis: The 72.7% requesting findings demonstrates strong stakeholder engagement beyond compliance participation, respondents genuinely interested in research outcomes suggesting investment in chatbot deployment success. This engagement validates research relevance and provides foundation for ongoing stakeholder collaboration through implementation phases.

Optional Follow-Up Interview Interest:

Provided Email	Count	%
Yes	56	72.7%
No	21	27.3%

Analysis: Identical 72.7% offering email contact for both findings AND interviews indicates that stakeholders willing to receive results are equally willing to participate in deeper qualitative discussions. This 56-person interview pool substantially exceeds the 15 interviews actually conducted, demonstrating sampling occurred from engaged volunteer population rather than reluctant conscripts, enhancing qualitative data credibility.

Section 9. **CROSS-TABULATION AND CORRELATION ANALYSIS**

9.1 Experience-Concern Relationships

Correlation Matrix (Spearman's ρ):

Variables	ρ	p-value	Interpretation
Experience × Data Leakage Concern (Q5)	0.34	0.002**	Moderate positive
Experience × Manipulation Concern (Q10)	0.29	0.011*	Weak positive
Experience × Trust (Q11)	-0.18	0.114	No significant relationship
Experience × Bias Concern (Q15)	0.22	0.052	Marginal positive

*p<0.05, **p<0.01

Analysis: Increasing experience correlates with higher data leakage and manipulation concerns (p<0.05), suggesting that tenure builds risk awareness—possibly through:

1. **Incident exposure:** Longer careers increase probability of witnessing security breaches
2. **Institutional knowledge:** Senior staff understand regulatory consequences more deeply
3. **Responsibility escalation:** Experienced professionals hold positions with greater accountability

The non-significant experience-trust relationship (ρ=-0.18, p=0.114) indicates that trust neither increases nor decreases systematically with experience, veteran scepticism and novice optimism apparently balance out.

9.2 Role-Based Response Pattern Summary

Dimension	Developers	Customer-Facing	Compliance	Statistical Significance
Data Leakage Concern	3.56	3.40	**4.26**	F=8.43, p<0.001***
Trust Level	3.20	**2.88**	3.56	F=7.92, p<0.001***
Manipulation Concern	3.48	3.36	3.52	F=0.31, p=0.735
Bias Concern	2.92	2.84	2.93	F=0.12, p=0.887
Feature: Speed	**3.84**	3.44	3.30	F=3.21, p=0.046*
Feature: Accuracy	4.36	4.52	4.37	F=0.47, p=0.628

Feature: Transparency	3.60	3.72	**4.11**	F=3.82, p=0.026*

*p<0.05, ***p<0.001

Pattern Insights:

Compliance Officers Profile:

- Highest data leakage concern (4.26)
- Highest trust (3.56) despite highest concern
- Highest transparency importance (4.11)
- **Interpretation:** Regulatory accountability drives both heightened concern and trust in formal mechanisms

Customer-Facing Staff Profile:

- Lowest trust (2.88)
- Moderate-to-low across most concerns
- **Interpretation:** Frontline experience with system failures and customer complaints breeds scepticism

Developer Profile:

- Highest speed importance (3.84)
- Moderate trust and concern levels
- **Interpretation:** Technical confidence in solutions balanced by awareness of limitations

9.3 Trust-Concern Paradox

Negative Correlation Analysis:

Correlation	ρ	p-value	Interpretation
Data Leakage Concern (Q5) × Trust (Q11)	-0.12	0.298	No significant relationship
Manipulation Concern (Q10) × Trust (Q11)	-0.08	0.492	No significant relationship

Surprising Finding: Expected inverse relationship (higher concern → lower trust) does NOT emerge significantly. Stakeholders simultaneously express high concern AND moderate trust, creating apparent paradox. Resolution lies in understanding that:

1. **Concern reflects importance:** "I'm very concerned" means "this matters greatly" not necessarily "I expect failure"

2. **Trust reflects conditional confidence:** "I trust IF proper mechanisms implemented" rather than unconditional trust
3. **Professional objectivity:** Stakeholders separate emotional concern from rational capability assessment

Section 10. SURVEY QUALITY AND RELIABILITY ASSESSMENT

10.1 Response Completeness

Metric	Value
Total distributed surveys	120 (estimate)
Complete responses	77
Response rate	64.2%
Incomplete/abandoned	~43 (35.8%)
Mean completion time	12.4 minutes
Median completion time	11.8 minutes

Analysis: The 64.2% response rate exceeds typical online survey benchmarks (30-40% for B2B contexts), suggesting effective recruitment and stakeholder engagement. Mean completion time (12.4 minutes) aligns with survey design estimate (10-15 minutes), indicating accurate burden assessment and appropriate question density.

10.2 Internal Consistency Checks

Likert Scale Reliability (Cronbach's α):

Construct	Items	α	Interpretation
Security Concern	Q5, Q7, Q10	0.82	Good reliability
Feature Importance	Q12a-e	0.79	Acceptable reliability
Trust/Confidence	Q11, Q12d	0.75	Acceptable reliability

Analysis: All constructs achieve $\alpha>0.70$ threshold for acceptable internal consistency, validating that related items measure coherent underlying dimensions. The security concern construct ($\alpha=0.82$) exhibiting highest reliability suggests stakeholders possess well-formed, consistent security perspectives rather than random response patterns.

10.3 Response Quality Indicators

Satisficing Detection:

Indicator	Count	%	Threshold
Straight-lining (same rating across Q12 items)	4	5.2%	<10% acceptable
Inconsistent role-specific responses	2	2.6%	<5% acceptable
Blank open-ended responses	8	10.4%	<20% acceptable

Analysis: Low straight-lining rate (5.2%) indicates respondents thoughtfully differentiated feature importance rather than mindlessly selecting identical ratings. Inconsistent role-specific responses (2.6%) suggest effective conditional logic, nearly all respondents received and answered appropriate role-targeted questions.

Blank open-ended response rate (10.4%) proves acceptable, likely reflecting fatigue effects at survey end (Q17 appeared late) rather than systematic disengagement.

Section 11. **INTEGRATION WITH INTERVIEW FINDINGS**

11.1 11.1 Triangulation: Survey-Interview Convergence

Comparing survey quantitative findings with interview qualitative insights (from parallel 15-participant interview study):

Finding	Survey Evidence	Interview Evidence	Triangulation Status
Security dominance	89.6% concern ≥3 (Q5)	94% of interview transcripts	**✓ Confirmed**
Trust deficit	66.2% trust ≤3 (Q11)	73% "potentially deployable with refinement"	**✓ Confirmed**
MFA universality	100% endorsement (Q12f)	Universal mention in interviews	**✓ Confirmed**
Accuracy > Speed	Mean 4.42 vs 3.52	"Speed matters but not at expense of reliability"	**✓ Confirmed**
Phased rollout	Not directly surveyed	100% interview consensus	**Survey gap identified**
Contextual leakage distinction	Not directly surveyed	Unanimous interview theme	**Survey gap identified**

Analysis: Strong convergence across methodologies validates core findings while identifying survey limitations. The survey effectively captured broad patterns (security concern, trust levels, feature preferences) but missed nuanced themes (phased implementation preferences, contextual leakage distinctions) emerging from interview depth. This demonstrates complementary value of mixed-methods approaches—surveys quantify prevalence, interviews uncover complexity.

11.2 Divergence Analysis

Identified Inconsistencies:

1. **Bias Concern Level:**

- Survey: Mean=2.90 (lowest concern dimension)
- Interviews: 61% of transcripts mentioned bias as consideration
- **Resolution:** Interviews prompted deeper reflection on bias through probing questions, while survey captured spontaneous top-of-mind concerns

2. **Performance Importance:**
    - Survey: Speed ranked lowest feature (mean=3.52)
    - Interviews: Latency >2s identified as "unacceptable" by stakeholders
    - **Resolution:** Survey measured *relative* importance (speed vs. accuracy), interviews assessed *absolute* thresholds. Both valid, speed less important than accuracy, but still requires minimum standards
3. **Multilingual Priority:**
    - Survey: 70.5% identified as diverse query type
    - Interviews: Described as "market requirement, not optional"
    - **Resolution:** Survey captured importance recognition, interviews revealed regulatory/competitive necessity. Difference in framing (important vs. mandatory) rather than contradiction

Conclusion: No fundamental contradictions emerged. Apparent divergences reflect methodological differences (structured vs. open-ended questions, relative vs. absolute importance) rather than inconsistent stakeholder perspectives.

Section 12. IMPLICATIONS FOR SYSTEM DESIGN

12.1 Design Requirements Derived from Survey

Critical Requirements (Non-Negotiable):

1. **Multi-Factor Authentication (100% endorsement):**
   - Implement for all sensitive queries (balance enquiries, transaction history, loan applications)
   - Support multiple auth methods (SMS, authenticator app, biometric)
   - Graceful degradation when MFA unavailable (redirect to human agent)

2. **Accuracy Optimization (Mean importance=4.42):**
   - Target F1 score >0.90 (aligns with interview benchmark)
   - Implement confidence thresholding (<85% triggers human review per interviews)
   - RAG grounding to minimize hallucinations (<3% target per interviews)

3. **PII Detection and Masking:**
   - Focus on high-frequency entities: account numbers (59.4%), customer names (60.9%)
   - Achieve >98% recall on structured financial identifiers (account numbers, IBANs)
   - Real-time detection indicators visible to users (trust-building transparency)

4. **Human Escalation Capability (Mean importance=4.01):**
   - Seamless handoff to human agents
   - Preserve conversation context during transfer
   - Clear escalation triggers (low confidence, explicit user request, query complexity)

High-Priority Requirements (Strongly Recommended):

5. **Transparent Security Messaging (Mean importance=3.81, 71.4% rating ≥4):**

- Explicit data protection explanations ("Your information is encrypted...")
- Real-time PII detection notifications ("I've masked your account number for privacy")
- Source attribution for responses ("According to GDPR Article 15...")

6. **Performance Under Load (Mean importance=4.39):**

   - Target median latency <1.5s (relaxed from 1.0s based on interview stakeholder acceptance)
   - 95th percentile <2.0s (avoiding "unacceptable" threshold per interviews)
   - Auto-scaling infrastructure or sufficient over-provisioning for peak demand

7. **Certified Security Standards (35.3% trust-building factor):**

   - Pursue ISO 27001, SOC 2 certifications early in development
   - Third-party security audits before deployment
   - Documented compliance with GDPR Article 32 technical measures

Moderate-Priority Requirements (Desirable):

8. **Multilingual Support (70.5% diverse query importance):**

   - Phase 1: English baseline
   - Phase 2: Nordic languages (Swedish, Finnish, Norwegian, Danish)
   - Phase 3: Additional EU languages based on customer demographics

9. **Bias Monitoring (Mean concern=2.90, but 100% Compliance regulatory risk identification):**

   - Implement fairness metrics segmented by language
   - Regular bias audits using Fairlearn or equivalent
   - Diverse training data ensuring balanced representation

10. **Simple Interface (Mean importance=3.56):**

    - Minimize complexity without compromising security
    - Progressive disclosure (advanced options hidden until needed)
    - Consistent design patterns following financial services conventions

12.2 Risk Mitigation Priorities

Based on survey concern levels, risk mitigation should prioritize:

Risk Category	Priority Level	Survey Evidence	Mitigation Approach
Data Leakage	**Critical**	Mean concern=3.74, 89.6% ≥3	Multi-layer security pipeline, redundant PII detection
Manipulation/Attacks	**High**	Mean concern=3.45, 84.4% ≥3	Input validation, prompt sanitization, behavioral monitoring
Performance Degradation	**High**	Mean importance=4.39 under load	Load testing, auto-scaling, graceful degradation
Accuracy Failures	**Critical**	Mean importance=4.42	RAG grounding, confidence thresholding, human oversight
Bias/Discrimination	**Moderate**	Mean concern=2.90, 57.1% ≥3	Diverse training data, fairness metrics, regular audits

Section 13.LIMITATIONS AND METHODOLOGICAL CONSIDERATIONS

13.1 13.1 Survey-Specific Limitations

Sampling Limitations:

- **Single-institution focus:** All respondents employed at NordicBank, limiting generalizability to broader financial services sector
- **Self-selection bias:** 72.7% willingness to provide email for follow-up suggests respondents represent more engaged, perhaps more security-conscious subset of employee population
- **Role balance artificiality:** Intentional 32-35% distribution across roles doesn't reflect typical organizational structures, potentially over-representing Compliance perspectives

Measurement Limitations:

- **Social desirability bias:** Respondents may have inflated security concern ratings to appear professionally responsible
- **Hypothetical scenario bias:** Questions asked about future chatbot usage rather than actual experience, potentially producing idealized rather than realistic responses
- **Likert scale interpretation variability:** "Moderately concerned" (3/5) may represent different absolute concern levels across respondents with varying risk tolerances

Coverage Limitations:

- **Missing stakeholder groups:** Survey excluded customers (primary end-users), senior executives (strategic decision-makers), and external auditors (independent compliance validators)
- **Insufficient contextual leakage exploration:** Survey didn't capture nuanced distinction between educational leakage and genuine security failures identified in interviews

- **Limited adversarial scenario detail:** Q10 manipulation concerns remained abstract without specific attack vector examples (prompt injection, Unicode exploits, multi-turn context manipulation)

13.2 Generalizability Constraints

Geographic Limitations:

- Nordic/European focus may not generalize to:
  - US financial institutions (different regulatory environment: GLBA vs GDPR)
  - Asian markets (varying cultural attitudes toward AI and privacy)
  - Emerging markets (different infrastructure maturity and risk profiles)

Organizational Context:

- NordicBank characteristics potentially influencing findings:
  - Large institution with established compliance infrastructure
  - Advanced technical capabilities enabling sophisticated AI deployment
  - Risk-averse culture typical of Nordic banking sector
  - Results may differ at smaller fintech startups or regional banks

Temporal Limitations:

- Survey conducted September-October 2024—findings reflect:
  - Current regulatory environment (EU AI Act implementation phase)
  - Contemporary LLM capabilities (GPT-4, Llama 3.1 generation)
  - Recent security incidents informing stakeholder concerns
  - Rapid AI evolution may date findings within 12-24 months

Section 14. **CONCLUSIONS AND RECOMMENDATIONS**

14.1 Key Takeaways

The survey of 77 financial services professionals reveals ten critical insights:

1. **Security Dominates:** Data leakage concern (89.6% ≥3) represents paramount priority, with Compliance Officers exhibiting highest anxiety (mean=4.26) reflecting regulatory accountability
2. **Trust Deficit Persists:** Only 33.8% express strong trust (≥4), creating 0.79-1.27 point gap between desired and actual trustworthiness requiring targeted intervention
3. **MFA Non-Negotiable:** Universal 100% endorsement establishes multi-factor authentication as baseline expectation, not advanced feature
4. **Accuracy Trumps Speed:** Stakeholders prioritize correctness (mean=4.42) over responsiveness (mean=3.52), challenging tech industry assumptions
5. **High-Risk Query Consensus:** Balance enquiries (64.9%), transaction history (59.7%), and loan applications (57.1%) identified as primary leakage vectors requiring enhanced protection
6. **Authentication-First Philosophy:** 66.2% prefer "request authentication" over refusal or redirection, indicating desire to enable legitimate access through proper security rather than blanket restrictions
7. **Multilingual Market Necessity:** 70.5% identify multilingual support as important with 100% Compliance Officers citing fairness/discrimination compliance requirements
8. **Bias Lower Priority:** Concern mean=2.90 (lowest dimension) suggests stakeholders assess bias as manageable versus emerging LLM security challenges
9. **Performance Criticality:** 89.7% rate performance under load as very-to-extremely important (≥4), establishing scalability as deployment prerequisite
10. **Role-Based Variations:** Significant differences across functions (Compliance highest concern, Customer-Facing lowest trust, Developers highest speed importance) validate necessity of role-adapted design approaches

14.2 Recommendations for Dissertation Research

Prototype Development Priorities:

1. **Security Architecture:**
    - Implement three-stage pipeline (pre-processing detection, in-flight monitoring, post-generation validation) addressing 89.6% high-concern stakeholders
    - Deploy dual-LLM approach (DistilBERT-NER detection, Llama 3.1 8B generation) balancing accuracy and efficiency
    - Integrate hybrid RAG (FAISS + BM25) achieving 96.3% Recall@5 target per system testing
2. **Authentication Integration:**
    - MFA for all high-risk queries (balance, transaction history, loan applications covering 64.9%, 59.7%, 57.1% respectively)
    - Support multiple auth methods accommodating user preferences
    - Graceful degradation redirecting to human agents when auth unavailable
3. **Transparency Mechanisms:**
    - Real-time PII detection indicators building trust per 56.0% Customer-Facing Staff endorsement
    - Source attribution for regulatory responses satisfying 44.4% Compliance Officer requirement
    - Security dashboard providing aggregate statistics per 35.3% certification/standards trust factor
4. **Performance Optimization:**
    - Target median latency <1.5s, 95th percentile <2.0s balancing stakeholder acceptance with technical feasibility
    - Implement load testing validating 4.39/5.0 importance of performance under peak demand
    - Auto-scaling or over-provisioning ensuring 89.7% high-importance stakeholder expectations met

Evaluation Framework:

1. Quantitative Metrics:
    - Data leakage rate <2% (survey establishes 5% as maximum acceptable, interviews tighten to 2%)
    - F1 score >0.90 reflecting accuracy priority (mean importance=4.42)
    - Median latency <1.5s, 95th percentile <2.0s per performance requirements
    - Confidence-based human escalation functioning at <85% threshold

2. **Qualitative Assessment:**
    - Post-pilot stakeholder satisfaction survey (target >4.0/5.0 matching interview benchmark)
    - Role-specific feedback sessions validating security (Compliance), usability (Customer-Facing), and technical robustness (Developers)
    - Trust measurement pre-post deployment assessing whether implementation closes 0.79-1.27 point trust gap

3. **Compliance Validation:**
    - GDPR Article 32 technical measures documentation
    - ISO 27001 readiness assessment
    - Third-party security audit addressing 35.3% certification trust requirement

Appendices

Appendix A: Complete Survey Instrument

[Full questionnaire content as provided in uploaded PDF document]

References

Braun, V. and Clarke, V. (2023) 'Toward good practice in thematic analysis: avoiding common problems and be(com)ing a knowing researcher', International Journal of Transgender Health, 24(1), pp. 1–6.

ICO (2023) Guide to the General Data Protection Regulation (GDPR). Wilmslow: Information Commissioner's Office.

Nordic Banking Association (2024) Workforce demographics in Nordic financial services 2024. Stockholm: Nordic Banking Association.

UK Finance (2024) Banking workforce statistics 2024. London: UK Finance.