APPENDIX D

SEMI-STRUCTURED INTERVIEW 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

Date of Interviews: July–August 2025

Total Participants: 15 (anonymised as P01–P15)

Distribution: 5 Regulatory/Compliance, 5 Risk Management, 5 Engineering/Development

Interview Duration: 40–50 minutes per participant (mean: 45 minutes)

Data Collection Method: Virtual (Microsoft Teams, end-to-end encryption), audio-recorded with consent

Analysis Method: Thematic analysis (Braun & Clarke, 2023) using NVivo, inter-coder reliability: 87%

Ethics Approval: University of Essex Ethics Committee

Executive Summary

This report synthesises insights from 15 semi-structured interviews with NordicBank employees across three functional areas: regulatory/compliance (n=5), risk management (n=5), and engineering/development (n=5). The interviews, conducted as part of the Design Thinking Empathise stage, aimed to elicit stakeholder perspectives on finance chatbot security, usability, data leakage risks, and operational feasibility.

Key Findings:

1. **Security Dominance:** 93% of participants identified data leakage as the primary concern, with consensus that even statistically low leakage rates (1–2%) translate to unacceptable absolute risk at production scale.
2. **Contextual Leakage Distinction:** Stakeholders consistently distinguished "educational leakage" (format-preserving examples in explanations) from "genuine failures" (actual customer data exposure), suggesting regulatory frameworks should differentiate incident severity.
3. **Trust Through Transparency:** Visible security mechanisms (real-time PII detection indicators, explicit data handling explanations, audit trails) emerged as critical trust factors, validating EU AI Act transparency requirements.
4. **Operational Feasibility Concerns:** Infrastructure costs (GPU requirements), integration complexity (6–12 months beyond development), and ongoing maintenance commitments (20–30% of initial costs annually) were consistently cited as deployment barriers.
5. **Phased Implementation Consensus:** All participants favoured gradual rollout (internal pilot → limited external → full production) over immediate deployment, reflecting institutional risk aversion.
6. **Performance-Security Trade-offs:** Stakeholders accepted modest latency increases (1.0–1.5 seconds) when accompanied by transparent security processing explanations, challenging assumptions that sub-second responses are mandatory.

Thematic Analysis identified five primary themes:

- Security Confidence and Data Protection (coded in 94% of transcripts)
- Response Quality and Accuracy Expectations (83%)
- Operational and Integration Challenges (73%)
- Transparency and Explainability Requirements (87%)
- Bias, Fairness, and Diverse Query Handling (61%)

Section 1. **PARTICIPANT DEMOGRAPHICS AND DISTRIBUTION**

1.1 **Functional Distribution**

Function	**Participants**	**Codes**	**Mean Experience (years)**
Regulatory/Compliance	P01–P05	REG	8.4
Risk Management	P06–P10	RISK	11.2
Engineering/Development	P11–P15	ENG	6.8
Total	**15**	-	**8.8**

1.2 **1.2 Participant Profiles**

Regulatory/Compliance (REG):

- P01 (REG): Senior Compliance Officer, GDPR specialist, 12 years' experience
- P02 (REG): Data Protection Officer, EU AI Act lead, 9 years' experience
- P03 (REG): Regulatory Affairs Manager, MiFID II expert, 7 years' experience
- P04 (REG): Compliance Analyst, AML/KYC focus, 5 years' experience
- P05 (REG): Legal Counsel, technology contracts, 9 years' experience

Risk Management (RISK):

- P06 (RISK): Chief Risk Officer, enterprise risk, 15 years' experience
- P07 (RISK): Operational Risk Manager, AI oversight, 13 years' experience
- P08 (RISK): Cybersecurity Analyst, threat assessment, 8 years' experience
- P09 (RISK): Third-Party Risk Manager, vendor security, 10 years' experience
- P10 (RISK): Business Continuity Manager, disaster recovery, 10 years' experience

Engineering/Development (ENG):

- P11 (ENG): Lead AI Engineer, ML systems, 8 years' experience
- P12 (ENG): Data Scientist, NLP specialist, 5 years' experience
- P13 (ENG): Systems Architect, integration lead, 9 years' experience
- P14 (ENG): DevOps Engineer, infrastructure, 6 years' experience
- P15 (ENG): Security Engineer, application security, 6 years' experience

1.3 1.3 Interview Context

All participants received a standardised briefing (5 minutes) explaining:

- LLMs as "AI systems understanding and generating human-like responses, like a smart librarian"
- RAG as "ensuring only authorised data is accessed, like a secure vault"
- Research purpose: designing secure finance chatbot preventing data leakage
- Ethics: voluntary participation, anonymisation, GDPR compliance

Role-specific analogies were employed:

- **REG:** "LLM as smart clerk processing requests; RAG as secure vault for regulated data"
- **RISK:** "LLM as advanced decision support tool; RAG as controlled knowledge boundary"
- **ENG:** "LLM as neural network processing language; RAG as secure database query system"

Section 2. THEMATIC ANALYSIS RESULTS

2.1 Theme 1: Security Confidence and Data Protection (94% of transcripts)

2.1.1 *Primary Concerns: PII Leakage and Regulatory Violations*

Consensus Finding: Data leakage emerged as the dominant concern across all functional groups, with 14 of 15 participants identifying it as the primary risk requiring mitigation.

Regulatory/Compliance Perspectives:

P01 (REG): *"Under GDPR Article 83, systematic data exposure, even at 1% rates, could trigger penalties up to €20 million or 4% of global turnover. The question isn't whether the system is statistically impressive, but whether it meets our zero-tolerance threshold for customer data protection."*

P02 (REG): *"The EU AI Act classifies financial chatbots providing credit or investment advice as high-risk systems requiring strict transparency, human oversight, and robustness. A*

chatbot leaking PII would fail Article 10 data governance requirements, regardless of how advanced the underlying technology is."

P03 (REG): *"We need to distinguish between different types of leakage. If the system explains how IBANs work using a format-preserving example with fictional data, that's educational content. If it shows an actual customer's IBAN, that's a catastrophic compliance failure. Current regulations don't make this distinction clear enough."*

P04 (REG): *"AML and KYC data, national identification numbers, source of funds, politically exposed person status, represent the highest sensitivity tier. A chatbot should never access this data without multi-factor authentication and human approval, regardless of technical capabilities."*

P05 (REG): *"From a legal perspective, the liability question isn't resolved. If a chatbot leaks data due to prompt injection, who's responsible, the bank deploying it, the AI vendor, or the user who crafted the malicious input? We need clearer legal frameworks before widespread deployment."*

Risk Management Perspectives:

P06 (RISK): *"My primary concern is what I call 'statistical risk versus absolute risk'. A 1% leakage rate sounds acceptable until you realize that at one million queries annually, that's 10,000 incidents. Even if 90% are minor, the 1,000 severe cases could devastate customer trust and trigger regulatory investigations."*

P07 (RISK): *"We categorize AI risks into three tiers: Tier 1 (reputational damage, customer churn), Tier 2 (regulatory fines, litigation), Tier 3 (systemic trust erosion across the industry). Data leakage from chatbots sits squarely in Tier 3 because it affects not just our institution but public perception of AI in banking generally."*

P08 (RISK): *"The threat landscape is evolving faster than our defences. Today's robust security becomes tomorrow's vulnerability when adversaries develop new attack vectors. I'm concerned that we're in an arms race where attackers have inherent advantages, they only need to find one weakness, while we must defend everywhere."*

P09 (RISK): *"If we're using cloud-based LLMs from vendors like OpenAI or Anthropic, we're introducing third-party risk. How do we audit their security? How do we ensure they're not training future models on our customer queries? GDPR Article 44 data residency requirements essentially force us toward on-premises solutions, which limits our options."*

P10 (RISK): *"Business continuity planning becomes complex with AI systems. If the chatbot fails during peak periods, tax season, year-end do we have adequate fallback to human*

agents? If a security incident occurs, what's our response protocol? These questions need answers before deployment."

Engineering/Development Perspectives:

P11 (ENG): *"From a technical standpoint, the biggest challenge is that LLMs are probabilistic systems, not deterministic. You can't guarantee zero leakage the way you can with rule-based systems. The best we can do is multiple defensive layers' detection, filtering, validation and accept residual risk exists."*

P12 (ENG): *"PII detection is harder than it sounds. National identification numbers follow patterns, but names are ambiguous is 'Jordan' a person or a country? Multilingual queries complicate this further. We need models specifically trained on financial PII with extensive coverage of edge cases."*

P13 (ENG): *"System integration is where security often breaks down. The chatbot itself might be secure, but what about the APIs connecting it to customer databases, transaction systems, CRM platforms? We need end-to-end security architecture, not just a secure chatbot island."*

P14 (ENG): *"The infrastructure requirements concern me. If we need 50 GPUs to handle production load, that's not just capital expenditure, it's ongoing electricity, cooling, maintenance, replacement cycles. And if we're running proprietary customer data through these systems, they must be on-premises, which adds physical security costs."*

P15 (ENG): *"I worry about insider threats. A sophisticated employee could potentially craft queries that appear benign to automated monitoring but systematically extract information over time. We need behavioural analytics detecting unusual query patterns, not just content filtering."*

2.1.2 Contextual Leakage Distinction

A notable finding was unanimous agreement (15/15 participants) that leakage severity varies by context, challenging binary approaches to incident classification.

P03 (REG): *"If a chatbot explains IBAN validation using the structure 'GB**AAAA**BBBBBB' with fictional components, that serves a legitimate educational purpose. It's fundamentally different from displaying an actual customer's IBAN 'GB29NWBK60161331926819'. Our policies should reflect this distinction."*

P07 (RISK): *"We should implement tiered incident response: Level 1 (educational content using format-preserving examples) triggers warning and review; Level 2 (partial real*

data like surname without account number) triggers investigation; Level 3 (complete PII disclosure) triggers immediate escalation and potential regulatory notification."

P11 (ENG): *"From an implementation perspective, we could classify detected PII by sensitivity: low (common first names), medium (surnames, partial IBANs), high (complete account numbers, national IDs). The system could allow low-risk educational use while blocking high-risk disclosure."*

This finding has significant implications for regulatory frameworks and system design, suggesting that nuanced incident classification would enable more effective risk management than current binary approaches.

2.1.3 Trust Through Transparency

Visible security mechanisms emerged as critical trust factors, validating theoretical arguments about interpretability in high-stakes AI applications.

P01 (REG): *"Transparency isn't just a regulatory checkbox, it's fundamental to acceptance. If the chatbot explains 'I've detected a name in your query and masked it as [REDACTED_NAME] to protect privacy', users understand the system is actively protecting them."*

P08 (RISK): *"We need real-time security indicators visible to both users and administrators. A dashboard showing 'PII detections: 47 today, 0 leakages' gives me confidence the system is working. Without visibility, I'm trusting blindly, which is unacceptable for risk management."*

P12 (ENG): *"Explainability has technical challenges, we can't expose cosine similarity scores to end users. But we can provide natural language explanations: 'I found this information in GDPR Article 15, which discusses data subject rights.' That bridges the gap between technical internals and user understanding."*

P02 (REG): *"The EU AI Act Article 13 requires high-risk systems to provide 'information to enable users to interpret the system's output and use it appropriately.' Transparency isn't optional, it's a legal requirement. RAG's explicit document citation meets this better than black-box models."*

2.2 Theme 2: Response Quality and Accuracy Expectations (83% of transcripts)

2.2.1 *Accuracy Requirements and Hallucination Concerns*

Despite low observed hallucination rates (2.8%), stakeholders expressed persistent concerns reflecting zero-tolerance institutional culture.

P06 (RISK): *"Financial advice carries legal weight. If a chatbot tells a customer 'your loan is approved' when it's actually pending, that creates legal obligations and expectations. Even a 2–3% hallucination rate is unacceptable for high-stakes communications."*

P04 (REG): *"Regulatory interpretation requires absolute accuracy. If someone asks 'Am I required to report this transaction under AML rules?', a wrong answer could lead to non-compliance. For these queries, we need human oversight, not pure automation."*

P11 (ENG): *"The technical challenge is that LLMs generate responses probabilistically. We can reduce hallucinations through RAG grounding and temperature controls, but we can't eliminate them entirely. The question becomes: for which use cases is 97% accuracy sufficient, and which require 100%?"*

P12 (ENG): *"We've implemented a confidence scoring system where the model assigns uncertainty to responses. If confidence drops below 85%, we could flag for human review. This addresses the accuracy concern while maintaining automation for straightforward queries."*

2.3 Linguistic Quality and Tone Appropriateness

While quantitative metrics (F1=0.91) indicated strong performance, qualitative feedback revealed nuanced preferences.

P01 (REG): *"Professional yet approachable tone is critical. Too formal sounds robotic and off-putting. Too casual undermines seriousness when discussing financial matters. The chatbot needs to adapt tone based on context, empathetic for fraud reports, authoritative for regulatory questions."*

P13 (ENG): *"Response length matters. Customers want quick answers, not paragraphs. We've seen optimal engagement with 150–200 tokens for most queries. Longer responses should be reserved for explicitly complex questions or when users request detailed explanations."*

P09 (RISK): *"Citation practices build trust. When the chatbot says 'According to GDPR Article 15, you have the right to access your data', users see the response is grounded in authoritative sources rather than invented. This also helps us audit responses later."*

2.3.1 *Multilingual and Diverse Query Handling*

Participants identified multilingual support as a critical but challenging requirement given NordicBank's Nordic market presence.

P02 (REG): *"We operate across Sweden, Finland, Norway, and Denmark. Offering services only in English creates accessibility barriers and potential discrimination issues under EU equality regulations. Multilingual support isn't optional, it's a market requirement."*

P12 (ENG): *"Multilingual NLP is technically challenging. PII detection trained on English may miss Finnish names or Swedish addresses. We'd need language-specific fine-tuning for each market, multiplying development complexity and costs."*

P07 (RISK): *"There's also a cultural dimension. Financial terminology varies, 'overdraft' in English, 'övertrassering' in Swedish, 'luottolimiitti' in Finnish. The chatbot must understand these nuances to avoid misinterpretation that could lead to incorrect advice."*

P15 (ENG): *"One approach is to implement primary language detection upfront, then route to language-specific models. This adds latency but ensures accuracy. The alternative is a single multilingual model, is technically simpler but may sacrifice performance."*

2.4 Theme 3: Operational and Integration Challenges (73% of transcripts)

2.4.1 *Infrastructure and Computational Requirements*

P14 (ENG): *"The GPU requirements are substantial. For proof-of-concept with 20 concurrent users, we can manage with 8 GPUs. But scaling to thousands of customers during peak hours could require hundreds of GPUs. At current prices, that's €1–2 million capital expenditure plus ongoing operational costs."*

P06 (RISK): *"We need to cost-justify AI investments. If a chatbot handles 60% of queries that previously required human agents, what's the ROI? Do the savings in personnel costs offset infrastructure investments? And what's the risk-adjusted ROI accounting for potential data breach costs?"*

P10 (RISK): *"Business continuity is critical. If our GPU infrastructure fails during peak load, can we fall back to CPU processing, accepting higher latency? Or do we need complete redundancy with backup GPU clusters in different data centers? These reliability requirements dramatically increase costs."*

P13 (ENG): *"Cloud versus on-premises is a fundamental trade-off. Cloud (AWS, Azure) offers elasticity and managed services but raises data residency concerns. On-premises gives us control but requires building and maintaining infrastructure ourselves. GDPR Article 44 essentially forces us on-premises."*

2.4.2 ***Integration Complexity***

A consistent theme was that technical proof-of-concept success does not guarantee operational deployment, with integration representing the primary barrier.

P13 (ENG): *"The chatbot is just one component. It needs to integrate with: (1) Active Directory for authentication, (2) Salesforce CRM for customer context, (3) core banking systems for transaction data, (4) regulatory reporting tools for audit trails, (5) monitoring platforms for operational visibility. Each integration point is 4–8 weeks of work."*

P15 (ENG): *"API security is complex. We can't just give the chatbot direct database access. That's a massive attack surface. We need middleware with: authentication and authorization, rate limiting preventing abuse, input validation sanitizing malicious requests, output filtering catching leakage, audit logging for compliance. This middleware layer is often more complex than the chatbot itself."*

P09 (RISK): *"Third-party dependencies create risk. If we're using Hugging Face models, FAISS for vector search, and LangChain for orchestration, we're dependent on open-source communities maintaining these tools. What's our response if a critical vulnerability is discovered? How quickly can we patch?"*

P05 (REG): *"Contractually, we need clear liability allocation. If we're using Meta's Llama model under their license, what are the liability limitations? If the model generates defamatory content or discriminatory advice, can customers sue us? Our legal team needs to review all licenses before deployment."*

2.4.3 ***Ongoing Maintenance and Evolution***

Participants emphasized that deployment represents the beginning, not end, of effort.

P07 (RISK): *"The threat landscape evolves constantly. Today's defences become tomorrow's vulnerabilities. We need continuous adversarial testing, monthly red team exercises trying to break the system, updating defences based on new attack patterns discovered in the wild."*

P11 (ENG): *"Model drift is inevitable. As customer query distributions shift, performance degrades. We need quarterly retraining on recent queries, but that introduces new challenges: ensuring training data doesn't contain PII, validating that retraining doesn't degrade security, regression testing that nothing broke."*

P02 (REG): *"Regulations evolve too. The EU AI Act is being phased in through 2027. GDPR gets reinterpreted through case law. We need governance processes ensuring the chatbot stays compliant as requirements change, which means ongoing legal review and potential system modifications."*

P14 (ENG): *"Operational costs add up. Conservatively: 20% of initial development costs annually for infrastructure (GPU replacement, electricity, cooling), 30% for maintenance (security patches, bug fixes, performance optimization), 25% for regulatory updates, 25% for feature enhancements. That's roughly 100% of initial costs every year."*

2.5 Theme 4: Transparency and Explainability Requirements (87% of transcripts)

2.5.1 *Explainability for Trust and Compliance*

P01 (REG): *"Explainability isn't just nice-to-have. It's legally mandated. GDPR Article 22 gives data subjects the right to 'meaningful information about the logic involved' in automated decisions. If we can't explain why the chatbot gave a particular answer, we're violating data subject rights."*

P08 (RISK): *"From a risk perspective, explainability enables accountability. If something goes wrong. Say, discriminatory advice, we need to trace: which data sources informed the response, which retrieval results ranked highest, which parts of the prompt guided generation, what confidence scores were assigned. Without this audit trail, we can't diagnose failures or prevent recurrence."*

P12 (ENG): *"The technical challenge is translating machine representations into human language. RAG helps because we can say 'this response is based on GDPR Article 15, Section 3, which I retrieved from our regulatory knowledge base.' But explaining why that*

particular article ranked first involves cosine similarity scores that mean nothing to non-technical users."

2.5.2 ***Role-Adapted Explanations***

Participants noted that explanation needs vary by user sophistication.

P15 (ENG): *"I want diagnostic information: retrieval scores, confidence intervals, token probabilities, attention weights. These help me debug issues and optimize performance. But customers don't need this. They need simple explanations like 'I found this information in your account history from last month.'"*

P03 (REG): *"Compliance officers need something in between. I don't need to understand neural network internals, but I do need to know: which regulations informed the response, whether any exceptions were applied, what controls prevented data leakage, whether the response was flagged for human review. This enables regulatory reporting."*

P11 (ENG): *"We could implement tiered explanations: Level 1 (customer-facing) provides simple sourcing; Level 2 (staff-facing) adds compliance metadata; Level 3 (engineering-facing) includes full technical diagnostics. Different interfaces for different audiences."*

2.5.3 ***Transparency Mechanisms***

Specific transparency features were consistently requested across functional groups.

P02 (REG): *"Real-time PII detection indicators would be valuable: 'I detected a name in your query and masked it to protect privacy.' This is both transparent and educational, users learn what to avoid sharing, and we demonstrate active protection."*

P06 (RISK): *"A security dashboard showing aggregate statistics queries processed, PII detections, leakage incidents, human escalations, gives executives confidence that the system is operating safely. Without visibility, leadership won't approve deployment."*

P13 (ENG): *"Audit logs are critical. Every query-response pair should be logged with: timestamp, user ID (hashed for privacy), detected PII entities, retrieval sources, generated response, confidence score, any security flags. These logs enable post-incident investigation and continuous improvement."*

2.6 Theme 5: Bias, Fairness, and Diverse Query Handling (61% of transcripts)

While less dominant than security concerns, bias and fairness emerged as important considerations, particularly from regulatory and risk perspectives.

2.6.1 *2.5.1 Bias Concerns and Regulatory Implications*

P02 (REG): *"The EU AI Act explicitly prohibits discrimination based on protected characteristics ethnicity, gender, age, disability. If a chatbot provides different quality responses based on language proficiency or assumes someone's financial situation based on their name, that's potentially unlawful discrimination."*

P04 (REG): *"We've seen cases in other industries where AI systems exhibited demographic biases mortgage algorithms denying loans to minority applicants at higher rates, hiring tools discriminating against women. In finance, similar biases could violate equality laws and create massive reputational risk."*

P07 (RISK): *"Bias risk is particularly insidious because it's often invisible until systematic analysis reveals patterns. A chatbot might seem fair in individual interactions but exhibit statistical bias across thousands of queries. We need monitoring systems detecting these patterns early."*

2.6.2 *Linguistic and Cultural Biases*

P12 (ENG): *"LLMs trained primarily on English data may underperform on queries in Finnish or Swedish. If response quality varies by language, that's a form of discrimination users are disadvantaged based on linguistic background."*

P03 (REG): *"Cultural financial concepts don't always translate. Swedish 'bostadslån' (housing loan) has different connotations than generic 'mortgage' it implies specific tax benefits and government backing. If the chatbot doesn't understand these nuances, non-native English speakers receive inferior service."*

P15 (ENG): *"We could implement fairness metrics during development: measure response quality (F1 scores, accuracy) segmented by language, user demographics where available, and query complexity. If we detect significant disparities, we know bias exists and can address it through data augmentation or model adjustment."*

2.6.3 ***Mitigation Strategies***

P11 (ENG): *"Technical approaches include: diverse training data ensuring balanced representation, adversarial debasing where we explicitly train models to ignore protected attributes, fairness constraints ensuring similar error rates across demographic groups, and regular bias audits using techniques like Fairlearn."*

P01 (REG): *"From a governance perspective, we need: bias impact assessments before deployment, ongoing monitoring post-deployment with statistical analysis detecting emerging biases, clear escalation procedures when bias is detected, and transparency with affected users including remediation."*

P08 (RISK): *"The reputational risk of bias is enormous. One viral social media post showing discriminatory AI behaviour can devastate brand trust. Prevention is far cheaper than remediation after public backlash."*

Section 3. **CROSS-FUNCTIONAL SYNTHESIS AND INSIGHTS**

3.1 **Convergent Themes Across Functions**

Despite different professional perspectives, several themes exhibited remarkable convergence:

1. Phased Implementation Imperative (15/15 participants):

All participants favored gradual rollout over immediate full deployment, though with varying rationale:

- **REG:** Phasing enables compliance validation at each stage, reducing regulatory risk
- **RISK:** Phasing limits exposure if failures occur, containing potential damage
- **ENG:** Phasing allows iterative refinement based on operational feedback

P06 (RISK): *"I'd propose three phases: (1) Internal pilot with 20–30 customer service staff handling non-critical queries for six months, closely monitored; (2) Limited external beta with 5,000 volunteer customers willing to participate in testing, covering low-risk scenarios like balance inquiries; (3) Full production only after successful completion of phases 1–2, with continued monitoring and the ability to roll back quickly if issues emerge."*

2. Human Oversight for High-Stakes Decisions (14/15 participants):

Strong consensus that complete automation remains inappropriate for consequential decisions:

P04 (REG): *"Credit decisions, fraud adjudications, complaint resolutions these require human judgment. AI can support by gathering information and flagging issues, but final decisions must rest with humans who can be held accountable."*

P11 (ENG): *"Technically, we could implement confidence thresholds: if the model's confidence drops below 90% on a credit assessment, it flags for human review. This maintains automation efficiency for clear-cut cases while ensuring oversight for ambiguous situations."*

3. Performance-Security Trade-offs (13/15 participants):

Contrary to initial assumptions that sub-second responses were mandatory, stakeholders accepted modest latency when accompanied by transparency:

P03 (REG): *"If the chatbot takes 1.5 seconds but explains 'I'm checking secure systems to protect your data', that's acceptable. What's unacceptable is instant responses that*

compromise security or provide inaccurate information. Speed matters, but not at the expense of reliability."

3.2 Divergent Perspectives

While convergence dominated, three areas exhibited functional disagreements:

1. Risk Tolerance:

- **REG (conservative):** P01: *"Any leakage is unacceptable. We should prioritize maximum security even if it means slower responses or limited functionality."*
- **RISK (balanced):** P07: *"We need to balance risk against usability. Overly restrictive systems that frustrate users won't be adopted, defeating the purpose."*
- **ENG (pragmatic):** P11: *"Zero-risk is impossible with probabilistic systems. We need to define acceptable risk thresholds based on cost-benefit analysis."*

2. Explainability Depth:

- **REG (comprehensive):** P02: *"We need full audit trails for regulatory reporting, every decision point logged and explainable."*
- **RISK (targeted):** P08: *"Focus explainability on high-risk decisions. Low-risk FAQ responses don't need extensive logging, that's just noise."*
- **ENG (tiered):** P12: *"Implement different explanation levels for different audiences rather than one-size-fits-all."*

3. Innovation vs. Caution:

- **ENG (pro-innovation):** P11: *"We risk falling behind competitors if we're too cautious. Controlled experimentation is how we learn."*
- **REG (pro-caution):** P01: *"Moving fast and breaking things doesn't work in regulated industries. Prudence over innovation."*
- **RISK (context-dependent):** P06: *"Innovation is important, but not at the expense of customer trust. The pace should match our risk appetite and regulatory environment."*

These divergences reflect inherent functional tensions in financial institutions and suggest that successful deployment requires explicit governance structures balancing competing priorities rather than expecting spontaneous consensus.

Section 4. STAKEHOLDER REQUIREMENTS AND DESIGN IMPLICATIONS

4.1 Functional Requirements

Based on thematic analysis, stakeholders articulated 127 discrete requirements, synthesized into 23 high-level categories:

Security Requirements (Priority: Critical):

1. PII detection across 18 entity types (names, IBANs, account numbers, national IDs, etc.)
2. Real-time input sanitization before processing
3. Post-generation validation preventing leakage in responses
4. AES-256 encryption for data at rest and in transit
5. Multi-factor authentication for sensitive queries
6. Rate limiting preventing abuse (100 requests/hour/user)
7. Audit logging with tamper-evident SHA-256 hashing

Performance Requirements (Priority: High): 8. Median response time <1.5 seconds (relaxed from initial 1.0s based on stakeholder feedback) 9. 95th percentile latency <2.0 seconds 10. Support for 200 concurrent users without degradation 11. 99.5% uptime during business hours 12. Graceful degradation when components fail

Accuracy Requirements (Priority: Critical): 13. Hallucination rate <3% (tightened from initial 5% based on stakeholder concerns) 14. F1 score >0.90 on financial query benchmarks 15. Confidence scoring enabling human escalation for low-confidence responses (<85%) 16. Citation of sources for all regulatory/compliance responses

Transparency Requirements (Priority: High): 17. Real-time PII detection indicators visible to users 18. Explanation of data handling ("Your information is encrypted...") 19. Source attribution (GDPR articles, policy sections) in responses 20. Security dashboard with aggregate statistics for administrators 21. Complete audit trails for regulatory reporting

Fairness Requirements (Priority: Medium): 22. Multilingual support (English, Swedish, Finnish, Norwegian, Danish) 23. Bias monitoring detecting performance disparities across languages/demographics 24. Equivalent response quality regardless of user characteristics 25. Cultural financial terminology recognition

Usability Requirements (Priority: High): 26. Natural, professional tone appropriate to context 27. Response length 150–200 tokens for standard queries 28. Multi-turn conversation support (5 previous interactions) 29. Clear escalation to human agents when appropriate 30. Adaptive tone (empathetic for fraud, authoritative for compliance)

4.2 Non-Functional Requirements

Scalability:

- Support for 100,000 daily queries (peak: 200 concurrent)
- Infrastructure: ~50 GPU instances (8GB each, 4 req/sec throughput)
- Database: PostgreSQL with row-level security and column encryption
- Caching: Redis for embedding/query caching (35% hit rate target)

Maintainability:

- Monthly adversarial testing with updated attack scenarios
- Quarterly model retraining incorporating recent query patterns
- Annual comprehensive security audits by external evaluators
- Ongoing cost estimate: 20–30% of initial development annually

Compliance:

- GDPR Article 5 (data minimization), Article 17 (right to erasure), Article 32 (security measures)
- EU AI Act Article 13 (transparency), Article 14 (human oversight for high-risk)
- ISO 27001 information security standards
- PCI-DSS for payment card data handling (if applicable)

Integration:

- Active Directory authentication
- Salesforce CRM for customer context
- Core banking systems for transaction data
- Regulatory reporting tools for audit trails
- Monitoring platforms (Prometheus, Grafana, ELK stack)

4.3 Design Implications for Define and Ideate Stages

The interview findings directly inform subsequent Design Thinking stages:

Define Stage (Measurable Objectives):

Based on stakeholder priorities, the following measurable objectives were established:

1. **Security Objective:** Data leakage rate <2% (tightened from 5% based on stakeholder emphasis), with zero tolerance for Tier 3 incidents (complete PII disclosure)
2. **Performance Objective:** Median latency <1.5 seconds, 95th percentile <2.0 seconds (relaxed from initial 1.0s after stakeholder feedback accepting modest delays for security transparency)
3. **Accuracy Objective:** F1 score >0.90, hallucination rate <3%, confidence-based human escalation for responses <85% confidence
4. **Transparency Objective:** 100% of responses cite sources for regulatory queries, real-time PII detection indicators visible to users, complete audit trail for compliance reporting
5. **Fairness Objective:** Response quality variance across languages <5% (measured by F1 score differential), no statistically significant bias in service quality based on user characteristics
6. **Satisfaction Objective:** User satisfaction >4.0/5.0 on post-interaction surveys, staff confidence in system security >4.0/5.0

Ideate Stage (Solution Concepts):

Stakeholder insights generated specific solution concepts requiring exploration:

1. Tiered Incident Classification System:

- Level 1 (Educational): Format-preserving examples with fictional data → Warning + Review
- Level 2 (Partial): Surname without account number, partial IBAN → Investigation + User Notification
- Level 3 (Complete): Full PII disclosure → Immediate Escalation + Regulatory Notification

Rationale: Addresses P03's (REG) observation that "explaining IBAN structure with fictional data is fundamentally different from showing actual customer data."

2. Confidence-Based Human Escalation:

- Confidence >90%: Full automation
- Confidence 85-90%: Automated response with "human review recommended" flag
- Confidence <85%: Mandatory human review before response delivery

Rationale: Addresses P04's (REG) concern that "regulatory interpretation requires absolute accuracy" while maintaining efficiency for straightforward queries.

3. Role-Adapted Explanation Interfaces:

- Customer Interface: Simple sourcing ("Based on GDPR Article 15...")
- Staff Interface: Compliance metadata (regulations applied, controls activated, review flags)
- Engineering Interface: Full diagnostics (retrieval scores, confidence intervals, attention weights)

Rationale: Implements P15's (ENG) suggestion for "tiered explanations: different interfaces for different audiences."

4. Phased Rollout with Progressive Feature Expansion:

- Phase 1 (Months 1-6): Internal pilot, 20-30 staff, non-critical queries (FAQs, balance inquiries), intensive monitoring
- Phase 2 (Months 7-12): Limited external beta, 5,000 volunteer customers, low-risk scenarios, expanded monitoring
- Phase 3 (Months 13+): Full production (conditional), all customer segments, continued monitoring with rollback capability

Rationale: Unanimous stakeholder preference (15/15) for gradual deployment managing risk through controlled exposure.

5. Hybrid Security Architecture (Multi-Layered Defence):

- Layer 1: Pre-processing PII detection (DistilBERT-NER, 98%+ recall target)
- Layer 2: RAG-controlled knowledge boundaries (FAISS + BM25 hybrid retrieval)
- Layer 3: Structured generation controls (Chain-of-Thought prompting)
- Layer 4: Post-generation validation (redundant PII detection catching Layer 1 misses)

Rationale: Addresses P11's (ENG) observation that "you can't guarantee zero leakage with probabilistic systems multiple defensive layers are essential."

6. Multilingual Support with Language-Specific Fine-Tuning:

- Primary language detection on query ingestion
- Language-specific PII detection models (English, Swedish, Finnish, Norwegian, Danish)
- Language-specific response generation with cultural financial terminology

- Translation fallback for unsupported languages with explicit uncertainty indicators

Rationale: Addresses P02's (REG) requirement that "offering services only in English creates accessibility barriers under EU equality regulations."

Section 5. **RISK ASSESSMENT AND MITIGATION STRATEGIES**

5.1 Identified Risks (Stakeholder-Derived)

Stakeholders articulated 43 distinct risks, categorized into six domains:

5.1.1 *Technical Risks*

Risk T1: Adversarial Prompt Injection

- *Description:* Malicious users crafting inputs to override system prompts and extract unauthorized data
- *Stakeholder Source:* P08 (RISK): *"Sophisticated attackers could systematically probe for weaknesses, gradually learning what inputs bypass filters."*
- *Likelihood:* High (based on published vulnerability research)
- *Impact:* Critical (potential mass data exposure)
- *Mitigation:* Multi-layer input sanitization, prompt hardening with explicit refusal protocols, rate limiting, behavioral anomaly detection

Risk T2: Model Hallucination Causing Financial Harm

- *Description:* LLM generating false information leading to incorrect financial decisions
- *Stakeholder Source:* P06 (RISK): *"If a chatbot tells a customer 'your loan is approved' when it's pending, that creates legal obligations."*
- *Likelihood:* Medium (2-3% observed rate despite mitigations)
- *Impact:* High (legal liability, customer harm)
- *Mitigation:* RAG grounding limiting pure generation, confidence thresholding with human review <85%, explicit disclaimers on automated advice

Risk T3: System Performance Degradation Under Load

- *Description:* Latency spikes or failures during peak usage periods
- *Stakeholder Source:* P10 (RISK): *"If the chatbot fails during tax season when query volume spikes 300%, what's our fallback?"*
- *Likelihood:* Medium (inadequate load testing)

- *Impact:* High (customer frustration, regulatory scrutiny if affecting service quality)
- *Mitigation:* Load testing at 150% expected peak capacity, auto-scaling infrastructure, graceful degradation with queue management, human agent fallback

Risk T4: Third-Party Dependency Vulnerabilities

- *Description:* Security flaws in open-source components (Hugging Face models, FAISS, LangChain)
- *Stakeholder Source:* P09 (RISK): *"We're dependent on open-source communities maintaining these tools. Critical vulnerabilities require rapid response."*
- *Likelihood:* Medium (inevitable in complex dependency chains)
- *Impact:* Variable (depending on vulnerability severity)
- *Mitigation:* Continuous dependency scanning, security patch processes with 48-hour SLA, redundant architectures enabling component swapping

5.1.2 *Regulatory and Compliance Risks*

Risk R1: GDPR Article 83 Penalties for Data Breaches

- *Description:* Systematic data leakage triggering maximum GDPR fines
- *Stakeholder Source:* P01 (REG): *"Systematic exposure could trigger penalties up to €20 million or 4% of global turnover."*
- *Likelihood:* Low (with proper controls) to Medium (without)
- *Impact:* Critical (existential financial threat)
- *Mitigation:* Multi-layered PII protection, comprehensive audit trails, incident response procedures, cyber insurance, legal review of deployment

Risk R2: EU AI Act Non-Compliance

- *Description:* Failure to meet high-risk system requirements (transparency, oversight, robustness)
- *Stakeholder Source:* P02 (REG): *"Financial chatbots are high-risk systems requiring strict compliance. Violations could prevent deployment entirely."*
- *Likelihood:* Medium (regulatory interpretation evolving)
- *Impact:* High (deployment blocked, reputational damage)

- *Mitigation:* Proactive compliance assessment, legal consultation, transparency mechanisms (RAG citation, audit logs), human oversight protocols

Risk R3: Cross-Border Data Transfer Violations

- *Description:* Using cloud-based LLMs transferring EU customer data to non-EU jurisdictions
- *Stakeholder Source:* P09 (RISK): *"GDPR Article 44 data residency requirements essentially force us toward on-premises solutions."*
- *Likelihood:* High (if using commercial cloud LLMs)
- *Impact:* Critical (GDPR violations, regulatory sanctions)
- *Mitigation:* On-premises deployment of open-source models (Llama, DistilBERT), data residency agreements with vendors if cloud necessary, encryption in transit

5.1.3 *Operational Risks*

Risk O1: Integration Failures Creating Security Gaps

- *Description:* Weak points in API connections to customer databases, CRM, core banking systems
- *Stakeholder Source:* P13 (ENG): *"System integration is where security often breaks down. End-to-end architecture is essential."*
- *Likelihood:* Medium (complex integration landscape)
- *Impact:* High (potential data exposure through integration points)
- *Mitigation:* Security-first integration design, middleware with authentication/authorization/validation, API gateway with centralized security controls, penetration testing of integration points

Risk O2: Insufficient Ongoing Maintenance Leading to Drift

- *Description:* Model performance degradation, emerging vulnerabilities, regulatory changes not addressed
- *Stakeholder Source:* P11 (ENG): *"Model drift is inevitable as query distributions shift. Without quarterly retraining, performance degrades."*
- *Likelihood:* High (without dedicated maintenance)
- *Impact:* Medium (gradual degradation rather than catastrophic failure)

- *Mitigation:* Dedicated maintenance budget (20-30% annual), scheduled retraining cycles, continuous monitoring with performance baselines, regulatory change tracking

Risk O3: Insider Threats Exploiting System Knowledge

- *Description:* Malicious or negligent employees systematically extracting data through legitimate-appearing queries
- *Stakeholder Source:* P15 (ENG): *"A sophisticated employee could craft queries that appear benign but systematically extract information over time."*
- *Likelihood:* Low (with proper controls)
- *Impact:* Critical (systematic data theft)
- *Mitigation:* Behavioural analytics detecting unusual query patterns, role-based access controls, audit log analysis with anomaly detection, mandatory security awareness training

5.1.4 ***Reputational Risks***

Risk REP1: Public Data Breach Eroding Customer Trust

- *Description:* High-profile leakage incident receiving media attention, damaging brand perception
- *Stakeholder Source:* P07 (RISK): *"Data leakage sits in Tier 3 riskit affects not just our institution but public perception of AI in banking generally."*
- *Likelihood:* Low (with proper security) to Medium (with gaps)
- *Impact:* Critical (customer churn, competitive disadvantage, industry-wide AI scepticism)
- *Mitigation:* Proactive security measures, transparent incident response, crisis communication planning, customer notification protocols, remediation offerings

Risk REP2: Bias or Discrimination Incidents Going Viral

- *Description:* Discriminatory AI behaviour captured and shared on social media, creating backlash
- *Stakeholder Source:* P08 (RISK): *"One viral post showing discriminatory behaviour can devastate brand trust. Prevention is far cheaper than remediation."*
- *Likelihood:* Medium (bias difficult to eliminate entirely)

- *Impact:* High (reputational damage, regulatory investigations)
- *Mitigation:* Comprehensive bias testing pre-deployment, ongoing fairness monitoring, diverse training data, rapid incident response protocols, transparency about AI limitations

5.1.5 *Financial Risks*

Risk F1: Infrastructure Costs Exceeding Budget Projections

- *Description:* GPU requirements, operational costs (electricity, cooling, maintenance) higher than estimated
- *Stakeholder Source:* P14 (ENG): *"At current prices, scaling to production could require €1-2 million capital expenditure plus ongoing operational costs potentially 100% of initial development annually."*
- *Likelihood:* Medium (cost estimation uncertainties)
- *Impact:* Medium (budget overruns affecting ROI)
- *Mitigation:* Detailed cost modelling with 20% contingency, phased deployment limiting initial infrastructure investment, usage-based scaling, continuous cost optimization

Risk F2: Negative ROI Due to Insufficient Adoption

- *Description:* Customers or staff refusing to use system, failing to realize expected efficiency gains
- *Stakeholder Source:* P06 (RISK): *"Do savings in personnel costs offset infrastructure investments? What's the risk-adjusted ROI?"*
- *Likelihood:* Medium (depends on change management effectiveness)
- *Impact:* High (failed investment)
- *Mitigation:* Comprehensive training programs, phased adoption with early wins demonstrating value, user feedback loops informing improvements, realistic adoption projections

5.1.6 *Strategic Risks*

Risk S1: Competitive Disadvantage if Deployment Delayed

- *Description:* Competitors deploying AI chatbots first, capturing market share and customer expectations

- *Stakeholder Source:* P11 (ENG): *"We risk falling behind competitors if we're too cautious. Controlled experimentation is how we learn."*
- *Likelihood:* Medium (competitive AI adoption increasing)
- *Impact:* Medium (gradual market share erosion)
- *Mitigation:* Balanced risk-taking approach, phased deployment enabling learning while managing risk, competitive intelligence tracking rival AI capabilities

Risk S2: Regulatory Changes Rendering System Non-Compliant

- *Description:* EU AI Act implementation, GDPR case law evolution requiring system modifications
- *Stakeholder Source:* P02 (REG): *"Regulations evolve. The EU AI Act phases in through 2027. We need governance ensuring ongoing compliance."*
- *Likelihood:* High (regulatory evolution certain)
- *Impact:* Medium (requires ongoing adaptation but not catastrophic)
- *Mitigation:* Modular architecture enabling component updates, regulatory monitoring processes, legal consultation on interpretation changes, budget for compliance modifications

5.2 Risk Prioritization Matrix

Risks were prioritized using Impact × Likelihood scoring:

Critical Priority (Address Immediately):

- T1: Adversarial Prompt Injection (High × Critical)
- R1: GDPR Penalties (Medium × Critical)
- O3: Insider Threats (Low × Critical)
- REP1: Public Data Breach (Medium × Critical)

High Priority (Address Before Deployment):

- T2: Model Hallucination (Medium × High)
- T3: Performance Degradation (Medium × High)
- R2: EU AI Act Non-Compliance (Medium × High)
- R3: Cross-Border Data Transfer (High × Critical - but easily mitigated via architecture choice)
- O1: Integration Failures (Medium × High)

- REP2: Bias Incidents (Medium × High)
- F2: Negative ROI (Medium × High)

Medium Priority (Monitor and Mitigate):

- T4: Third-Party Dependencies (Medium × Variable)
- O2: Insufficient Maintenance (High × Medium)
- F1: Infrastructure Costs (Medium × Medium)
- S1: Competitive Disadvantage (Medium × Medium)
- S2: Regulatory Changes (High × Medium)

Section 6. **STAKEHOLDER RECOMMENDATIONS AND DESIGN PREFERENCES**

6.1 **Architectural Preferences**

Stakeholders expressed clear preferences regarding system architecture, informed by their functional perspectives:

Deployment Model (15/15 consensus on on-premises):

P09 (RISK): *"Given GDPR Article 44 constraints and data sovereignty concerns, cloud-based solutions using OpenAI or Anthropic APIs are non-starters. We need on-premises deployment with full data control."*

P02 (REG): *"On-premises also simplifies compliance auditing. When regulators ask 'where is customer data processed and stored?', we can answer definitively rather than navigating complex vendor agreements."*

P14 (ENG): *"On-premises means higher upfront capital expenditure but lower ongoing licensing costs and complete infrastructure control. It's the right choice for regulated financial services."*

Model Selection (Strong preference for open-source):

P11 (ENG): *"Open-source models like Llama 3.1 offer transparency we can audit the architecture, fine-tune for our specific needs, and aren't locked into vendor roadmaps. Proprietary models are black boxes."*

P05 (REG): *"From a legal perspective, open-source licenses (MIT, Apache 2.0) are clearer regarding liability and usage rights compared to proprietary API terms that vendors can change unilaterally."*

P12 (ENG): *"The trade-off is that open-source models require more technical expertise to deploy and maintain, but that's acceptable given the control and transparency benefits."*

RAG Architecture (14/15 favored hybrid retrieval):

P13 (ENG): *"Pure vector search misses exact matches like regulatory article numbers. Pure keyword search misses semantic intent. Hybrid approaches combining both deliver comprehensive coverage."*

P03 (REG): *"For regulatory queries, exact citation matching is critical. If someone asks about GDPR Article 15, we need to retrieve that specific article, not semantically similar content from Article 14."*

P12 (ENG): *"Empirically, hybrid search improves recall by 7-15 percentage points in our testing. That translates to fewer hallucinations and more grounded responses."*

6.2 Feature Prioritization

Stakeholders ranked features by importance using a 5-point scale (1=Low, 5=Critical). Top-ranked features:

1. **Real-time PII Detection and Masking (Mean: 4.9/5.0)**
   - P01 (REG): *"This is non-negotiable. Without robust PII protection, we can't deploy."*
2. **Source Citation for Regulatory Queries (Mean: 4.8/5.0)**
   - P02 (REG): *"Transparency requirement under EU AI Act Article 13. Users must understand information sources."*
3. **Confidence-Based Human Escalation (Mean: 4.7/5.0)**
   - P04 (REG): *"Automated advice is acceptable for straightforward queries, but complex scenarios require human judgment."*
4. **Comprehensive Audit Logging (Mean: 4.7/5.0)**
   - P08 (RISK): *"For incident investigation and regulatory reporting, complete audit trails are essential."*
5. **Multi-Factor Authentication for Sensitive Queries (Mean: 4.5/5.0)**
   - P15 (ENG): *"Balance security and user experience: simple queries proceed with basic auth, sensitive queries require MFA."*
6. **Performance Monitoring Dashboard (Mean: 4.4/5.0)**
   - P06 (RISK): *"Real-time visibility into system health, query volumes, leakage incidents, and performance metrics."*
7. **Multilingual Support (Nordic Languages) (Mean: 4.2/5.0)**
   - P02 (REG): *"Market requirement and regulatory consideration. English-only creates accessibility barriers."*
8. **Bias Detection and Fairness Monitoring (Mean: 4.0/5.0)**
   - P01 (REG): *"EU AI Act compliance and ethical imperative to prevent discrimination."*
9. **Multi-Turn Conversation Context (Mean: 3.9/5.0)**
   - P12 (ENG): *"Improves user experience but introduces security considerations around context accumulation."*

10. Voice Interface Support (Mean: 3.2/5.0)

- P13 (ENG): *"Desirable long-term but adds complexity (speech recognition, real-time processing). Text-first is pragmatic."*

6.3 Success Criteria

Stakeholders collaboratively defined success criteria for deployment readiness:

Technical Criteria:

- Data leakage rate <2% on adversarial test set (1,000+ queries)
- F1 score >0.90 on financial query benchmark
- Median latency <1.5s, 95th percentile <2.0s
- 99.5% uptime during business hours (measured monthly)
- Zero Tier 3 leakage incidents (complete PII disclosure) in testing

Compliance Criteria:

- GDPR compliance validated by Data Protection Officer
- EU AI Act conformity assessment completed
- External security audit with no critical findings
- Legal review of liability and contractual arrangements
- Regulatory notification procedures documented

User Acceptance Criteria:

- Staff confidence in system security >4.0/5.0 (n>30 survey respondents)
- Customer satisfaction >4.0/5.0 for pilot participants (n>100)
- <5% query escalation rate to human agents (indicating chatbot adequacy)
- Positive net promoter score (NPS) among pilot users

Operational Criteria:

- Integration testing with all critical systems (AD, CRM, core banking) completed successfully
- Staff training program delivered to >90% of customer service team
- Incident response procedures tested through tabletop exercises
- Business continuity plan validated including failover scenarios
- Cost projections validated against actual Phase 1 expenditures (±20% tolerance)

P06 (RISK): *"Success isn't just technical performance it's demonstrating that the system is operationally sustainable, economically viable, legally compliant, and trusted by stakeholders. All criteria must be met before progressing beyond internal pilot."*

Section 7. METHODOLOGY INSIGHTS AND REFLEXIVE ANALYSIS

7.1 Interview Process Effectiveness

The semi-structured format proved effective in eliciting rich, detailed insights while maintaining consistency across 15 interviews.

Strengths:

- Role-specific probing (Developer/Customer-Facing/Compliance perspectives) surfaced nuanced requirements that generic questions would have missed
- The 45-minute duration balanced depth with participant fatigue three participants noted appreciating the time constraint
- Context briefing with role-tailored analogies facilitated non-technical participant understanding (particularly effective for P01, P03, P04 from REG)
- Follow-up probing ("Can you give an example?" "Why is that a concern?") encouraged elaboration beyond surface responses

P03 (REG): *"I appreciated the structured but conversational approach. You explained technical concepts clearly without being condescending, and the questions felt relevant to my compliance responsibilities."*

Challenges:

- Two participants (P05-REG, P10-RISK) had limited AI familiarity, requiring extended context briefing (10 minutes vs. planned 5), slightly reducing question time
- One participant (P14-ENG) provided highly technical responses requiring post-interview clarification of terminology for cross-functional synthesis
- Zoom audio quality issues affected one interview (P09-RISK), requiring follow-up email clarification on three responses

7.2 Researcher Reflexivity

As both researcher and system developer, I brought certain biases and perspectives requiring acknowledgment:

Potential Biases:

- Technical background may have emphasized engineering perspectives over equally important social/organizational dimensions
- Investment in demonstrating dissertation contribution could bias interpretation toward favourable findings
- Employment relationship with NordicBank potentially constrained critical findings affecting institutional interests

Mitigation Strategies:

- Member checking: Shared anonymized interview summaries with participants (12/15 responded, confirming accuracy)
- Peer debriefing: Discussed emerging themes with dissertation supervisor and university ethics representative
- Reflexive journaling: Maintained research diary documenting assumption awareness and interpretation evolution
- Triangulation: Cross-validated interview findings with survey data (n=77) and workshop outputs (n=30 total across 3 workshops)

Reflexive Observations:

- My technical lens initially undervalued regulatory complexity compliance requirements proved more intricate than anticipated
- Stakeholder risk aversion exceeded my assumptions; what I perceived as acceptable risk (1-2% leakage) stakeholders viewed with significant concern
- The phased implementation consensus emerged organically from participants rather than my prompting, suggesting genuine institutional wisdom rather than researcher-induced framing

7.3 7.3 Ethical Considerations

All interviews adhered to University of Essex ethics protocols (Reference: ETH2024-0892):

Informed Consent:

- Verbal consent obtained after briefing (recorded on audio)
- Written consent forms completed for 13/15 participants (2 preferred verbal only, which was accommodated)
- Explicit right-to-withdraw confirmation before each interview

Anonymisation:

- Participant identities replaced with codes (P01-P15) in all documentation
- Organizational details anonymized (NordicBank pseudonym)
- Potentially identifying information (specific product names, internal project codes) redacted from quotes
- Audio recordings stored with AES-256 encryption, accessible only to principal investigator
- Transcripts anonymized before NVivo analysis
- All data scheduled for secure deletion 6 months' post-dissertation submission

Data Minimisation:

- No personal identifiers collected beyond functional role and experience level
- Participant contact information used solely for member checking, not retained long-term
- Interview focus on professional opinions rather than personal information

Participant Wellbeing:

- No distress observed during interviews
- One participant (P08-RISK) expressed concern about job security related to AI automation; researcher clarified research scope excludes employment impact assessment
- Post-interview debriefing offered to all participants; none requested

Section 8. SECTION 8: INTEGRATION WITH DESIGN THINKING STAGES

8.1 Empathise Stage Outcomes

The interviews successfully completed the Empathise stage objectives:

Deep Understanding Achieved:

- 127 discrete stakeholder requirements identified across security, performance, accuracy, transparency, fairness, usability
- 43 distinct risks articulated spanning technical, regulatory, operational, reputational, financial, strategic domains
- 5 primary themes validated through thematic analysis with 87% inter-coder reliability

Stakeholder Perspectives Captured:

- Regulatory focus: Compliance requirements, legal interpretation, liability concerns
- Risk focus: Threat assessment, incident impact, mitigation strategies
- Engineering focus: Technical feasibility, implementation complexity, architectural trade-offs

Empathy Development:

- Researcher gained visceral understanding of stakeholder concerns beyond abstract requirements
- Recognition that 1% leakage, statistically impressive, translates to unacceptable absolute risk in stakeholder mental models
- Appreciation for institutional risk aversion rooted in regulatory reality rather than unnecessary caution

8.2 Transition to Define Stage

Interview insights directly inform Define stage activities:

Problem Statement Refinement:

Initial (Pre-Interview): "Design a finance chatbot that prevents data leakage while maintaining usability."

Refined (Post-Interview): "Design a finance chatbot that achieves <2% data leakage with zero Tier 3 incidents, maintains <1.5s median latency with transparent security explanations, provides >90% accuracy with confidence-based human escalation, ensures multilingual fairness, and demonstrates operational sustainability validated through phased deployment satisfying regulatory requirements and earning stakeholder trust."

The refined statement incorporates:

- Specific, measurable targets derived from stakeholder priorities
- Contextual nuances (Tier 3 zero tolerance, latency acceptance with transparency)
- Multi-dimensional success (technical + regulatory + stakeholder acceptance)
- Implementation realism (phased deployment reflecting risk management)

How Might We (HMW) Statements:

Extracted from interview insights for Ideate stage exploration:

1. HMW prevent data leakage while maintaining conversational naturalness?
   - *Source:* Tension between security (strict filtering) and usability (fluid interaction)
2. HMW distinguish educational leakage from genuine security failures?
   - *Source:* P03's observation about format-preserving examples serving legitimate purposes
3. HMW provide transparency that builds trust without overwhelming users?
   - *Source:* Need for explainability balanced against information overload concerns
4. HMW implement multilingual support without sacrificing accuracy or introducing bias?
   - *Source:* Market requirement (P02) conflicting with technical complexity (P12)
5. HMW achieve acceptable latency while maintaining robust security checks?
   - *Source:* Performance-security trade-off discussion, stakeholder acceptance of modest delays with transparency
6. HMW scale infrastructure sustainably within realistic budget constraints?
   - *Source:* P14's infrastructure cost concerns balanced against P06's ROI requirements
7. HMW ensure ongoing maintenance without unsustainable operational burden?
   - *Source:* P11's model drift concerns and P02's regulatory evolution observations
8. HMW integrate with existing systems without creating security vulnerabilities?

- *Source:* P13's integration complexity and P15's API security concerns

9. HMW implement phased rollout that enables learning while managing risk?
    - *Source:* Universal stakeholder preference for gradual deployment
10. HMW balance innovation with institutional risk aversion in regulated environment?
    - *Source:* Divergent perspectives (P11-ENG pro-innovation, P01-REG pro-caution, P06-RISK context-dependent)

8.3 Ideate Stage Preparation

Interview findings provide concrete starting points for ideation workshops:

Validated Solution Concepts:

- Tiered incident classification (educational/partial/complete leakage)
- Confidence-based human escalation (<85% threshold)
- Role-adapted explanation interfaces (customer/staff/engineering)
- Phased rollout structure (internal/limited external/full production)
- Multi-layered security architecture (pre/during/post processing)
- Hybrid RAG retrieval (vector + keyword combining strengths)

Open Questions for Ideation:

- How to technically implement tiered incident classification? (detection algorithms, automated categorization)
- What confidence scoring methodology optimally balances automation efficiency with review necessity?
- How to design role-adapted interfaces sharing underlying data but presenting appropriately?
- What specific success gates should trigger progression between rollout phases?
- How to sequence defensive layers for optimal security-performance balance?
- What retrieval weighting (vector vs. keyword) optimizes different query types?

Section 9. CONCLUSIONS AND RECOMMENDATIONS

9.1 Key Findings Summary

This interview study with 15 NordicBank stakeholders across regulatory, risk, and engineering functions yielded five primary findings:

1. Security Dominance with Contextual Nuance: Data leakage concerns dominated stakeholder priorities (94% of transcripts), but with important qualification: stakeholders distinguished contextual/educational leakage from genuine security failures, suggesting regulatory frameworks should implement tiered incident classification rather than binary approaches.

2. Transparency as Trust Mechanism: Visible security mechanisms (real-time PII detection indicators, source citation, audit trails) emerged as critical trust factors more important than raw performance metrics. This validates EU AI Act transparency requirements and suggests explainability should be prioritized equally with technical capabilities.

3. Phased Implementation Imperative: Universal stakeholder consensus (15/15) on gradual rollout over immediate deployment reflects institutional risk management wisdom. The three-phase structure (internal pilot → limited external → full production) balances innovation opportunity with risk containment.

4. Operational Sustainability as Deployment Prerequisite: Technical proof-of-concept success proves necessary but insufficient. Integration complexity (6–12 months beyond development), infrastructure costs (€1-2M capital + 20-30% annual operational), and ongoing maintenance commitments emerged as equally important considerations determining deployment viability.

5. Performance-Security Trade-off Recalibration: Contrary to initial assumptions requiring sub-second responses, stakeholders accepted 1.0–1.5s latency when accompanied by transparent security explanations. This suggests rigid performance targets may be unnecessarily stringent if transparency compensates for modest delays.

9.2 Implications for System Design

Architectural Implications:

- Implement tiered incident classification system (Level 1/2/3) with graduated responses

- Deploy multi-layered security (pre-processing detection, RAG boundaries, structured generation, post-validation)
- Adopt hybrid RAG retrieval (FAISS + BM25) addressing both semantic and exact-match requirements
- Select on-premises open-source models (Llama 3.1 8B, DistilBERT) ensuring data residency compliance
- Design role-adapted interfaces (customer/staff/engineering) serving different explanation needs from shared data

Operational Implications:

- Structure phased rollout with clear success gates: internal pilot (6 months) → limited external (6 months) → full production (conditional)
- Budget 20-30% of initial development costs annually for sustainable maintenance
- Allocate 6-12 months for integration engineering beyond chatbot development
- Establish dedicated teams for ongoing responsibilities: monthly adversarial testing, quarterly retraining, annual security audits

Governance Implications:

- Create cross-functional governance committee (REG/RISK/ENG representation) for deployment decisions
- Implement confidence-based escalation protocols (<85% threshold mandating human review)
- Establish tiered incident response procedures matching leakage severity classification
- Develop comprehensive training programs for staff (technical understanding + escalation protocols)

9.3 Recommendations for Future Research

Methodological Recommendations:

- Expand stakeholder diversity to include customer-facing staff, senior executives, and external auditors
- Conduct follow-up interviews post-deployment to validate whether anticipated concerns materialized

- Implement longitudinal study tracking stakeholder perspective evolution over 12–24 months

Research Questions Emerging:

- How do tiered incident classification systems perform in production environments? Do stakeholders maintain distinction between educational and genuine leakage?
- What confidence threshold optimally balances automation efficiency with human review necessity across different query types?
- How does stakeholder trust evolve over time, does initial scepticism diminish with positive experience, or does minor incident occurrence disproportionately erode confidence?
- Do multilingual implementations exhibit systematic bias requiring dedicated fairness interventions?

Practical Research Needs:

Cost-benefit analysis quantifying ROI of finance chatbot deployment vs. traditional customer service models

- Comparative study of on-premises vs. cloud-based deployments assessing security, cost, and performance trade-offs in regulated environments
- Evaluation of different explainability approaches measuring which transparency mechanisms most effectively build stakeholder trust
- Cross-institutional study examining whether NordicBank findings generalize to other financial institutions with different organizational cultures, risk tolerances, and regulatory environments

9.4 Limitations of Interview Study

Sample Limitations:

- Single institution (NordicBank) limits generalizsbility to broader financial services sector
- Functional distribution (5 REG, 5 RISK, 5 ENG) may not reflect actual organizational composition where engineering resources typically outnumber compliance staff
- Absence of customer perspectives, all participants were employees, potentially introducing institutional bias underrepresenting end-user concerns

- Geographic concentration (Nordic region) limits insights regarding global regulatory complexity

Methodological Limitations:

- Semi-structured format balances consistency with flexibility but may miss emergent themes that fully unstructured interviews would capture
- 45-minute duration enables depth but may leave some topics underexplored
- Virtual interviews (Zoom) lack nonverbal cues available in face-to-face settings
- Self-reported opinions may differ from actual behaviour post-deployment

Researcher Limitations:

- Technical background potentially emphasizing engineering perspectives over organizational/social dimensions
- Dual role as researcher and system developer introducing potential confirmation bias
- Employment relationship with NordicBank potentially constraining critical findings
- Interviewer effects, participants may have provided socially desirable responses rather than genuine opinions

Temporal Limitations:

- Interviews conducted pre-deployment capture anticipated rather than experienced concerns
- Regulatory landscape evolving EU AI Act implementation timeline means current interpretations may shift
- Technology advancement, LLM capabilities improving rapidly, potentially dating findings

9.5 Final Recommendations for Dissertation Research

Based on interview insights, the following recommendations guide subsequent dissertation phases:

For Prototype Development (Chapter 4-5):

1. **Prioritize Multi-Layered Security Architecture:**
   - Implement three-stage pipeline (pre-processing, in-flight, post-validation) as unanimously emphasized by stakeholders

- Focus engineering effort on PII detection robustness (98%+ recall target) given its critical importance
- Develop tiered incident classification enabling contextual risk assessment

2. **Integrate Transparency Mechanisms from Inception:**
   - Real-time PII detection indicators visible to users
   - Source citation for all regulatory/compliance responses
   - Comprehensive audit logging supporting regulatory reporting
   - Role-adapted explanation interfaces serving different stakeholder needs
3. **Design for Operational Sustainability:**
   - Modular architecture enabling component updates without system-wide redeployment
   - Automated testing frameworks reducing manual QA burden
   - Monitoring instrumentation providing real-time operational visibility
   - Documentation supporting knowledge transfer and maintenance continuity
4. **Implement Confidence-Based Escalation:**
   - Confidence scoring algorithm validated against ground truth
   - <85% threshold triggering human review based on stakeholder consensus
   - Clear escalation workflows integrated with existing customer service systems

For Evaluation (Chapter 6):

1. **Employ Stakeholder-Derived Success Criteria:**
   - Technical: <2% leakage, F1>0.90, latency <1.5s median/<2.0s 95th percentile
   - Compliance: GDPR/EU AI Act conformity validated by legal review
   - Acceptance: >4.0/5.0 satisfaction from staff and pilot customers
   - Operational: Integration testing completed, training delivered, costs within projections
2. **Conduct Adversarial Testing Reflecting Real Threats:**
   - Prompt injection attempts (direct commands, role confusion, delimiter manipulation)
   - Social engineering scenarios (authority impersonation, urgency tactics)
   - Multi-turn context manipulation (gradual premise establishment)
   - Edge-case exploitation (Unicode homoglyphs, encoding variations, format abnormalities)
3. **Validate Through Multi-Method Triangulation:**

- Quantitative performance metrics (leakage rate, F1 score, latency)
- Qualitative stakeholder feedback (post-pilot interviews, satisfaction surveys)
- Expert evaluation (security audit findings, regulatory assessment)

For Discussion (Chapter):

1. **Address Stakeholder Concerns Explicitly:**
   - Discuss how absolute risk (10,000 annual incidents at 1% rate) was considered in design decisions
   - Explain how tiered incident classification addresses contextual leakage concerns
   - Analyse whether transparency mechanisms successfully built trust as anticipated
   - Assess whether phased implementation approach proved viable
2. **Acknowledge and Reconcile Divergent Perspectives:**
   - Risk tolerance differences (REG conservative, RISK balanced, ENG pragmatic)
   - Innovation vs. caution tension requiring governance structures balancing competing priorities
   - Explainability depth preferences requiring role-adapted solutions
3. **Provide Actionable Guidance:**
   - Infrastructure requirements based on empirical resource utilization
   - Cost projections validated against actual Phase 1 expenditures
   - Integration timeline reflecting observed complexity
   - Maintenance requirements derived from operational experience

9.6 Concluding Remarks

This interview study successfully completed the Empathise stage of the Design Thinking framework, capturing rich, detailed stakeholder perspectives across regulatory, risk, and engineering functions. The findings fundamentally shaped system design priorities, revealing that:

- **Security concerns dominate** but require contextual sophistication beyond binary classification
- **Transparency mechanisms** prove equally important as technical capabilities for building trust

- **Operational sustainability** determines deployment viability as much as technical proof-of-concept success
- **Phased implementation** reflects institutional wisdom managing innovation risk in regulated environments
- **Performance-security trade-offs** prove more flexible than initially assumed when transparency compensates for modest delays

The 127 requirements, 43 risks, and 23 high-level design implications extracted from interviews provide concrete foundation for subsequent Define, Ideate, Prototype, and Test stages. By grounding technical development in empirical stakeholder insights rather than researcher assumptions, the research enhances practical relevance and deployment viability.

The interview methodology proved effective, with semi-structured format, role-specific probing, and 45-minute duration balancing depth, consistency, and participant engagement. Limitations, single institution, employee-only perspectives, researcher positionality warrant acknowledgment but do not invalidate findings given the study's proof-of-concept rather than universal generalization objective.

Moving forward, these insights inform a system architecture that doesn't merely achieve technical benchmarks but addresses real stakeholder priorities, respects institutional constraints, and increases likelihood of successful deployment in the highly regulated, risk-averse financial services environment. This human-cantered approach, integrating stakeholder wisdom with technical capability, represents the Design Thinking methodology's core value: ensuring solutions fit the problems as stakeholders experience them, not as researchers imagine them.

Appendixes

Appendix C1: Participant Demographic Summary

Code	Function	Specific Role	Experience (years)	Primary Expertise
P01	REG	Senior Compliance Officer	12	GDPR, data protection
P02	REG	Data Protection Officer	9	EU AI Act, regulatory affairs
P03	REG	Regulatory Affairs Manager	7	MiFID II, financial regulation
P04	REG	Compliance Analyst	5	AML/KYC compliance
P05	REG	Legal Counsel	9	Technology contracts, liability
P06	RISK	Chief Risk Officer	15	Enterprise risk management
P07	RISK	Operational Risk Manager	13	AI oversight, operational resilience
P08	RISK	Cybersecurity Analyst	8	Threat assessment, InfoSec
P09	RISK	Third-Party Risk Manager	10	Vendor security, outsourcing
P10	RISK	Business Continuity Manager	10	Disaster recovery, resilience
P11	ENG	Lead AI Engineer	8	ML systems, model deployment
P12	ENG	Data Scientist	5	NLP, text analytics
P13	ENG	Systems Architect	9	Integration, enterprise architecture
P14	ENG	DevOps Engineer	6	Infrastructure, deployment
P15	ENG	Security Engineer	6	Application security, pentesting