AUTOSUMM: A Comprehensive Framework for LLM-Based Conversation Summarization

Abhinav Gupta [*], **Devendra Singh** [*], **Greig Cowan**[*], **N Kadhiresan**[*], **Siddharth Srivastava**[*], **Sriraja Yagneswaran**[*], **Yoages Kumar Mantri**[*]
Data Science and Innovation
NatWest Group
{abhinav.gupta1, greig.cowan, Yoageskumar.Mantri, Kadhiresan.N, Siddharth.Srivastava, Yagneswaran.Sriraja}@natwest.com,
Devendra.Singh2@natwestmarkets.com

Abstract

We present AUTOSUMM, a large language model (LLM)-based summarization system deployed in a regulated banking environment to generate accurate, privacy-compliant summaries of customer-advisor conversations. The system addresses challenges unique to this domain, including speaker attribution errors, hallucination risks, and short or low-information transcripts. Our architecture integrates dynamic transcript segmentation, thematic coverage tracking, and a domain specific multi-layered hallucination detection module that combines syntactic, semantic, and entailment-based checks. Human-in-the-loop feedback from over 300 advisors supports continuous refinement and auditability.

Empirically, AUTOSUMM achieves a 94% factual consistency rate and a significant reduction in hallucination rate. In production, 89% of summaries required no edits, and only 1% required major corrections. A structured model version management pipeline ensures stable upgrades with minimal disruption. We detail our deployment methodology, monitoring strategy, and ethical safeguards, showing how LLMs can be reliably integrated into high-stakes, regulated workflows.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing and have spawned numerous applications. Organizations are racing to incorporate LLMs into their use cases seeking to replicate the success and deliver transformational value. However, deployment of LLM-based solutions presents distinct challenges that are further exacerbated in regulated environments, such as banks, where customer and data privacy need to be sanctimoniously preserved, and standards of ethics and governance have to be met. Domain adaptation, monitoring, detecting hallucinations and explainability are additional challenges that need to be considered. (Meskó and Topol, 2023, Wu et al., 2024, Mökander et al., 2024, Zhao et al., 2024, Wan et al., 2024, Kim et al., 2023, Das et al., 2024)

Using LLMs to summarize larger volumes of textual information into a smaller, more manageable size is useful for many business applications and processes. In the financial domain, research and work on summarization have focused on documents such as earnings reports, product and process descriptions. The dynamic nature of human conversations coupled with domain-specific context poses challenges in summarizing conversations and extracting key information. Recent studies on financial text summarization (Mukherjee et al., 2022, Khatuya et al., 2024) highlight progress, but further innovation is needed to meet regulatory and accuracy requirements.

We showcase AUTOSUMM, a successfully deployed end-to-end solution that uses an LLM to summarize various customer conversations that occur during a customer's banking relationship journey. Our solution features a robust summarization capability designed to meet bespoke requirements and support a wide range of conversation scenarios such as meetings and multi-participant dialogues. The solution ecosystem includes modules that monitor and manage changes to data and model performance to drive operational efficiency. [*]

A key insight from our implementation is the importance of maintaining balance across contrasting parameters. Incorporating the right level of human oversight with automation is one such instance. AUTOSUMM leverages LLMs for initial summaries while incorporating systematic feedback from users to mitigate risks.

The primary contribution of this work is a practical, end-to-end solution that serves as a template for deploying LLM-based systems at scale within regulated enterprise environments. In addition, we

[*] The authors' names are listed alphabetically, with all contributing equally.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 500–509
July 28-30, 2025 ©2025 Association for Computational Linguistics

present custom techniques for data quality assessment, model and data monitoring, and hallucination detection, all tailored to incorporate domain-specific constraints and operational realities. Our findings also offer empirical insights from large-scale production deployment in the banking sector, highlighting both technical challenges and process-level learnings. Collectively, these contributions demonstrate that with appropriate guardrails, LLM-based summarization can be reliably and compliantly integrated into high-stakes, regulated workflows.

2 Solution Overview

AUTOSUMM has been designed and developed for summarizing customer-advisor conversations to improve operational efficiency, cost-reduction, and enhance customer engagement. Figure 1 illustrates the overall architecture of the solution that comprises of Core Summarization module along with other modules to support operational and governance requirements. The AUTOSUMM solution is currently deployed for one business area, processing approximately 12,000 customer conversations/month and delivering summaries to over 300 advisors. The generated summaries maintain consistency across advisors and have removed the necessity for them to manually compose follow-up call notes, resulting in an average time savings of 15 minutes per call.

Audio recorded from phone calls and meetings between customers and banking advisors is transcribed by an off-the-shelf transcription service. The Core Summarization module processes the transcripts and generates summaries, which are then presented to the advisors for review. Summaries and transcripts are saved to the Customer Relationship Management (CRM) system, establishing a record of each client conversation.

2.1 Data Specification and User Requirements

Our solution processes diverse call transcriptions, varying in length, noise, ambiguity, and type. These include short internal calls and lengthy annual reviews, in addition to routine check-ins and portfolio transfers. To ensure consistent output, a standardized summary format is utilized. To gain a deeper understanding of business needs within specific areas and summary format, a workshop was held with relationship managers (advisors) to identify the most critical thematic extracts. From this session, six key themes were identified, as shown in Table 1: actions, hard facts, soft facts, queries and status, financial objectives, and portfolio positions. The LLM prompt is specifically tailored to focus on these six themes during the transcript summarization process. The summaries must be concise and presented in bullet points, emphasizing the two most important actions and covering all six key themes (an example summary is included in Appendix A.1). Furthermore, summaries must be delivered within 60 minutes, adhere to word limits based on call duration (as outlined in Table 2), and accurately reflect numeric and financial details.

Table 1: Summary Themes and Descriptions

Themes	**Description**
Soft Facts	Insights into the client's broader personal situation and evolving financial needs
Actions	Client requests and the commitments made by the agent in response
Financial Objectives	Specific financial goals and aspirations set forth by the client
Portfolio Position	Client's current investments and relevant financial factors
Queries and Status	Client's questions and the responses provided by the advisor
Hard Facts	Client's risk tolerance and strategic financial planning

Table 2: Call Duration Statistics

Call Duration (in minutes)	**Expected summary length (in words)**	**Approx. input transcript length (in words)**	**Distribution of call volume (in %)**
(0,15]	100-150	1100	82
(15,30]	250-500	3500	14
(30,60]	500-600	6600	3
> 60	600-750	12000	1

2.2 Core Summarization Module

The primary function of this module is to generate summaries for the conversation transcripts. The summarization is triggered every 15 minutes and all available transcripts in the queue are processed in a batch. The transcripts undergo content moderation checks and pre-processing prior to the summarization model. Pre-processing involves analysis to determine if a transcript needs to be broken into smaller chunks as well as anonymizing names of customers and agents. Summarization is performed using an LLM with prompts designed to meet the format and content requirements.

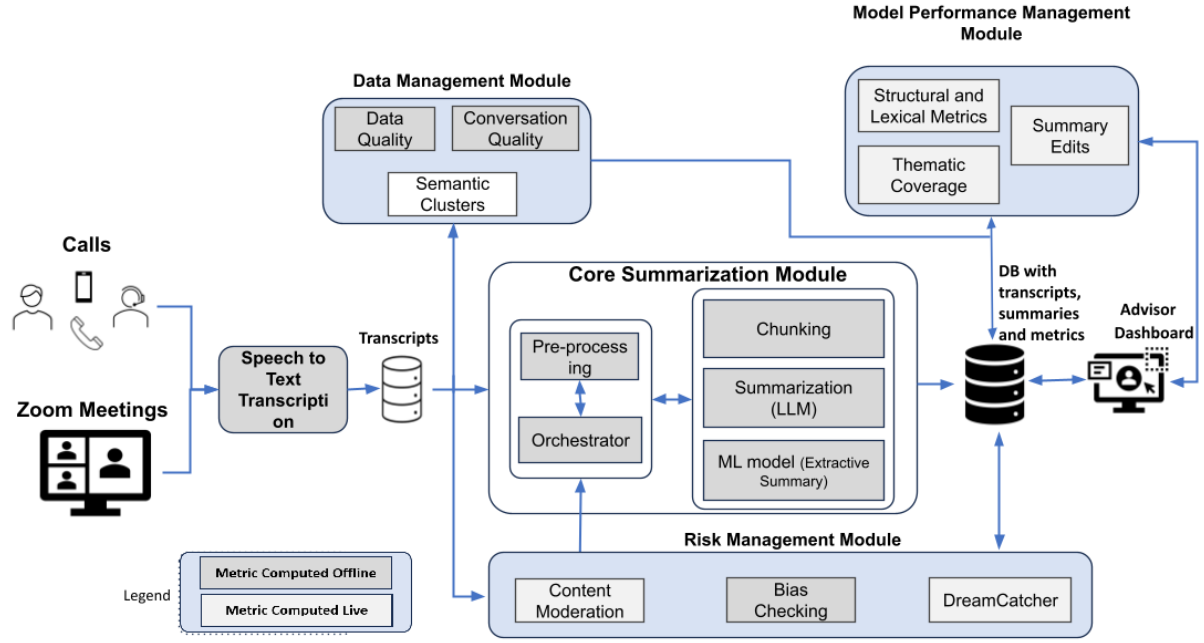

Figure 1: AUTOSUMM System architecture comprising of four modules: Core Summarization, Data Management, Model Performance Management, and Risk Management

Model Selection

Choosing an appropriate model required balancing cost-per-million-tokens with performance metrics such as hallucination rates, toxicity, fairness, and confidentiality safeguards. Initial decisions referenced publicly available benchmarks (Hughes, 2025) and experimentation evaluations. Based on these, OpenAI's GPT-3.5 Turbo (16K) (OpenAI, 2023) was chosen for initial deployment. Open-source models were not considered for deployment due to concerns about long-term support for hosting and maintenance.

Prompt Design

We use a *Chain of Thoughts (CoT)* (Wei et al., 2022) style prompt design to guide the model in a step-by-step reasoning process. This ensures the summarization is thorough and logically consistent, as opposed to simply truncating or paraphrasing large segments of text in a single pass. This approach is especially beneficial for capturing and key themes and prioritizing them in the summary. Besides, the LLM is also instructed to disambiguate and correct speaker labels, typically tagging them as "customer" or "advisor," minimizing attribution errors in the final summary.

ML Model - Extractive Summary

In cases where the LLM API is unavailable, an extractive fallback mechanism employs a fine-tuned BERT (Wolf et al., 2021) classifier to assign utterances to the desired themes. This ensures continuity of service and reliable summary generation.

Chunking

While large context windows generally suffice, certain transcripts exceed practical limits—especially when multiple conversations are consolidated or when discussions shift across departments. To address this, an in-house semantic-chunking algorithm, inspired by (Lattisi et al., 2022), was developed to detect significant context shifts, segment the transcript, and provide individual summaries for each segment. These partial summaries are then merged to form a coherent final summary.

2.3 Validation

The absence of a reference dataset motivated us to look at alternative approaches to validation of the generated summaries. We used metrics such as percentage of named entities captured in the summary as a proxy to measure information capture in the summary. We were also able to obtain 200 manually typed contact notes from past conversations to compare with the LLM-generated summaries. Despite these notes being highly inconsistent, we obtained a *ROUGE-1* score of 0.44.

Subjective Evaluation

A subset of the advisors were asked to evaluate the summary along three key dimensions: *precision* (faithful representation of facts and statements), *recall* (inclusion of all critical information from the conversation), and *coherence* (clarity and ease of reading). The evaluation was carried out by 22 advisors who evaluated 150+ summaries using a 5-point scale: *[Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree]*. 85% of the summaries received a "*Strongly Agree*" rating.

3 Supporting Modules

The successful deployment of AI solutions requires a robust ecosystem throughout their life cycle. For our AUTOSUMM solution, we developed and integrated modules that address three critical domains: Data Management, Model Performance, and Risk Mitigation.

3.1 Data Management

Our Data Management module includes utilities that examine the aspects of quality and information present in the conversation data to quantify and track it across various dimensions.

Data and Conversation Quality

AUTOSUMM uses an LLM-based evaluation framework to assess the quality of input conversations across three key dimensions: *linguistic quality*, *conversation flow*, and *agent response quality*.
Linguistic quality evaluates grammar, clarity, and lexical richness to ensure the transcripts meet professional standards required in banking applications.
Conversation flow is assessed using semantic embeddings to capture coherence and topic continuity across dialogue turns, helping identify fragmented or disjointed exchanges that may affect summary accuracy.
Agent response quality is evaluated using LLM-generated scores based on contextual relevance, timeliness, helpfulness, and inferred customer sentiment. These scores are also used to provide advisors with targeted feedback for performance improvement.
These metrics enable consistent benchmarking, early detection of issues with transcription quality, and feedback on the quality of conversations. Examples are included in Appendix A.2.

Semantic Clustering

To represent the thematic distribution in conversation data, we identified high-level topic clusters through a four-step methodology: First, we created a reference dataset of 600,000 sentences from conversations spanning two months; Second, we generated 384-dimensional sentence embeddings using SBERT (all-MiniLM-L6-v2) (Face, 2021a); Third, we applied Uniform Manifold Approximation and Projection (UMAP v0.5.6) (McInnes et al., 2020) for dimensionality reduction; Finally, we employed HDBSCAN (McInnes et al., 2017) for cluster generation. This process yielded 77 distinct semantic clusters, each representing topically similar content. The cluster topics were determined using a LLM. The distribution of instances across clusters establishes metrics for expected topic spread, enabling monitoring of data trends and providing explainability for distribution shifts.

Data Drift Monitoring

To complete the data management cycle, we track quality and information metrics daily. The Kolmogorov-Smirnov (KS) d-statistic quantifies divergence between reference and live data distributions. Threshold values for this statistic are established through 5,000 bootstrap iterations on both reference and live datasets.

Tables 3 and 4 demonstrate insights captured during holiday season monitoring. A significant increase in KS statistic indicated drift, with the contributing clusters reflecting expected seasonal conversation patterns.

Table 3: Data monitoring insights during the holiday season

Time Window	**Dec-H1**	**Dec-H2**	**Jan-H1**	**Jan-H2**
KS Stat	0.027	**0.046**	0.025	0.017

Table 4: Clusters with highest changes in Dec-H2

Change Trend	**Variation**
Personal and Non-Financial conversations (Holiday plans and task urgency)	**+15.1%**
Loans, debts, and real estate financial structures	-6.5%
Pensions and retirement planning	-6.2%
Interest rates and financial market fluctuations	-5.9%
Banking processes, transactions, and bank interactions	-4.8%

3.2 Model Performance Management

Evaluating summary quality without reference outputs or explicit user feedback poses a challenge. AUTOSUMM employs a multi-faceted approach to track and refine performance.

Structural and Lexical Metrics

These metrics encompass both structural elements, such as paragraph count, bullet points, sentence structure and word count, which help identify deviations over time, and lexical metrics, which assess the distribution of key entity categories (e.g., Date, Name, Time, Money, Location, Organization) to ensure that information is represented in a balanced and comprehensive manner.

Thematic Coverage

The distribution of constituent themes present in the summary are extracted for a reference dataset as a baseline. The topic distribution is then tracked across summaries, to identifying biases or systemic issues when expected themes either vary significantly or are completely absent.

Summary Edits

Since direct user feedback is limited, we analyze advisor edits to the summaries as an implicit feedback mechanism, providing quantitative insights for continuous improvement. Table 5 shows the results for one such analysis, conducted on over 300 summaries over a two-week period.

Table 5: Analysis of advisor edits to the summaries

Level of Summary Edits	Observed Cases
No Edits	89%
Minor Edits: Grammar, contextual additions while retaining key points	8.5%
Moderate Corrections: A few(< 5) words are edited.	1.5%
Major Corrections: Sentences including key information rewritten	1%

3.3 Risk Management

Operating within a banking environment requires rigorous risk management to protect customer privacy and data integrity. Beyond standard technology controls for unauthorized access and information leakage, we implemented specific measures to mitigate model output risks.

Bias Mitigation, Content Moderation

Customer and agent names are anonymized prior to summarization and re-inserted afterward, with periodic checks ensuring consistency of generated summaries across all users. Inappropriate content in input data are masked prior to summarization.

Hallucination Detection - DreamCatcher

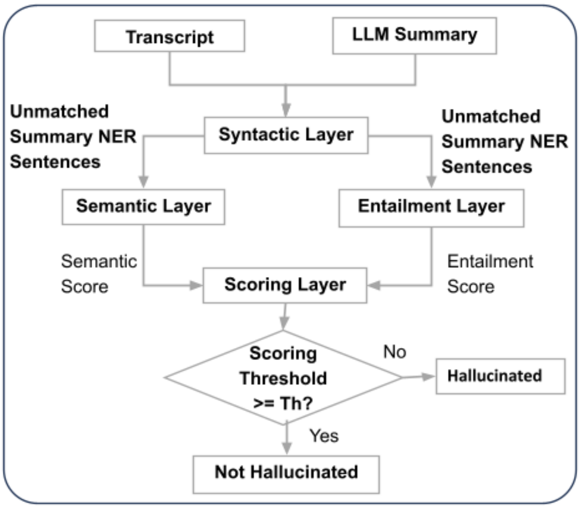

Figure 2: DreamCatcher: A multi-layered approach for hallucination detection.

We have developed a package *DreamCatcher* for detecting factual misrepresentations and hallucinations in summaries. The multi-layered approach is illustrated in Figure 2 and described as follows:
Syntactic Layer pivots hallucination checks around 18 types of named entities (e.g., MONEY, PERSON, DATE) extracted using a custom spaCy pipeline trained with a RoBERTa transformer model (Face, 2021b). Entities from summaries are compared with transcript entities using fuzzy matching for non-numeric entities.
Semantic Layer assesses semantic alignment between mismatched summary sentences and transcript segments. Transcripts are segmented via sliding window technique, with embeddings generated using Sentence-BERT. Cosine similarity is used to identify the highest matches.
Entailment Layer verifies logical consistency using a Bidirectional Auto-Regressive Transformer trained on Multi-Genre Natural Language Inference (BART MNLI (Face, 2023)) to evaluate whether summary information logically follows from the transcript.
Scoring Layer quantifies the extent of hallucination by aggregating scores from semantic and entailment layers.

The approach is validated on a human-labelled dataset, and the results shown in Table 6 indicate a high accuracy.

Table 6: Hallucination Detection Confusion Matrix

Actual Predicted	Not Hallucinated	Hallucinated
Not Hallucinated	69	1
Hallucinated	0	6

4 Deployment Process and Challenges

We implemented a phased deployment strategy, starting with a targeted pilot serving 10 advisors. This controlled introduction allowed for close performance monitoring and revealed practical considerations not evident during development. The pilot phase revealed several issues requiring remediation before full-scale deployment.

4.1 Observations - Pilot Deployment

During the pilot, we discovered that advisors couldn't consistently provide explicit feedback on each summary. In response, we introduced an edit-tracking feature to passively collect user modifications as a proxy for summary quality, which became integral to our ongoing performance monitoring framework.

Summaries occasionally included the phrase "*No Financial Information Present*" despite containing expected content, creating user confusion. Investigation showed this occurred primarily in conversations lacking explicit monetary figures. We adjusted prompt engineering parameters to eliminate this inconsistency.

Shorter conversations exhibited summaries with lower quality and more hallucinations. Since these brief interactions typically lacked significant information exchange, we implemented filtering to exclude transcripts with fewer than 50 words or under 3 minutes from summarization.

4.2 Model Version Management

To ensure smooth model transitions and minimize disruptions from OpenAI's LLM updates (OpenAI, 2025) (please refer Appendix B.1, Table 7), AUTOSUMM implements a structured version management strategy. A model upgrade during development underscored the need for a robust transition plan.

We maintain a reference dataset to benchmark new models, detect performance variations and conduct impact assessments upon release. Proactive plans are designed to seamless model transitions and minimize service interruptions, For more details please refer Appendix B.2, B.3.

4.3 Metrics Collection and Monitoring

While initial metrics for hallucination detection, data quality, and thematic coverage aided post-deployment analysis, we identified the need for more proactive monitoring. Effective production oversight requires real-time anomaly detection, stage-specific performance tracking, automated alerts for timely intervention, and comparative metrics across model versions. These enhancements ensure operational stability and drive continuous optimization.

4.4 Human-in-the-Loop Integration

Human oversight is crucial for quality assurance and risk mitigation. Advisor feedback enables early detection of model failures, sensitive content, and optimization opportunities while fostering user trust and transparency. This approach aligns with financial regulatory requirements, ensuring appropriate governance of AI-generated content.

5 Conclusion

This paper presents AUTOSUMM, a production-ready LLM-based summarization system for financial customer-advisor conversations. Designed to meet the operational, regulatory, and ethical demands of the financial domain, the system addresses challenges such as speaker attribution errors, short transcripts, and hallucinations through solutions like entity-aware processing, transcript filtering, and the DreamCatcher module.

Human-in-the-loop oversight ensures quality and compliance, with real-time advisor feedback integrated into system refinement. A structured model versioning strategy minimizes deployment disruptions, while custom monitoring tracks performance, drift, and thematic coverage. Ethical considerations—privacy, fairness, transparency, and accountability—are embedded throughout the system lifecycle.

AUTOSUMM provides a practical framework for safely and effectively deploying LLMs in regulated environments, demonstrating that operational impact and compliance can be achieved together.

Acknowledgments

The authors wish to express their gratitude to Zachery Anderson, Chief Data and Analytics Officer, and Graham Smith, Head of Data Innovation, for their invaluable support and for fostering a constructive research environment that was essential throughout the AUTOSUMM project. We also extend our thanks to Sacheen Patel and Alyssa Anderson from Wealth Businesses for their insightful feedback from a business perspective, which significantly enhanced the project's outcomes. Additionally, we thank Grant Falconer and Faye Drummond for their support and enablement that helped to manage the challenges and ensured smooth delivery of the solution. Lastly, we would like to extend our appreciation to all colleagues who contributed to the development of the project.

Ethical Considerations

AUTOSUMM has been deployed within a regulated financial institution, where ethical concerns around privacy, fairness, and responsible automation are critical. Below, we outline key ethical dimensions and our corresponding mitigations:

Data Privacy and Consent The conversation transcripts used for training and evaluation were collected under institutional agreements with appropriate customer disclosures. No personally identifiable information (PII) was retained in training data; names and sensitive fields were anonymized prior to model processing. Data access was restricted to authorized personnel, and all processing occurred within the bank's secure infrastructure, in compliance with GDPR and internal data governance policies.

Bias and Fairness Given the risk of demographic, occupational, or financial bias in LLM outputs, we implemented pre- and post-generation checks. Names and gendered terms were anonymized prior to summarization, reducing exposure to learned biases. We conducted periodic reviews of summary outputs to detect patterns of exclusion or stereotyping.

Transparency and Accountability To support explainability, the system logs all inputs and generated summaries along with metadata (model version, prompt variant, edit history). The DreamCatcher module provides sentence-level hallucination flags, supported by entailment and semantic similarity scores. These signals are exposed to users and reviewers during post-call audits, enabling traceability and informed review.

Human Oversight and Model Governance Summaries are reviewed by financial advisors before being saved to customer records. This oversight loop helps catch factual errors, omissions, and contextually inappropriate content. We also maintain clear version control over models and prompts; any change undergoes pre-deployment evaluation against reference benchmarks and advisor feedback. Governance boards review deployment plans to ensure alignment with regulatory expectations (e.g. suitability, fairness, and auditability under financial regulations).

Limitations and Responsible Use AUTOSUMM is not used in decision-making or product recommendations. Summaries are designed to aid documentation and post-call follow-up—not to replace judgment or communication. We do not claim complete factual accuracy in generated outputs, and all summaries are clearly marked as AI-generated drafts subject to advisor validation.

Broader Societal Impact The system was deployed to improve documentation efficiency, not to replace human roles. By reducing administrative burden, advisors spend more time on relationship-building. Nevertheless, we acknowledge concerns about automation displacing cognitive tasks and continuously engage with internal stakeholders to ensure responsible, assistive use.

References

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*.

Hugging Face. 2021a. all-minilm-l6-v2: A lightweight transformer model for sentence embeddings. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2024-04-11.

Hugging Face. 2021b. Roberta: A robustly optimized bert pretraining approach. https://huggingface.co/docs/transformers/en/model_doc/roberta. Accessed: 2024-07-10.

Hugging Face. 2023. Bart-large-mnli: A bart model fine-tuned for multi-class natural language inference. https://huggingface.co/facebook/bart-large-mnli. Accessed: 2024-06-14.

Simon Hughes. 2025. Vectara hallucination leaderboard. Accessed: 2025-03-15.

Subhendu Khatuya, Koushiki Sinha, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2024. Instruction-guided bullet point summarization of long financial earnings call transcripts. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

T. Lattisi, D. Farina, and M. Ronchetti. 2022. Semantic segmentation of text using deep learning. *Computing and Informatics*, 41:78–97.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ Digital Medicine*, 6(1):120.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics*, 4:1085–1115.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10893–10906. Association for Computational Linguistics.

OpenAI. 2023. Introducing gpt-3.5 turbo with 16k token limit.

OpenAI. 2025. Deprecations - openai documentation. `https://platform.openai.com/docs/deprecations`. Accessed: 2025-03-03.

Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2024. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on Chinese legal domain. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5030–5041, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Pierric Cistac, Nathan Hunter, Julien Plu, and et al. 2021. Transformers: State-of-the-art natural language processing. `https://huggingface.co/docs/transformers/model_doc/bert`. Accessed: 2024-01-10.

Li Wu, Jun Xu, Sherry Thakkar, Matthew Gray, Yu Qu, Dong Li, and Weida Tong. 2024. A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document. *Regulatory Toxicology and Pharmacology*, 149:105613.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

A Appendix

A.1 Example Summary

Example

Summary:
- The agent apologized for missing the customer's call earlier due to a meeting.
- The customer discussed their holiday in the Maldives for their son's birthday last month.
- The agent noted a request to increase the customer's credit card limit to 10,000 GBP.
- The customer expressed frustration and disbelief over the valuer's decision, mentioning that their London property is freehold.
- The customer is planning to invest approximately GBP 2.5 million in real estate in Singapore. They must complete legal formalities within the next two months.
- In November of last year, the client's husband was diagnosed with cancer and is undergoing treatment at a hospital.
- The client's son is going to MIT next year, which will cost them 5,000 dollars monthly, including all expenses.

Actions:
- Agent to submit request for increasing credit card limit.
- Agent to escalate the valuation related issue with the Mortgage Team.

A.2 Data and Conversation Quality Examples

A.2.1 Snippet of a Good-Quality Conversation

Grammatically correct sentences without spelling or transcription errors. Response is relevant to the query and does not deviate from the topic.

Customer-Agent Exchange

Customer: "I'd like to explore a new investment option because I've recently changed jobs and want more flexibility."
Agent: "Absolutely. Let's review your current portfolio and discuss how the job change might affect your risk profile and liquidity needs."

A.2.2 Low Quality Conversation Snippet

Poorly transcribed sentences with missed words and incorrect grammar.

Customer-Agent Exchange

Customer: "i need help with my 401k contributions im not sure if uh um doing right or ah how much should be putting each"
Agent: "We can help with that. Their are different opinions on contribution amounts, but typically we recommend about 15% of you're income. When did you last adjusted your contributions?"

B Model Version Management Strategy

B.1 Model Deprecation Timelines

Table 7: LLM deprecation timelines and context lengths

GPT model name	Deprecation date	Context length
GPT 3.5 Turbo 0613	September 2024	16000 tokens
GPT 3.5 Turbo 1106	December 2024	16000 tokens
GPT 3.5 Turbo 0125	February 2025	16000 tokens
GPT 4o Mini	Not Known	128000 tokens

B.2 Risk Mitigation and Deployment Strategy

To ensure smooth deployment, we adopt a staged rollout strategy:

1. Phase 1: Offline Testing – Running controlled experiments with historical data.
2. Phase 2: Shadow Mode Deployment – Generating outputs from both old and new models in parallel without affecting users.
3. Phase 3: Limited Production Rollout – Gradually increasing the proportion of queries handled by the new model.
4. Phase 4: Full Deployment – Transitioning entirely once performance benchmarks are met.

Additionally, we version control model prompts to ensure that refinements are systematically evaluated before deployment. If a newer model exhibits unexpected deviations, we have rollback procedures in place to revert to a stable version.

B.3 Continuous Monitoring Post-Deployment

Post-deployment, we track performance via:

- Daily Summarization Quality Audits: Automated and manual reviews to detect performance drifts.

- User Feedback Loop: Continuous collection of feedback from financial advisors to fine-tune the system.
- Data Drift Analysis: Monitoring changes in input conversation patterns that may impact model behavior.

This systematic version management approach ensures that AUTOSUMM remains resilient to LLM deprecations, providing stable and high-quality summarization outputs without disrupting financial workflows.