Call Support Copilot: A Reproducible Multimodal System for Speech Emotion Recognition, Intent Understanding, and Agent Assistance

Khoshimov Rakhmatillokhon
University of Zurich
Dept. of Informatics
rakhmatillokhon.khoshimov@uzh.ch

Dmitry Rudshin
University of Zurich
Dept. of Informatics
dmitry.rudshin@uzh.ch

Yanyang Luo
University of Zurich
Dept. of Informatics
yanyang.luo@uzh.ch

Abstract

We present Call Support Copilot, a reproducible multimodal system that integrates automatic speech recognition, speech emotion recognition, machine translation, spoken language understanding, and client knowledge retrieval in a single dashboard for customer support agents. Built from publicly accessible pretrained models and standard benchmarks, the system transcribes speech with Whisper-family ASR (Radford et al., 2023; Klein, 2023), detects caller affect in valence-arousal-dominance terms (Mehrabian, 1996; Russell, 1980), classifies intents from a banking-domain inventory of 77 categories (Casanueva et al., 2020), and retrieves client records from a database. Evaluation shows strong component performance: 6.6% word error rate on LibriSpeech (Panayotov et al., 2015), 91.7% macro-F1 on SUPERB ER session1 (IEMOCAP subset, n=6) (Yang et al., 2021; Busso et al., 2008), 42.98 BLEU (Papineni et al., 2002; Post, 2018) for German–English translation, and 87.0% accuracy on BANKING77 intent classification. End-to-end benchmarking of the core pipeline achieves faster-than-real-time throughput with mean real-time factor 0.67–0.71. All model identifiers, configurations, and evaluation scripts are documented in the accompanying repository, supporting reproducibility in line with the SwissText 2026 theme.

1 Introduction

Customer service call centers remain essential infrastructure for financial institutions, healthcare providers, and commercial enterprises (Gao et al., 2019). The quality of these interactions depends critically on human agents' ability to understand callers quickly, respond to their emotional states appropriately, and access relevant account information without disrupting the conversation flow. Traditional call center technology provides basic telephony functions but offers little support for these cognitive demands (Ram et al., 2018; Hosseini-Asl et al., 2020).

Recent advances in speech and language processing have created opportunities to address these limitations. Models such as Whisper (Radford et al., 2023) and efficient implementations such as faster-whisper (Klein, 2023) achieve robust speech recognition across diverse acoustic conditions, building on foundational work in self-supervised speech representation learning (Baevski et al., 2020; Hsu et al., 2021). Speech emotion recognition systems trained on corpora like IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019) can detect affective states following the dimensional affect framework (Russell, 1980; Mehrabian, 1996). Neural machine translation enables cross-lingual communication (Fan et al., 2021; Junczys-Dowmunt et al., 2018), while intent classification models trained on domain-specific datasets such as BANKING77 (Casanueva et al., 2020) can identify caller needs from transcribed speech, extending work in spoken language understanding (Young et al., 2013).

However, most published systems focus on individual components rather than integrated pipelines, and reproducibility remains a significant challenge in NLP research (Belz et al., 2021; Dodge et al., 2019).

This paper presents Call Support Copilot, a system that combines these technologies in an integrated dashboard for customer support agents. Our contributions are twofold. First, we present a **reproducible** implementation assembled from publicly accessible model checkpoints and resources, with model identifiers, configurations, and evaluation scripts documented in an accompanying repository. Second, we provide quantitative evaluation across all major system components using established benchmarks, showing that strong performance is achievable with an accessible, modular architecture without training ASR, MT, or SER models from scratch. In line with the SwissText 2026 theme of

Reproducible NLP, we focus on a pipeline that can be replicated end-to-end using publicly accessible resources.

2 Related Work

Call-center assistance systems are closely related to task-oriented dialogue and conversational AI, where systems combine speech or text understanding with dialogue-state tracking, retrieval, and response or action selection (Young et al., 2013; Gao et al., 2019; Ram et al., 2018). Datasets such as MultiWOZ (Budzianowski et al., 2018) and BANKING77 (Casanueva et al., 2020) support evaluation of dialogue and intent-understanding modules, but they do not by themselves evaluate the full operational chain from audio input to agent-facing recommendations.

Our system follows a modular integration strategy rather than proposing a new model architecture. This places it between component-level benchmark work on ASR (Radford et al., 2023; Panayotov et al., 2015), MT (Junczys-Dowmunt et al., 2018; Post, 2018), SER (Schuller, 2018; Latif et al., 2021), and intent classification (Casanueva et al., 2020), and applied systems that combine these components in a practical workflow. The comparison to Casanueva et al. (Casanueva et al., 2020) anchors the SLU result against a published BANKING77 baseline, while the other reported metrics are intended as reproducible component checks rather than claims of state-of-the-art performance.

Speech emotion recognition has particular reproducibility and validity concerns. IEMOCAP is widely used, but recent analyses highlight issues around modality dependence, recording quality, ambiguous labels, and misclassifications that may be unsurprising even to human annotators (Probol and Mieskes, 2023). We therefore treat the IEMOCAP-derived SER result as a limited benchmark sanity check, not as evidence that the system generalizes to spontaneous customer-support speech.

3 System Architecture

Call Support Copilot processes audio input through five interconnected modules (Figure 1).

Speech Recognition. Audio recordings are processed through Whisper (Radford et al., 2023) using `faster-whisper` (Klein, 2023) with CTranslate2 and int8 quantization, building on advances in self-supervised speech representations (Baevski et al., 2020; Hsu et al., 2021; Graves et al., 2006). Voice activity detection (Silero Team, 2021) filters silence. The system supports WAV, MP3, M4A, FLAC, and OGG formats with FFmpeg conversion.

Machine Translation. When the detected language differs from English, transcripts pass through Helsinki-NLP MarianMT (Junczys-Dowmunt et al., 2018) transformer models (Vaswani et al., 2017) with SentencePiece tokenization (Kudo and Richardson, 2018) trained on OPUS parallel data (Tiedemann, 2012). Six source languages are supported: German, Dutch, French, Spanish, Italian, and Portuguese.

Speech Emotion Recognition. Audio is analyzed using MERaLiON-SER-v1 (MERaLiON Team, 2025), outputting categorical emotions following Ekman (Ekman, 1992) and PAD dimensions (Mehrabian, 1996; Russell, 1980). Seven categories are classified: neutral, happy, sad, angry, fearful, disgusted, and surprised. A sliding window approach (4 s windows, 2 s overlap) handles varying durations (Schuller, 2018), producing emotion timelines with deep learning representations (Latif et al., 2021).

Spoken Language Understanding. Intent classification uses DistilBERT (Sanh et al., 2019), a knowledge-distilled variant of BERT (Devlin et al., 2019), fine-tuned on BANKING77 (Casanueva et al., 2020) with 77 intent categories. Low-confidence predictions are flagged as out-of-domain (Larson et al., 2019). Slot extraction uses regex patterns for amounts, dates, reference IDs, and merchant names, tracking missing required slots (Henderson et al., 2014).

Action Generation. Caller phone numbers are used to query a client database. The system then combines intent predictions, emotion states, extracted slots, and client records to generate suggested actions (Gao et al., 2019), including intent-specific recommendations, slot-based follow-ups, and emotion-triggered de-escalation guidance, following task-oriented dialogue practice (Young et al., 2013; Budzianowski et al., 2018).

4 Evaluation Setup

We evaluate the integrated system with fixed model identifiers and released repository scripts, matching the configurations used to generate the bundled result artifacts. For ASR and MT, we report mean per-

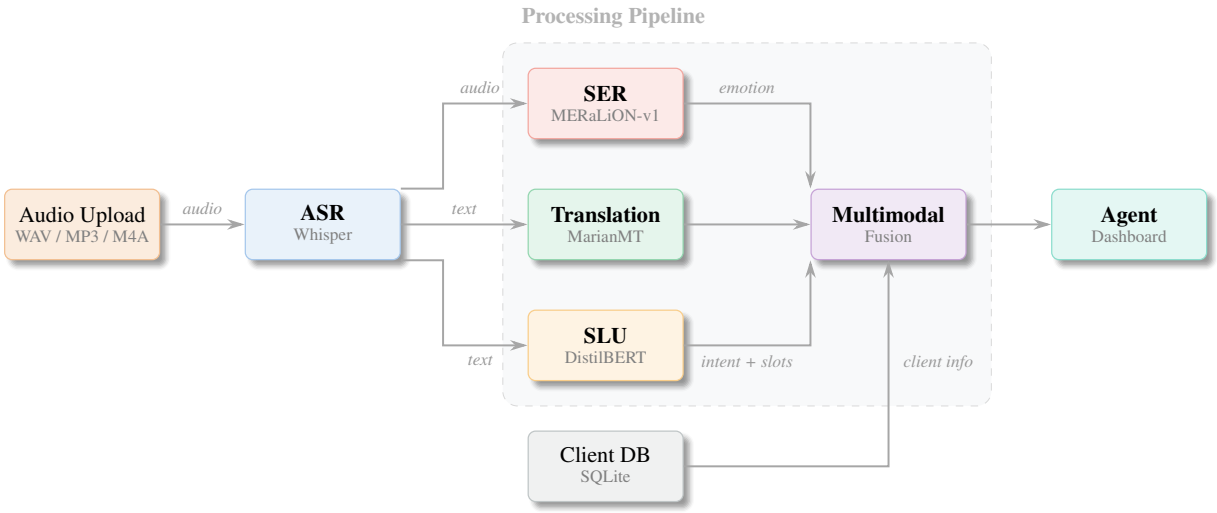

Figure 1: Call Support Copilot system architecture. Uploaded audio flows through ASR (Whisper) which feeds three parallel branches: Speech Emotion Recognition (SER) for affective state detection, Machine Translation (MT) for multilingual support, and Spoken Language Understanding (SLU) for intent classification and slot extraction. These signals are fused with client records retrieved from the knowledge base to generate context-aware recommendations displayed on the agent dashboard.

formance over 50 randomly sampled items with a fixed seed from LibriSpeech test-clean (Panayotov et al., 2015) and WMT16 German–English test data. For SER, we use the SUPERB ER session1 split (Yang et al., 2021), which exposes a small IEMOCAP-derived four-class benchmark (Busso et al., 2008) in our workflow. For latency, we run the full pipeline on LibriSpeech and SUPERB ER utterances and record wall-clock times for ASR, SER, MT, and database lookup.

The reported component scores are measured at component boundaries. ASR WER is computed from audio against reference transcripts, MT BLEU is computed on WMT text pairs, SER macro-F1 is computed on the SUPERB ER audio examples, and SLU accuracy is computed on BANKING77 text rather than on ASR-generated transcripts. The end-to-end experiment measures runtime through the actual pipeline, but it does not estimate downstream accuracy after ASR error propagation. This distinction is important: the deployed dashboard does pass ASR output to downstream modules, but the present paper reports technical component benchmarks and pipeline latency rather than a full cascaded-error evaluation.

Metrics are chosen to reflect standard practice for each module. ASR uses word error rate (WER), MT uses corpus BLEU (Papineni et al., 2002; Post, 2018), SER uses macro-F1 over the four mapped emotion labels, and SLU uses accuracy and macro-F1 together with per-intent precision, recall, and F1. End-to-end efficiency is summarized with real-time factor (RTF), where values below 1 indicate faster-than-real-time processing. Table 1 summarizes the core results.

5 Results and Discussion

Speech Recognition. On LibriSpeech test-clean (Panayotov et al., 2015), our system achieves **6.6% WER** (95% CI [4.4%, 9.2%]) on 50 samples. Errors arise mainly from proper nouns and archaic vocabulary. This is competitive with reported Whisper performance on clean speech (Radford et al., 2023).

Machine Translation. German-to-English achieves **42.98 BLEU** (Papineni et al., 2002; Post, 2018) on 50 sentence pairs with 0.86 s mean inference time, in line with Helsinki-NLP MarianMT benchmarks on WMT data (Junczys-Dowmunt et al., 2018; Tiedemann, 2012).

Speech Emotion Recognition. On SUPERB ER session1 (IEMOCAP subset, four-class labels, n=6) (Yang et al., 2021; Busso et al., 2008), full-audio processing achieves **91.7% macro-F1**. Table 2 shows dimensional predictions consistent with the PAD model (Mehrabian, 1996; Russell, 1980).

Intent Classification. On BANKING77 (Casanueva et al., 2020), DistilBERT (Sanh et al., 2019) achieves **87.0% accuracy** and 0.863 macro-F1. Table 3 shows per-category results for

Component	Dataset	Metric	Value
ASR	LibriSpeech	WER	6.61%
MT	WMT DE–EN	BLEU	42.98
SER	SUPERB ER	Macro-F1	0.917
SLU	BANKING77	Accuracy	87.0%
System	Libri+SUPERB	RTF	0.67–0.71

Table 1: Summary of evaluation results across all system components on standard benchmarks, plus end-to-end real-time factor (RTF) for the core pipeline.

Emotion	V	A	D
Happy	0.85	0.70	0.68
Neutral	0.47	0.48	0.52
Sad	0.24	0.08	0.18
Angry	0.28	0.84	0.75

Table 2: Mean valence (V), arousal (A), and dominance (D) predictions for each emotion category on IEMOCAP.

key banking intents. For comparison, Casanueva et al. report 85.19% with USE+ConveRT on this benchmark (Casanueva et al., 2020).

End-to-End Latency. We benchmark the core pipeline (ASR + SER + optional MT + DB lookup). Real-time factor (RTF) values below 1 indicate faster-than-real-time throughput. On LibriSpeech test-clean (n=10), mean RTF is **0.71**; on SUPERB ER (n=6), mean RTF is **0.67**. SER dominates latency due to sliding-window analysis.

The results show that the architecture is strongest where mature pretrained models can be integrated with limited task-specific adaptation. ASR and MT remain within expected benchmark ranges despite the lightweight engineering stack, and the SLU component improves on the 85.19% BANKING77 USE+ConveRT baseline reported by Casanueva et al. (Casanueva et al., 2020). This supports the core claim of the paper: a practical customer-support copilot can be assembled from publicly accessible building blocks without retraining the full stack.

The evaluation also clarifies the current bottlenecks. SER contributes the largest share of end-to-end latency because sliding-window inference scales with utterance duration, and the reported SER score is derived from a very small SUPERB ER sample in the released results. The runtime numbers are therefore encouraging for uploaded-call processing, but they do not yet establish readiness for live streaming deployments. Likewise, intent prediction is strong on BANKING77, yet real customer calls remain harder because ASR errors,

Intent Category	P	R	F1
transaction_charged_twice	.889	1.00	.941
lost_or_stolen_card	.804	.925	.860
refund_not_showing_up	.947	.900	.923
card_arrival	.756	.850	.800
activate_my_card	1.00	.925	.961
verify_my_identity	.745	.875	.805
Macro avg (77 cls)	.875	.870	.863

Table 3: Intent classification results for selected banking intents on the BANKING77 test set (Casanueva et al., 2020).

spontaneous speech, and slot omissions compound downstream uncertainty.

An important observation is that the component scores should not be interpreted independently of pipeline coupling. In the deployed interface, translation quality depends directly on ASR transcription fidelity, and intent classification inherits both lexical errors and disfluencies from the upstream recognizer. This means that the strongest standalone module is not automatically the most useful one operationally: modest ASR degradation on accented or noisy calls can propagate into wrong intent predictions or misleading summaries even when the classifier remains strong on clean BANKING77 text. For customer-support settings, cross-component robustness is therefore at least as important as any single benchmark number.

6 Reproducibility

In the spirit of reproducible NLP (Belz et al., 2021; Dodge et al., 2019), all core models and resources used in the system are publicly accessible:

- **ASR**: `Systran/faster-whisper-base` via `faster-whisper` (Klein, 2023) with int8 quantization
- **MT**: Helsinki-NLP MarianMT `opus-mt-{de,nl,fr,es,it,pt}-en`
- **SER**: `MERaLiON-SER-v1` (MERaLiON Team, 2025)
- **SLU**: `distilbert-base-uncased` fine-tuned on BANKING77
- **VAD**: Silero VAD (Silero Team, 2021)

The system is implemented in Python using Streamlit, PyTorch, and HuggingFace Transformers. Audio preprocessing converts input to 16 kHz mono using `soundfile` and FFmpeg. The client database uses SQLite. All

evaluation scripts, result files, and model configurations are included in the accompanying repository released with the paper: https://github.com/rakhmatillokhon-khoshimov/call-support-copilot.

To make the submission auditable rather than merely runnable, the repository records model identifiers, quantization choices, dataset slices, and summary CSV/JSON outputs for each experiment. This is especially relevant for a modular system, where implementation choices such as VAD preprocessing, CPU quantization, or label remapping can materially affect downstream scores. By exposing these settings together with the measured outputs, the paper makes it possible to verify not only that the code executes, but also that the reported benchmark numbers correspond to a specific, inspectable evaluation protocol.

7 Conclusion

We presented Call Support Copilot, a reproducible multimodal system integrating speech recognition (Radford et al., 2023; Klein, 2023), emotion detection (MERaLiON Team, 2025), translation (Junczys-Dowmunt et al., 2018), and intent understanding (Casanueva et al., 2020) to assist customer support agents. Evaluation on established benchmarks shows strong component performance with practical latency, achieved without training ASR, MT, or SER models from scratch. The modular implementation supports both practical deployment and research extension, and aligns with the goal of making NLP systems reproducible (Belz et al., 2021).

Limitations

Our evaluation samples are small due to computational constraints (50 samples for ASR and MT, and 10/6 samples for latency on LibriSpeech/SUPERB ER). The SER model, trained on acted IEMOCAP speech (Busso et al., 2008), may not generalize to spontaneous customer-service conversations (Schuller, 2018; Poria et al., 2019). IEMOCAP-derived evaluation should also be interpreted cautiously because prior work reports issues around recording quality, task ambiguity, and the relative strength of text-only emotion models (Probol and Mieskes, 2023). Slot extraction patterns are English-specific and rely on hand-crafted regex rules rather than learned extractors. The supported MT setup is limited to six Western and Central European source languages (German, Dutch, French, Spanish, Italian, and Portuguese), excluding many low-resource, non-European, and Swiss-relevant language varieties. The current system processes uploaded recordings rather than streaming live audio, which limits real-time deployment scenarios. ASR may exhibit demographic disparities for accented speech (Ardila et al., 2020), and emotion recognition may reflect cultural biases in the training data (Ekman, 1992). Future work includes streaming ASR with incremental emotion updates, speaker diarization, CRM integration, broader language coverage, and large language models (Brown et al., 2020) for more flexible action generation.

The present study evaluates technical components rather than end-user outcomes. We do not measure whether agents resolve issues faster, make fewer mistakes, or perceive the dashboard as trustworthy under realistic workload. Likewise, the database layer is demonstrated on a controlled SQLite setup rather than in a production CRM environment with authentication, logging, and policy constraints. These omissions matter because deployment quality depends not only on benchmark performance, but also on usability, integration overhead, and organizational acceptance. A stronger follow-up study should therefore pair component benchmarks with human-centered evaluation in a realistic support workflow.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Anya Belz, Simon Mille, and David M Howcroft. 2021. ReproGen: A first step towards a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 338–350.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-

shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ – a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerber, Milica Gasic, and Matthew Henderson. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2185–2194.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*, 13:127–298.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 263–272.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, and 1 others. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Guillaume Klein. 2023. faster-whisper. GitHub repository.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, and 1 others. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 1311–1316.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W Schuller. 2021. Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 14(2):1634–1654.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

MERaLiON Team. 2025. MERaLiON-SER: Robust speech emotion recognition model for english and SEA languages. *arXiv preprint arXiv:2511.04914*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, and Erik Cambria. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Nadine Probol and Margot Mieskes. 2023. Emotions in spoken language – do we need acoustics? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 71–84, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nuber, Bo Xu, Ming Yang, and 1 others. 2018. Conversational AI: The science behind the Alexa prize. *arXiv preprint arXiv:1801.03604*.

James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Silero Team. 2021. Silero VAD: Pre-trained enterprise-grade voice activity detector. GitHub repository.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankun Chang, Guan-Ting Lin, and 1 others. 2021. SUPERB: Speech processing universal PERformance benchmark. *arXiv preprint arXiv:2105.01051*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

A Supported Translation Languages

Source Lang.	Model	Avg. Time (s)
German (DE)	opus-mt-de-en	0.86
Dutch (NL)	opus-mt-nl-en	0.82
French (FR)	opus-mt-fr-en	0.79
Spanish (ES)	opus-mt-es-en	0.81
Italian (IT)	opus-mt-it-en	0.84
Portuguese (PT)	opus-mt-pt-en	0.83

Table 4: Supported translation language pairs with Helsinki-NLP MarianMT models (Junczys-Dowmunt et al., 2018) and average inference times on CPU.

B Slot Extraction Details

Following standard approaches in spoken language understanding (Henderson et al., 2014), regex patterns extract: amounts with currency symbols (CHF, USD, EUR); dates in ISO, European, and natural formats; reference IDs; card last-four digits; and merchant names.

C Emotion Categories

Following Ekman (Ekman, 1992) and PAD (Mehrabian, 1996; Russell, 1980): Neutral (moderate valence, low arousal), Happy (high valence, moderate arousal), Sad (low valence, low arousal), Angry (low valence, high arousal, high dominance), Fearful (low valence, low dominance), Disgusted (low valence), Surprised (variable valence, high arousal).

D Action Generation Rules

Intent-specific rules following task-oriented dialogue best practices (Young et al., 2013; Budzianowski et al., 2018): Chargeback triggers transaction confirmation and dispute workflow. Lost card triggers freeze, replacement, and fraud review. Angry emotion prepends de-escalation guidance. Missing slots generate follow-up prompts (Henderson et al., 2014).