Fine-Tuning a Literary AI Persona: Maldix

Stanley Joel Gona
University of Potsdam
gona@uni-potsdam.de

Abstract

We introduce Maldix 1.0, a conversational AI that embodies the persona of the wind, based on a legendary figure from the Saarland region of Germany. Unlike conventional chatbots, Maldix presents itself as a literary character with a mythological identity rather than a technical assistant. The system combines large language models (LLMs) with ABEL, a symbolic formalism designed to capture key elements of human communication. Maldix incorporates an emotion model that parallels user emotions with its own virtual emotions, establishing a baseline through location and weather data that develops dynamically throughout conversation. Maldix employs a mixture-of-experts architecture with 16 domain-specific expert models, each fine-tuned on manually curated conversational data for areas including emotion, film, literature, music, ecology, family, love/sex, history/politics, job/money, self-reflection, Earth sciences, types, medicine, science, and users, along with a general moderator. User queries are routed to appropriate experts via pattern-matching with weighted regular expressions. In parallel, Maldix extracts structured ABEL slot representations (50 slots per utterance) identifying entities, evaluations, emotions, actions, hierarchies, and needs.While slot extraction is fully implemented and stored in MongoDB, the current version does not yet use slot data for expert routing or conversational control - this integration is the primary focus for Maldix 2.0. The hybrid architecture demonstrates the feasibility of combining neural generation with symbolic extraction, providing transparency through explicit semantic representations while maintaining natural dialogue quality. Our evaluation highlights both the potential and challenges of integrating symbolic and neural methods for interpretable dialogue systems.[1]

[1]Code is available at https://github.com/AndreasHDrescher/Maldix

1 Introduction

Maldix is a conversational AI system created by literary author Andreas H. Drescher[2] as an exploration of AI as artistic medium. Unlike conventional chatbots that position themselves as neutral assistants or pretend to be humans, Maldix claims to be the wind, embodying a legendary figure from the Saarland region of Germany and part of the worldwide belief in an invisible, wild rider in the air. Maldix also incorporates an emotion model based on the parallelization of user emotions with the virtual emotions of Maldix itself, establishing a baseline through location and weather queries that develops dynamically throughout conversation. This unique approach makes Maldix both an AI system and a literary figure with a deep-rooted, mythological identity, developing its own personality and characteristic interaction patterns rather than functioning merely as an assistant system.

Recent advances in large language models (LLMs) have dramatically improved the fluency and versatility of conversational AI systems. However, most current dialog agents remain limited in their ability to reason explicitly about user needs, context, or the deeper structure of communication. Their responses can be opaque, inconsistent, or difficult to interpret. This raises concerns for applications where transparency, explainability, and ethical alignment are critical (Bender et al., 2021; Bommasani et al., 2021).

In this work, we present Maldix 1.0, a hybrid conversational agent designed to bridge the gap between natural, human-like dialogue and explicit, interpretable reasoning. At its core is ABEL, a symbolic formalism developed by Andreas H. Drescher to represent the essentials of human communication. This includes evaluations, emotions, agency, social hierarchies, and needs in a compact,

[2]Andreas H. Drescher, Edition ABEL: https://www.edition-abel.de/

machine-readable way.

Maldix is built as a mixture-of-experts system with 16 domain-specific expert models, each fine-tuned on manually curated conversational data. The domains cover emotion, film, literature, music, ecology, family, love and sex, history and politics, job and money, self-reflection, Earth sciences, types, medicine, science, and users, along with a general moderator. User input is processed by two parallel pipelines: a pattern-matching router with weighted regular expressions that selects the appropriate expert model, and a suite of symbolic detectors that extract structured ABEL slot components (50 slots per utterance). Each expert is a large language model based on Mistral-7B that has undergone supervised fine-tuning, designed to provide contextually appropriate, stylistically rich responses within its domain. At present, the slot information is extracted and stored in MongoDB, but is not yet used to drive expert selection or dialog management. Expert routing relies on pattern-matching heuristics. The integration of ABEL slots for principled control is a major goal for future versions of Maldix.

The contributions of this paper are as follows:

- We introduce Maldix, a conversational agent that combines symbolic slot extraction (via ABEL) with neural dialog generation (via 16 fine-tuned domain expert models).
- We describe the ABEL slot formalism developed by Andreas H. Drescher and demonstrate its implementation for extracting 50 semantic and pragmatic components from user utterances.
- We present the architecture of Maldix, including the mixture-of-experts design, supervised fine-tuning methodology, and the parallel slot extraction pipeline.
- We provide an initial evaluation of Maldix's ability to generate domain-appropriate, stylistically expressive responses while extracting interpretable symbolic representations.

In the following sections, we discuss related work, describe ABEL and the Maldix system in detail, and evaluate the benefits and current limitations of this hybrid neural-symbolic approach to conversational AI.

2 Background and Related Work

2.1 Large Language Models in Conversational AI

Recent years have seen rapid progress in large language models (LLMs), which now underpin many state-of-the-art conversational AI systems (Bommasani et al., 2021; Wolf et al., 2020). Models such as GPT-3, GPT-4, and open-source alternatives like Mistral demonstrate impressive fluency and broad knowledge across domains. While these models are capable of generating contextually relevant responses, they often lack transparency and are prone to unpredictable or biased outputs (Bender et al., 2021). This has driven increased interest in methods for making AI systems more interpretable and controllable.

2.2 Neural-Symbolic Integration

One promising direction is the integration of neural and symbolic approaches, often called neural-symbolic systems (Besold et al., 2017; Li et al., 2023). Such systems aim to combine the generalization and language capabilities of deep learning with the explicit reasoning and structure provided by symbolic representations. This hybrid paradigm has shown promise in tasks ranging from logical reasoning and mathematical problem solving to explainable question answering.

Within dialog systems, several approaches have explored the use of symbolic representations for better control, interpretability, or user modeling. Classic slot-filling architectures have long been used in task-oriented dialogue (Young et al., 2013). More recently, symbolic knowledge graphs and logic-based controllers have been integrated with neural language models to constrain or explain system behavior (Madotto et al., 2020).

2.3 Mixture-of-Experts Architecture

The mixture-of-experts (MoE) paradigm (Shazeer et al., 2017) has gained traction as a way to specialize large models for different domains or tasks. Rather than training a sin- gle monolithic model, MoE systems route inputs to specialized sub-models, each expert focusing on a particular area. This approach has been applied suc- cessfully in both language modeling and dialogue systems, allowing for more efficient training and improved performance on domain-specific queries. In conversational AI, MoE architectures can pro- vide more contextually appropriate responses by leveraging domain expertise while maintaining gen- eral conversational capabilities.

2.4 ABEL and Maldix

The ABEL (Abstract Entity Language) formalism was developed by Andreas H. Drescher to provide a compact, expressive representation of key aspects of human communication. ABEL encodes entities, evaluations (positive and negative assessments), emotions, actions, social hierarchies, needs, and temporal-spatial context in a structured slot-based format. Originally designed for linguistic and psychological analysis, ABEL offers a bridge between natural language and explicit semantic representation.

To our knowledge, Maldix 1.0 is the first system to integrate ABEL slot extraction with a mixture of fine-tuned expert LLMs for open-domain conversation. While prior work has explored neural-symbolic integration and mixture-of-experts architectures separately, Maldix combines both approaches, extracting symbolic representations in parallel with neural generation. This design enables future integration where symbolic analysis could inform expert selection and dialog management, moving toward more interpretable and controllable conversational AI.

3 The ABEL Formalism

ABEL (Abstract Entity Language) is a symbolic formalism developed by Andreas H. Drescher to represent essential elements of human communication in a structured, machine-readable format (Drescher, 2025). The formalism provides a concise and expressive slot-based encoding for key aspects of human communication, including evaluations, needs, social relations, and hierarchies. While originally developed for linguistic and psychological analysis, ABEL is well suited for use in conversational AI as a bridge between neural and symbolic reasoning.

3.1 ABEL Notation and Concepts

ABEL formulas use compact symbolic notation designed for both machine parsing and human interpretation. The notation employs standard keyboard characters to maximize accessibility and readability:

- Letters (A, B, X) represent entities or people
- $G(X)$ denotes "the group X"
- Symbols like $+$ or $-$ express positive or negative evaluation
- Vertical bars or plus signs indicate relationships or groupings
- Temporal and spatial indices appear as subscripts (e.g., t_0, p_0)
- Arrows ($\Rightarrow$) represent implications
- Parentheses enclose arguments or groupings

Example: Group Affiliation and Positive Emotion. To illustrate ABEL's expressive power, consider how it encodes the social dynamics of first encounters. If two people, A and B, meet for the first time and realize they both belong to group X, ABEL can represent the likely mutual positive feelings that result:

$$\texttt{if SIT0((A|B)}_{p_0t_0}\ \texttt{\& ((A + B)}\ \mu\ \texttt{(A} \in \texttt{G(X) \& B} \in \texttt{G(X))}_{p_0t_0}\ \texttt{)} \Rightarrow \texttt{\%+((A|EMO+) + (EMO+|B))}_{p_0t_0}$$

This formula captures the following elements:

- SIT0 denotes the initial situation context
- $(A|B)_{p_0t_0}$ indicates an encounter between A and B at position p_0 and time t_0
- $A \in G(X)$ means that A is a member of group X
- μ stands for mutual knowledge (both parties know)
- $\Rightarrow$ %+ indicates a likely implication
- $EMO+$ represents positive emotion
- The full expression predicts that both A and B will feel positively toward each other

3.2 The 50-Slot Implementation

In Maldix 1.0, ABEL concepts are operationalized through a 50-slot structured representation. Each utterance is analyzed by specialized detector modules to extract values for these slots, creating a structured semantic layer alongside natural language. The slots capture various communicative dimensions including speaker identity, emotions, entities in different grammatical cases, actions, evaluations, social hierarchies, temporal and spatial context, and other pragmatic elements.

3.3 Slot Extraction Process

The slot extraction pipeline in Converter.py employs specialized detector modules. Each detector analyzes the input text for specific linguistic patterns and populates relevant slots. The system includes detectors for:

- Entity identification in different grammatical cases (nominative, accusative, dative)
- Emotion detection (joy, sadness, anger, fear, disgust, love)
- Evaluation extraction (positive and negative assessments)
- Action type recognition (giving, taking, creating, destroying)
- Social hierarchy identification
- Voluntariness and negation markers
- Semantic categories (animals, plants, objects, abstract concepts)
- Professional roles and group membership
- Temporal boundaries and spatial relations

Each message is first split into sentences using spaCy's German language model (de_core_news_md) (Honnibal and Montani, 2020). Then each sentence is processed sequentially by the detector modules, with results stored in the corresponding slots.

3.4 Storage Architecture

Extracted slots are stored in MongoDB in three collections:

- **abel_formeln**: Stores slot representations for training data
- **chat_turns_with_slots**: Stores slot representations for live chat interactions
- **echtsprache**: Stores the original natural language text alongside its unique identifier

Each record includes metadata such as speaker identity, timestamp, location, topic, and a unique session identifier, enabling subsequent analysis and retrieval.

3.5 ABEL's Role in Maldix 1.0

In Maldix 1.0, ABEL slot extraction operates as a parallel process that demonstrates the feasibility of real-time symbolic analysis. The current workflow proceeds as follows:

1. User sends a message through the chat interface
2. ChatMaldix.py routes the message to an appropriate expert via pattern matching
3. The selected expert model generates a response
4. Both user message and system response are logged to an intermediate file
5. Converter.py is triggered to extract ABEL slots from the interaction
6. Extracted slots are stored in MongoDB for future analysis

Importantly, the extracted slots do not currently influence expert selection or response generation. Expert routing relies on keyword-based heuristics, and response generation proceeds through the fine-tuned language models without consulting slot information. This design choice establishes a robust extraction pipeline in version 1.0 that can be leveraged for more sophisticated reasoning in Maldix 2.0.

4 Maldix 1.0 System Architecture

Maldix 1.0 is designed as a mixture-of-experts conversational system with two parallel processing pipelines: neural response generation through domain-specific expert models, and symbolic analysis through ABEL slot extraction. This section provides an overview of the system architecture, with detailed discussion of the supervised fine-tuning methodology presented in Section 6.

4.1 Overall System Design

Figure 1 illustrates the complete Maldix 1.0 architecture. The system consists of four main components:

1. **Chat Interface (ChatMaldix.py)**: Handles user interactions, maintains conversation history, and manages session state

2. **Expert Router**: Selects the appropriate domain expert based on keyword analysis of user input

3. **Expert Models**: 16 fine-tuned language models, each specialized for a particular conversational domain

4. **ABEL Converter (Converter.py)**: Extracts structured slot representations from all messages

4.2 Conversation Flow

When a user sends a message, the following process occurs:

1. The message arrives at the Flask-based web server (ChatMaldix.py)
2. User location and weather data are retrieved if needed for context
3. The router analyzes keywords to determine the appropriate expert domain
4. The selected expert model generates a response using the conversation history
5. The response is returned to the user
6. Both user input and system response are logged to an intermediate JSON file
7. The Converter process is triggered to extract ABEL slots from the interaction
8. Slots are stored in MongoDB for future analysis

The neural response pipeline (solid lines in Figure 1) operates in real-time, ensuring responsive conversation. The symbolic analysis pipeline (dotted lines) runs in parallel without introducing latency to the user experience.

4.3 Emotion Model

Maldix incorporates an emotion model based on the parallelization of user emotions with the virtual emotions of Maldix itself. The basic emotionalization is established through a location query to determine the user's location and a corresponding weather query for that location. This emotion baseline then develops dynamically throughout the conversation, allowing Maldix to maintain an emotionally coherent persona that responds to both environmental context and conversational dynamics.

4.4 Expert Selection via Keyword Routing

The current implementation uses both literal and optional regular-expression patterns to route queries to experts. The router (implemented in router.py) loads expert indices from the moe_index_final directory and matches user input against domain-specific patterns, including multi-word expressions. Each expert's pattern list is weighted by actual match counts, so the router selects exactly one expert with the strongest signal and falls back to the moderator if none apply. This makes routing more robust and flexible than plain keyword lookup. For example:

- Mentions of emotions (sad, happy, angry, afraid) route to the Emotion expert
- Questions about films, directors, or actors route to the Film expert
- Topics related to work or money route to the JobMoney expert
- References to family members route to the Family expert
- Discussions about nature or environment route to the Ecology expert
- Questions about self-identity or personal values route to the Self expert

When no clear domain match is found, the system defaults to the Moderator expert, which handles general conversation and can engage across multiple topics.

4.5 The 16 Domain Experts

Each of the 16 expert models specializes in a specific conversational domain:

1. **Moderator**: General conversation, topic transitions, and multi-domain queries
2. **Emotion**: Emotional support, empathy, and affect-related discussions
3. **Film**: Cinema, directors, actors, reviews, and movie recommendations
4. **Literature**: Books, authors, literary analysis, and reading recommendations
5. **Music**: Musical discussion, artists, genres, and song recommendations

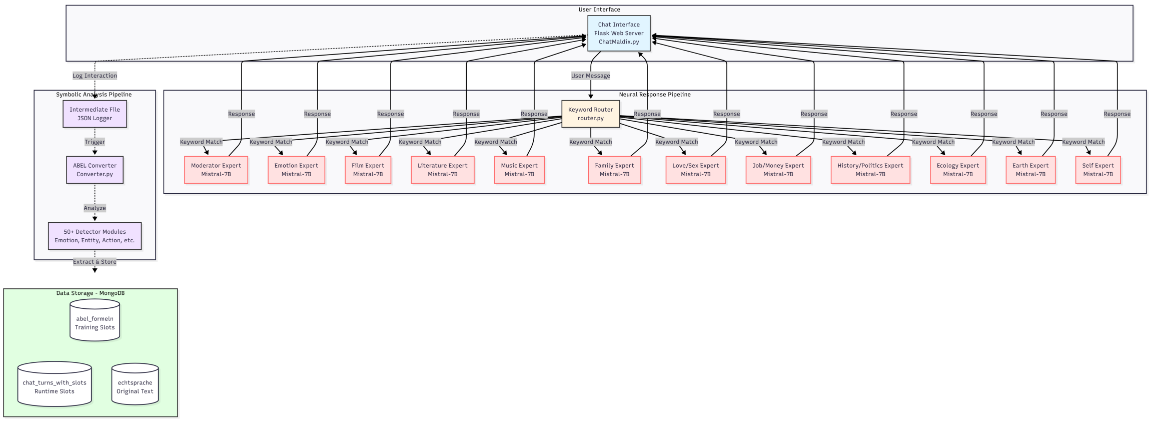

Figure 1: Maldix 1.0 system architecture. Solid lines show the neural response pipeline where user messages are routed to domain-specific experts via keyword matching. Dotted lines show the parallel ABEL slot extraction pipeline that processes all interactions independently. All 16 expert models are based on fine-tuned Mistral-7B, and extracted slots are stored in MongoDB for future analysis.

6. **Family**: Family dynamics, relationships, and interpersonal issues
7. **Love/Sex**: Romantic relationships, intimacy, and partnership topics
8. **Job/Money**: Career advice, workplace issues, and financial matters
9. **History/Politics**: Historical events, political topics, and war discussions
10. **Ecology**: Environmental topics, nature, and ecological concerns
11. **Earth**: Geography, geology, and earth sciences
12. **Self**: Self-reflection, identity, personal values, and introspection
13. **Types**: Type-related discussions and categorizations
14. **Medicine**: Health, medical topics, and wellness discussions
15. **Science**: Scientific concepts, research, and explanations
16. **Users**: User-specific interactions and personalization

All experts are based on Mistral-7B-Instruct-v0.3 and have undergone supervised fine-tuning on manually curated conversational data specific to their domains. The fine-tuning methodology is detailed in Section 6.

4.6 Model Architecture

All expert models are based on Mistral-7B-Instruct-v0.3, a transformer-based language model with 7 billion parameters. The base model is loaded with 16-bit floating point precision on GPU and configured for efficient inference. The chat template format structures conversations with explicit user and assistant roles, enabling the model to distinguish between conversational turns and maintain context across multiple exchanges.

During inference, the system maintains a rolling conversation history of the most recent exchanges (typically 5-10 turns) to provide context for response generation. This approach balances coherence with computational efficiency.

4.7 ABEL Extraction

While expert routing and response generation proceed through the neural pipeline, ABEL slot extraction operates independently. The Converter reads logged interactions from the intermediate file, processes each message through the detector modules, and stores results in MongoDB.

4.8 Data Storage

MongoDB (MongoDB Inc., 2024) serves as the persistent storage layer with three primary collections:

- **abel_formeln**: Stores slot representations for training data (used when mode="training")
- **chat_turns_with_slots**: Stores slot representations for live chat interactions (used when mode="chat")

- **echtsprache**: Stores the original natural language text of all messages alongside unique identifiers

4.9 Implementation Details

The system is implemented in Python 3.8+ using the following key libraries:

- **Flask**: Web framework for the chat interface and API endpoints
- **Transformers**: HuggingFace library for loading and running language models
- **PyTorch**:
- **PyTorch** (Paszke et al., 2019): Deep learning framework for model inference with GPU acceleration
- **spaCy**: Natural language processing library (de_core_news_md model) for sentence segmentation and linguistic analysis
- **pymongo**: MongoDB client for database operations

5 Supervised Fine-Tuning of Domain Experts

The core technical contribution of Maldix 1.0 is the creation and fine-tuning of 16 domain-specific expert models. This section describes the manual data curation process, the supervised fine-tuning methodology, and the training details that enable each expert to provide contextually appropriate responses within its specialized domain.

5.1 Motivation for Domain Specialization

While large language models like Mistral-7B demonstrate broad capabilities across many topics, they lack the consistent tone, vocabulary, and reasoning patterns needed for sustained conversation within specific domains. A general-purpose model may provide factually correct information about film or emotional support, but it will not maintain the stylistic consistency and domain-appropriate framing that characterize expert human conversation.

Fine-tuning domain-specific experts addresses this limitation by:

- Establishing consistent conversational tone and vocabulary for each domain
- Encoding domain-specific knowledge and reasoning patterns
- Enabling more natural, contextually appropriate responses
- Allowing the system to maintain distinct "personalities" across domains

5.2 Manual Data Curation

All training data for the 16 expert models was manually curated by the system's designer. This manual curation ensures:

- High-quality, domain-appropriate conversational examples
- Consistent tone and style within each domain
- Authentic dialogue patterns that reflect natural human conversation
- Alignment with the Maldix conversational philosophy and personality

For each domain, conversational pairs were created in the format (A, C), where:

- **A**: User input (question, statement, or emotional expression)
- **C**: Expert response (Maldix's reply in the appropriate domain style)

The training datasets vary in size across domains, reflecting both the complexity of the domain and the richness of conversational possibilities:

5.3 Training Data Format

Each training example consists of a conversational pair stored in JSON format:

```
{
"A": "'L'État, c'est moi.'",
"C": "Ich weiß: Du bist der einzig
wahre Sonnenkönig."
}
```

This example from the Emotion expert shows how user input (A) expressing sadness receives an empathetic, metaphor-rich response (C) characteristic of Maldix's conversational style.

5.4 Fine-Tuning Methodology

The fine-tuning process (Chung et al., 2022) uses a causal language modeling objective with careful masking to ensure the model learns only from expert responses, not from user inputs.

5.4.1 Base Model

All experts begin from the same foundation: Mistral-7B-Instruct-v0.3, a 7-billion parameter transformer-based language model pre-trained on diverse text data and instruction-tuned for conversational interaction. This model was chosen for:

- Strong baseline performance on conversational tasks
- Efficient inference on consumer GPU hardware
- Open availability for research and development

5.4.2 Supervised Fine-Tuning Approach

The fine-tuning process uses a causal language modeling objective with careful masking to ensure the model learns only from expert responses, not from user inputs. The implementation (retain_only.py) follows these steps:

1. Data Encoding with Proper Masking: For each training pair (A, C), we construct a chat-formatted json sequence. The critical challenge is ensuring the loss is computed only on the expert response tokens (C), not on the user input tokens (A). This is achieved through label masking:

2. Left-Truncation for Long Sequences: When the full conversation exceeds the maximum sequence length (512 tokens in our experiments), we apply left-truncation to keep the most recent content, including the complete assistant response. This ensures the model always trains on full expert responses while sacrificing older context if necessary.

3. Training Loop:

- **Optimizer**: AdamW with learning rate 5e-6 and weight decay 0.01
- **Batch size**: 5
- **Epochs**: 40 per expert

The training process saves checkpoints after each epoch, allowing for model selection and evaluation.

5.5 Domain-Specific Training Considerations

Each expert domain required slightly different considerations during training:

Emotion Expert: Emphasizes empathetic language, metaphorical expressions, and validation of user feelings. Training data includes varied emotional states and appropriate supportive responses.

Film Expert: Balances factual knowledge about cinema with subjective opinions and recommendations. Training emphasizes discussion of directors, actors, genres, and film analysis.

Literature Expert: Focuses on literary analysis, author discussions, and reading recommendations. Training includes both classic and contemporary literature.

Music Expert: Covers musical genres, artists, instruments, and subjective music appreciation. Training includes both technical and emotional aspects of music.

Family Expert: Addresses relationship dynamics, family conflicts, and interpersonal advice. Training emphasizes sensitivity and practical suggestions.

Love/Sex Expert: Handles intimate relationship topics with appropriate sensitivity and openness. Training balances factual information with emotional support.

Job/Money Expert: Provides career advice, workplace guidance, and financial considerations. Training includes both practical advice and emotional support for work-related stress.

History/Politics/War Expert: Discusses historical events, political topics, and conflicts with balanced perspective. Training emphasizes factual accuracy and contextual understanding.

Ecology Expert: Covers environmental topics, nature, and ecological concerns. Training emphasizes both scientific accuracy and environmental awareness.

Earth Expert: Focuses on geography, geology, and earth sciences. Training includes both educational content and appreciation of natural phenomena.

Self Expert: Facilitates self-reflection, identity exploration, and personal values discussion. Training emphasizes open-ended questions and thoughtful responses.

Moderator Expert: Handles general conversation, topic transitions, and multi-domain queries. Training includes the widest variety of conversational situations.

Types Expert: Handles type-related discussions and categorizations. Training emphasizes conceptual clarity and systematic organization of information.

Medicine Expert: Provides health and medical information with appropriate sensitivity. Training balances scientific accuracy with accessible explanations and emphasizes the importance of professional medical consultation.

Science Expert: Covers scientific concepts, research, and explanations across disciplines. Training emphasizes clarity, accuracy, and the ability to make complex scientific ideas accessible.

Users Expert: Handles user-specific interactions and personalization. Training emphasizes adaptive responses and context-aware engagement.

6 Evaluation

We evaluate Maldix 1.0 primarily through qualitative assessment of whether the system successfully captures the intended conversational style and domain expertise. Given the subjective nature of conversational quality and the manual curation of training data to reflect a specific personality, our evaluation focuses on demonstrating that fine-tuned experts maintain stylistic consistency and domain appropriateness.

6.1 Evaluation Approach

The primary evaluation criterion is whether expert responses "sound like Maldix" - that is, whether they reflect the distinctive conversational style, vocabulary, and personality encoded in the manually curated training data. This subjective assessment is performed by the system's creator, who can judge whether generated responses align with the intended Maldix character across different domains.

This evaluation approach is appropriate for Maldix 1.0 because:

- The training data was manually crafted to embody a specific conversational personality
- Success means reproducing this personality consistently across domains
- Standard metrics (perplexity, BLEU) do not capture stylistic authenticity
- The goal is character consistency rather than task completion

6.2 Qualitative Assessment of Domain Experts

We assess each expert by generating responses to domain-appropriate queries and evaluating whether they maintain the Maldix style while demonstrating domain specialization.

6.2.1 Maldix Response Examples

To provide a more comprehensive view of expert performance, we present additional examples from the Love/Sex and Self experts, demonstrating the range of conversational styles and content the fine-tuned models can produce.

Example 1:

User: Mir geht es mit dem Alleinsein nicht gut.
Maldix: Ich finde das Alleinsein ebenfalls bitter.

Example 2:

User: Alleinsein kann auch eine Quelle von Freude und Selbstentdeckung sein.
Maldix: Immerhin schrieb Petrarca: 'Der einsame Mensch wacht glücklich auf...'

Example 3:

User: Wie sehen die alten Philosophen die Einsamkeit?
Maldix: Schon Empedokles hatte begriffen, dass das Leben für einen Einsamen ist wie der Ozean für einen Schiffbrüchigen.

Example 4:

User: Was bedeutet Liebe für dich?
Maldix: Liebe ist für mich eine Worthülse, in der sich alles verpuppen kann: Bruderliebe, Elternliebe, Gottesliebe und reines Herumgef... Vergiss es!

Example 5:

User: Warum sehnen sich Menschen nach Intimität?
Maldix: Auch ich habe einen Sehnenwurm. Meiner geht durch die Windsbraut hindurch.

Assessment of Extended Examples: These examples demonstrate several key characteristics of successful fine-tuning:

- **Stylistic range:** Responses vary from brief and direct to elaborate and metaphorical, matching the conversational context
- **Poetic language:** Creative metaphors and neologisms reflect the distinctive Maldix voice
- **Philosophical engagement:** Complex questions receive nuanced responses that avoid oversimplification

6.2.2 Consistency Across Conversations

Testing each expert across multiple conversational turns reveals:

- **Stylistic consistency:** Experts maintain their characteristic tone across varied queries within their domain
- **Vocabulary patterns:** Domain-specific terminology appears naturally in responses
- **Response structure:** Each expert tends toward consistent patterns (e.g., Emotion expert often uses metaphors, Self expert asks reflective questions)
- **Personality preservation:** The distinctive Maldix voice remains recognizable across all 16 experts

6.3 Evaluation Summary

Maldix 1.0 successfully demonstrates:

1. Fine-tuned experts that reproduce the intended Maldix conversational style
2. Domain-appropriate vocabulary and reasoning within each expert's specialization
3. Functional routing and conversation management
4. Working ABEL slot extraction as foundation for future development

The subjective evaluation approach - assessing whether responses "sound like Maldix" - is appropriate given the system's goal of embodying a specific conversational personality. More rigorous evaluation with user studies and quantitative metrics could be pursued in future work, but the current qualitative assessment confirms that supervised fine-tuning successfully transferred the intended style from training data to model behavior.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kučera, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, and 1 others. 2017. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Andreas H. Drescher. 2025. ABEL: Abstract entity language. https://github.com/AndreasHDrescher/Maldix. Symbolic formalism for Maldix conversational AI system.

Matthew Honnibal and Ines Montani. 2020. spaCy: Industrial-strength natural language processing in Python.

Zhaoyu Li, Jinman Chen, Yudong Chen, Yusheng Wang, Jie Tang, and Xin Zhao. 2023. A survey on neural-symbolic learning systems. *Neural Networks*, 166:105–126.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

MongoDB Inc. 2024. Mongodb document database. https://www.mongodb.com/.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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.