Can LLM Agents Identify Spoken Dialects like a Linguist?

Tobias Bystrich[♡] **Lukas Hamm**[♡] **Maria Hassan**[♡]
Lea Fischbach[♢] **Lucie Flek**[♣♠] **Akbar Karimi**[♣♠]

[♡]Department of Computer Science, University of Bonn, Germany
[♢]Research Center Deutscher Sprachatlas, Marburg University, Germany
[♣]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
[♠]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
s5tobyst@uni-bonn.de ak@bit.uni-bonn.de

Abstract

Due to the scarcity of labeled dialectal speech, audio dialect classification is a challenging task for most languages, including Swiss German. In this work, we explore the ability of large language models (LLMs) as agents in understanding the dialects and whether they can show comparable performance to models such as HuBERT in dialect classification. In addition, we provide an LLM baseline and a human linguist one. Our approach uses phonetic transcriptions produced by ASR systems and combines them with linguistic resources such as dialect feature maps, vowel history, and rules. Our findings indicate that, when linguistic information is provided, the LLM predictions improve. The human baseline shows that automatically generated transcriptions can be beneficial for such classifications, but also presents opportunities for improvement. Code is available on GitHub[1].

Keywords: Spoken dialect classification, German dialects, LLM agents

1. Introduction

The capabilities of Large Language Models (LLMs) have advanced substantially across a wide range of linguistic tasks (Şahin et al., 2020; Srivastava et al., 2023; Chi et al., 2025). However, their performance in low-resource and linguistically grounded tasks, such as spoken dialect identification, remains underexplored. Understanding dialectal variation is important both for linguistic theory, which seeks to explain how language varieties diverge, and for speech technology, where dialect awareness can improve recognition and translation systems. A professional linguist can often identify an unfamiliar dialect by reasoning over linguistic cues or by consulting phonetic transcriptions, which reveal systematic variation through isoglosses on dialect maps (Wolk and Szmrecsanyi, 2018; Tavakoli et al., 2019; Lameli et al., 2020; Lameli, 2022). While some have utilized acoustic modeling (Fischbach et al., 2025a) or text-based methods (Dolev et al., 2024; Peng et al., 2024) for dialect analysis, whether modern LLMs can differentiate between dialects using phonetic representations is under-explored.

In this paper, we investigate whether LLMs can identify Swiss German dialects when provided with automatic phonetic transcriptions of dialectal speech as textual input. Specifically, we ask: *Can LLMs and LLM agents use phonetic patterns to infer dialect identity, and how does this capability compare to that of human linguists?* We use a corpus of automatically transcribed dialectal recordings (Dogan-Schönberger et al., 2021; Plüss et al., 2023) and evaluate the models on dialect classification. We find that, for binary classification of dialects, fine-tuning an encoder-based model still outperforms the LLM approaches. However, the agentic approach considerably improves upon a single LLM model.

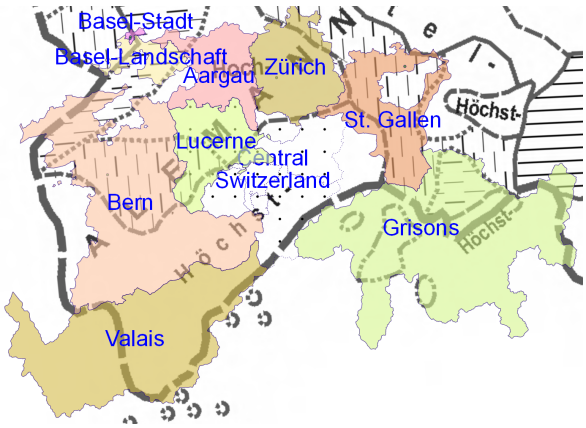

Figure 1: Graphic for SwissDial classes and Innerschweiz made with REDE-SprachGIS and dialect region background (Besch et al., 2008).

Our contributions are: (1) a manual examination of LLM dialect identification using phonetic transcriptions, uncovering how models justify dialect distinctions; (2) an encoder-based model baseline for dialect classification to compare performance with a large language model; (3) comparing the linguistic capabilities of GPT-4o mini and GPT-5, high-

[1] https://github.com/caisa-lab/dialect-classification-with-llm-agents

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 83–92
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

lighting their respective strengths and limitations. These contributions shed light on the limits of LLMs in dialect identification and how far current models approximate human experts when grounded in phonetic representations.

2. Related Work

Large language models (LLMs) have recently demonstrated strong capabilities in linguistic tasks. Foundation models such as GPT-4 (OpenAI, 2023) and multi-agent frameworks like LangGraph (Duan and Wang, 2024) have been used to implement reasoning pipelines, incorporating techniques such as chain-of-thought prompting (Wei et al., 2022) and tool-using agents (Schick et al., 2023). These approaches have shown that LLMs can engage in stepwise inference and apply symbolic reasoning to linguistic tasks, including syntactic analysis, translation, and semantic inference. Despite these advances, the capabilities of LLMs have not been systematically examined in the context of spoken dialect identification. Works such as Wolk and Szmrecsanyi (2018), Tavakoli et al. (2019), Lameli et al. (2020), and Lameli (2022) demonstrate how phonetic transcriptions and dialect maps reveal systematic regional variation through isoglosses, enabling expert linguists to infer dialect affiliation from phonological cues. Fischbach et al. (2025a) investigate acoustic modeling while Dolev et al. (2024) and Peng et al. (2024) utilize text-based methods for dialect classification. The International Phonetic Alphabet (IPA) (Association, 1999) encodes fine-grained pronunciation details and has been widely used in phonetic automatic speech recognition (ASR) models. A number of studies have explored phonetic ASR systems (Xu et al., 2022; Li et al., 2020). Taguchi et al. (2023) introduce MultIPA, a language-agnostic transcription model trained on carefully selected languages to mitigate irregular correspondences between the orthography and the phones. Evaluation of phonetic ASR typically relies on phone error rate (PER), complemented by feature-based and linguistic evaluations (Bystrich, 2025).

Encoder-based speech models form a complementary line of work. For instance, HuBERT (Hsu et al., 2021) performs strongly on dialect classification (Sullivan et al., 2023) and related ASR and phone(me) recognition tasks (Yang et al., 2021), but these models capture primarily acoustic similarity rather than linguistic knowledge. As a result, we explore the capabilities of LLMs in spoken dialect identification from phonetic representations, offering a comparison between encoder models, generative models and a human linguist.

3. Dataset Preparation

While the LLM agent evaluation requires an evaluation set, the HuBERT model needs both fine-tuning and evaluation sets. Therefore, after introducing our utilized datasets in this section, we describe the preparation process.

3.1. Corpora

We use two Swiss German datasets: **SwissDial** (Dogan-Schönberger et al., 2021) and **STT4SG-350** (Plüss et al., 2023) referred to hereafter as **STT**. **SwissDial** contains recordings and manual transcriptions in vernacular orthography for eight dialects: Aargau (AG), Bern (BE), Basel (BS), Grisons/Graubünden (GR), Lucerne/Luzern (LU), St. Gallen (SG), Valais/Wallis (VS), and Zürich (ZH). Each dialect is represented by one speaker who translated sentences from Standard German into the local dialect. The material covers various topics, including news, Wikipedia articles, weather reports, and short stories. **STT** is a substantially larger dataset comprising approximately 343 hours of recordings from 316 speakers across seven dialect regions. Speakers were instructed to read Standard German sentences and render them in their own dialect. The dataset includes metadata on speaker age, gender, and origin and was used under license for this study. While STT provides a broad and demographically diverse sample, making it more representative overall, SwissDial is smaller, with only one speaker per dialect, but it offers richer orthographic details. Together, the two datasets cover a wide range of dialectal and topical variation, including domains such as politics, news, science, and everyday life.

3.2. Label Creation for STT Dataset

To make the classes more consistent across both datasets, we mapped the labels from STT to the 8 SwissDial labels. While Aargau, Lucerne and St. Gallen are not among the dialect region labels in STT, we approximated these by using the existing dialect regions in conjunction with the canton (see Figure 1). We approximated Aargau by using the dialect regions of Zürich and Bern in conjunction with the canton Aargau. We expect that this class cannot easily be predicted since multiple dialect regions occur in the canton Aargau. For the class Luzern/Lucerne, a simpler and likely accurate approximation was possible. The Lucerne dialects are grouped among the STT label ”Innerschweiz” and can further be limited to the canton Lucerne. For St. Gallen, we took the segments labeled with ”Ostschweiz” and the canton St. Gallen.

# Train	Accuracy	# Class predictions		Macro-F1	Accuracy per class	
		High	Highest		High	Highest
400	66.25%	13	67	61.91%	32.5%	100%
4000	**66.25%**	39	41	**66.25%**	65%	67.5%

Table 1: HuBERT baseline for the binary task with best hyperparameters on test data from SwissDial

# Train	Accuracy	Most / least class predictions		Macro-F1
800	11.25%	SG (35)	BE (0)	8.67%
8000	**26.25%**	BS & GR (20)	LU (2)	**22.85%**

Table 2: HuBERT baseline for the 8-class task with best hyperparameters on test data from SwissDial. Abbreviations for dialects: St. Gallen (SG), Bern (BE), Basel (BS), Grisons/Graubünden (GR), Lucerne/Luzern (LU).

Setting	AG	BE	BS	GR	LU	SG	VS	ZH
8-class (800)	40%	0%	10%	0%	0%	30%	10%	0%
8-class (8000)	0%	0%	70%	50%	10%	40%	20%	20%

Table 3: Accuracy per class for HuBERT (8-class) with best hyperparameter configuration on test data from SwissDial. Abbreviations for dialects: Aargau (AG), Bern (BE), Basel (BS), Grisons/Graubünden (GR), Lucerne/Luzern (LU), St. Gallen (SG), Valais/Wallis (VS), and Zürich (ZH).

3.3. Dataset Splits

To prepare the training and test data, all audio files were converted to MP3 format with a sampling rate of 16 kHz. The data were then sampled and divided into three splits each for training, validation, and testing, ensuring that each split contained an equal number of instances for all labels. Following preliminary experiments, all subsequent training sets (for fine-tuning HuBERT) were drawn exclusively from STT, while test and validation sets were created for both STT and SwissDial. Since STT already provides three independent splits, we sampled from these to construct our training, validation, and test sets. For SwissDial, which is a parallel corpus, we ensured that no sentence overlap occurred across the different splits. Because hyperparameter searches and overfitting control are computationally demanding, we prepared both small and large training sets for the 8-class and 2-class classification tasks. For the 8-class task, training set sizes were 800 (small) and 8000 (large); for the 2-class task, 400 and 4000 samples were used, respectively. The small sets were employed for hyperparameter tuning, workflow development, and stability testing, while the large sets were each used for a single full training run. Test splits were fixed to 80 segments due to the high inference time required by the agent. Validation sets followed the same procedure. Although the STT4SG-350 dataset includes speakers from Aargau, they are distributed across multiple broader dialect regions (Basel, Berne, Central Switzerland, and Zurich) rather than being categorized as a standalone set. To specifically evaluate this dialect, the validation set was limited to 80 segments, and 10 additional segments identified as Aargau-origin were sampled from the original training data to ensure a representative sample.

3.4. Classification Setup

Since the eight classes in the SwissDial dataset (8-class problem) were found to be too challenging and not all linguistically defined, we define a 2-class problem. Results for the binary and 8-class problems are shown using the HuBERT baseline in Tables 1 and 2 as motivation for designing a 2-class problem. As we can see in Table 2, the overall performance for the 8-class problem is way below the binary one. Considering the individual classes in Table 3, we can see that many classes are not recognized by this model.

3.5. Conversion into 2-class Problem

Faced with lower accuracies and higher overfitting for this problem than expected, we developed a simpler classification problem. Previous work such as Fischbach et al. (2025b) excluded dialect transition areas and predicted the traditional German dialect regions using a Wiesinger-based map (Besch et al., 2008) as a reference. We take this as a justification for using the dialect regions from Besch et al. (2008), High Alemannic and Highest Alemannic,

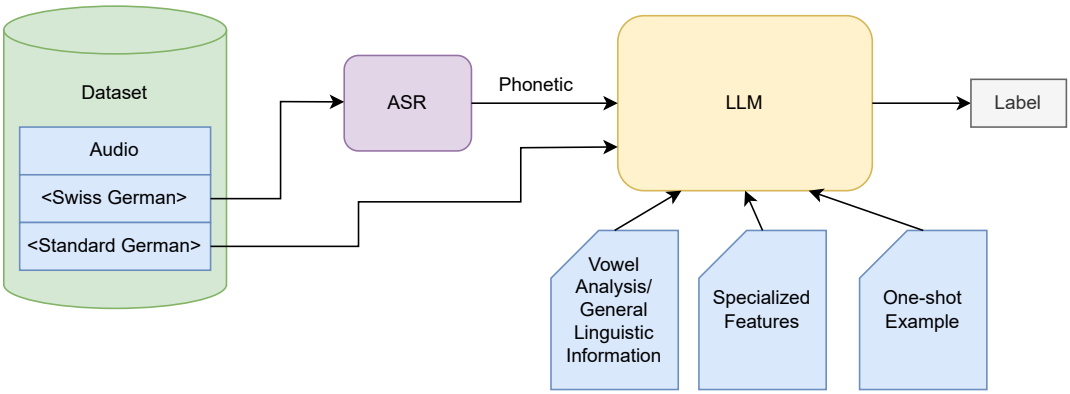

Figure 2: The agentic framework for dialect analysis

and removing transition areas such as Low Alemannic and parts of High and Highest Alemannic.

We used the REDE project SprachGIS[1] to determine the exclusions and mappings to the 2-class problem. Therefore, the new class High Alemannic can only contain Zürich, Aargau and Lucerne. Aargau contains Bernese dialect, however, we excluded Bern as a political unit due to its large overlap with Highest Alemannic (see Figure 1). The canton Lucerne, having some overlap with a transition area and minor overlap with Highest Alemannic, is the least neatly mapped inclusion. We decided in favor of its inclusion since it seems to have considerably smaller overlap than the excluded classes and since not including it might limit its ability to actually classify High Alemannic given a high bias towards Zürich.

For Highest Alemannic, the overlap of Grisons/-Graubünden with the High Alemannic region made its exclusion important (see Figure 1). This only left Valais for SwissDial. For STT, increasing the variance was possible by using Innerschweiz, limited to cantons totally situated in the Highest Alemannic region. We validated this choice by comparing the performance on an STT vs. a SwissDial test set (see Section 5).

All 2-class datasets have an equal number of segments for High and Highest Alemannic, while the source classes (Aargau, Lucerne, Zürich; Valais, Innerschweiz_Highest) get an equal number within their class, barring a difference of 1 due to discrete numbers. Innerschweiz_Highest is not used for the 8-class problem.

4. Methodology

Figure 2 shows the design for our dialect agent. In this section, we describe its components.

4.1. ASR Model

We used the XLSR-53 version[2] of Wav2Vec2Phoneme (Xu et al., 2022) for phonetic transcription. The reported phone error rate (PER) for zero-shot transcription without an LLM is 33.3% (Xu et al., 2022). To answer how well a human linguist can classify dialects based on these transcriptions, we perform an evaluation with human baseline in Section 6.

4.2. Base Prompt and Query

Base Prompt Box

You are now a linguist who needs to identify dialects based on feature descriptions and observations. As a linguist, linguistic reasoning is far more important than coding, coding is likely unnecessary. You are provided transcriptions in the IPA that were generated automatically as well as translations into Standard German to help you interpret the dialect transcription. Please consider the topics with a focus of a linguist, a dialectologist and a phoneticist. Vowels and consonants must always be considered in their word. It makes no sense to just check for the presence of specific phones. Instead, they must be considered in relation to the morpheme in question. **[THIS PART IS FOR BINARY PROBLEM]** Your first task is to identify a Swiss German dialect into one of two dialect regions: High Alemannic, Highest Alemannic. Please output as your final reply only the name of the dialect region. **[THIS PART IS FOR 8-CLASS PROBLEM]** Your first task is to identify a dialect as one of several Swiss German dialects. Please output your reply as the two letter short form (ag, be, bs, gr, lu, sg, vs, zh) for (Aargau, Bern, Basel, Grisons/Graubünden, Lucerne/Luzern, St. Gallen, Valais/Wallis, Zürich).

[1] https://regionalsprache.de

[2] https://huggingface.co/docs/transformers/model_doc/wav2vec2_phoneme

[v eː r h aɪ t d iː v oː n n ɑː b ɔ t l ə r
a n iː r ə ʃ l ɛ t ʃ t ə a r b aɪ t s t ɔ k
p s y ɛ x t ʊ n ə f iː r ɪ ʃ p ʊ l iː t
ɪ ʃ l aʊ f b a n ts r uː k l ɛ k t]

Figure 3: Sample transcription with Highest Alemannic features. Orthographic transcription: *Wir haben Yvonne Beutler an ihrem letzten Arbeitstag besucht und auf ihre politische Laufbahn zurückgeschaut* [*-gelugt*]

The base prompts for both problems can be found in Base Prompt Box. The boldfaced sentences are not part of the prompt, but only to show which part was used for which problem. The query we provide for the model for classification is preceded by "[USER]" and consists of the phonetic transcription in the IPA and a Standard German translation. Figure 3 shows a sample transcription.

4.3. Linguistic Information

As input, we provide the model with linguistic information to enable it to go beyond its pretrained knowledge. Some of this information include the dialect features derived from maps since the model might not be able to recognize distinguishing features of Swiss German dialect regions from IPA transcriptions. In addition, we feed the model an explanation of the linguistic information in plain English. To further improve the accuracy, we also provide historical vowel information, since vowel sound changes are highly important. This is provided as a table. We give the model IPA charts for vowels and consonants, to help achieve a more definition-consistent IPA interpretation of the model. Finally, we provide the model with a sample reference evaluation so that it can learn from the thought process of a linguist.

4.4. LLM

We used OpenAI's GPT-4o mini via API. We set the temperature to zero to maximize truthful outputs and minimize unpredictable factors. Since non-mini GPT versions by OpenAI such as GPT-5 could not be used via API, we randomly compared some outputs via ChatGPT with GPT-5 as the backbone.

4.5. LangGraph

We used a LangGraph agent for our dialect classification task. It uses the same linguistic information and prompt but breaks it down into different nodes. Our system consists of 2-3 nodes where each of them uses a different prompt to send requests to the LLM (GPT-4o mini). We structured the LangGraph agent using multiple **nodes**. The node **vowel and consonant analysis** focuses on vowel- and consonant-related cues and returns an object with per-class confidence scores and brief reasoning. And the **specialized features analysis** considers broader phonological features and returns an object with class probabilities and a final prediction. The state object passed between nodes includes inputs (audio filename, ASR transcription, Standard German transcription), intermediate analysis (vowel analysis and dialect features analysis results).

5. Experiments and Results

We tested all of the models with balanced 80 test samples from SwissDial. The HuBERT Baseline was trained and validated on STT to avoid overfitting. We used a balanced small training split of 400 segments for the hyperparameter search and then trained with a balanced training split of 4000 segments using the same hyperparameters. As an additional validation for HuBERT, we additionally tested it on a STT split to compare the difference to the SwissDial data.

5.1. Baselines

LLM Baseline Our LLM baseline is similar to Figure 2 without the additional linguistic information, with only the data and the phonetics from the ASR model as inputs. It has the same model and base prompt as the main dialect agent, intended as a fair comparison with the encoder-based classifier and to show whether our additional information can improve the LLM's performance.

HuBERT HuBERT uses a training approach similar to that of BERT (Devlin et al., 2019). The raw audio signal is first segmented into short frames of approximately 20–25 ms. These frames are then represented by Mel-frequency Cepstral Coefficients (MFCC) features, which are clustered via k-means into up to 500 groups. The resulting cluster assignments serve as pseudo-labels (cluster IDs) for self-supervised pre-training.

During training, the audio frames are processed by a convolutional waveform encoder, producing latent representations. After parts of these representations are masked, they are fed into a Transformer network (Vaswani et al., 2017). The Transformer is trained to predict the original cluster IDs of the masked segments, thereby learning robust and context-aware speech representations.

For classification, we used a two-layer head, consisting of 128 units in the first GELU layer and

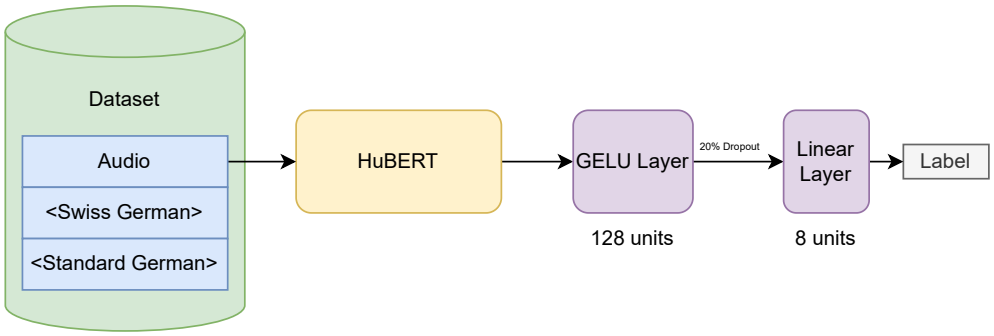

Figure 4: HuBERT baseline architecture

lr	Batch	Accuracy	# Class predictions		Macro-F1	Accuracy per class	
			# High	# Highest		High	Highest
10^{-3}	8	52.5%	61	19	50.99%	70%	35%
	16	46.25%	61	19	42.27%	72.5%	20%
	32	56.25%	25	55	54.66%	37.5%	75%
	64	62.5%	22	58	60.5%	40%	85%
10^{-4}	8	50%	76	4	37.3%	95%	5%
	16	53.75%	67	13	47.8%	87.5%	20%
	32	47.5%	70	10	38.9%	85%	10%
	64	**66.25%**	**13**	**67**	**61.91%**	**32.5%**	**100%**

Table 4: Results of hyperparameter search for the HuBERT baseline on test data from SwissDial.

Experiment Configuration	**Accuracy (%)**
Base Prompt Only (2 runs average)	47.8
Specialized Features Analysis Node	47.5
Vowel & Consonant Analysis Node	55.0
Vowel & Consonant Node + Specialized Features Node (2 runs average, no IPA explanations from GPT)	**58.0**

Table 5: Accuracy of different LangGraph experiment configurations. We use the best results for comparison with other methods.

8 units in the final linear output layer. Additionally, a dropout rate of 0.2 was applied between the first and last classifier-head layer. The number of training epochs was determined using early stopping. We set a minimum threshold of 3000 training examples, translated into the corresponding number of epochs, to allow for initial mistakes in the early stages. Beyond this threshold, training was stopped once the development loss exceeded the lowest dev loss observed across all epochs. Figure 4 shows the diagram for the HuBERT baseline.

We used the HuBERT model to compare the agentic approach to more conventional speech classification approaches. Using the multilingual version mHuBERT-147 (Boito et al., 2024) ensured that Germanic languages were included during pretraining. After doing a hyperparameter search shown in Table 4, we choose the best results for comparison with other methods (lr = $1e{-}4$, batch size = 64).

Human Linguist The last binary test set of 80 segments was analyzed by one of the authors with linguistic experience based on the same linguistic information that the LLM is provided with and the same auto-generated transcriptions. They used only the same linguistic information, aiming to avoid any personal opinion. If the evidence did not suffice for determining a class, the prediction was counted as neither correct nor incorrect to avoid any biases.

5.2. LangGraph Findings

We investigated various LangGraph configurations to study the impact of prompt engineering and amount of information provided. Experiments revealed that using only the base prompts gives an average accuracy of 47.8 percent (Table 5). However, when the additional vowel & consonant and specialized features nodes are added to the setup, accuracy improves to 58% on average and 62.5% for the best run. A prediction imbalance was also observed which was that for an 80 sample dataset with an even 40/40 split for a 2-class problem, the

Model	Accuracy	# Class predictions		Macro-F1	Accuracy per class	
		High	Highest		High	Highest
HuBERT	66.3%	39	41	66.3%	65%	67.5%
LLM	47.8%	30	50	47%	35.6%	60%
Human	72.5%	33	46	72.3%	80%	65%
Agentic	58%	23	57	56.3%	37.5%	78.8%

Table 6: Results for the 2-class problem

Manual Transription	**ChatGPT (GPT-5)**	**GPT 4o-mini**
[ ] Die [overlaps with next word]	[d oː] Die	[d i] Die
[d oː t r ʊ] da[d]raus [overlaps with next word]	[t r ʊ] daraus	[d aː r ʊ] daraus
[ʃ l ɪ p ə n d] schlüpfenden	[ʃ l ɪ p ə n d] schlüpfenden	[ʃ l ʊ p f ə n d] schlüpfenden
[r aʊ p ɑː] Raupen	[r aʊ p ɑː] Raupen	[r aʊ p ɑː] Raupen
[ɛ r m ɛ r ə n] ernähren	[ɛ r m ɛ r ə n] ernähren	[ɛ r n ɛː r ə n] ernähren
[z ɪ x] sich	[z ɪ x] sich	[z ɪ x] sich
[f oː d ɛ r ɜ] von der	[f oː d ɛ r] von der	[f oː n] von
		[d iː s ə] dieser
[p f l a n ts ə] Pf anze	[ɜ p f l a n ts ə] Pf anze	[p f l a n ts ə] Pf anze

Figure 5: Alignments between human baseline and GPT models

model predicted 57 instances of ”Highest” and only 23 instances of ”High”.

5.3. Overall Comparison

Table 6 shows the results for the baseline models, agentic framework, and the human baseline for the 2-class problem. The HuBERT baseline model gives an accuracy and macro-F1 of 66.3% for the 2-class problem. Compared to the other classifiers, HuBERT achieved the best macro-F1 and the second-best accuracy after the human baseline, which makes it the highest-achieving non-human model in this metric. The most and least predicted classes show well-distributed predictions (41/39) that are close to the original 50/50 distribution of the ground truth test set. For the LLM-based approaches, two runs were averaged with the standard deviation of 4.1 percentage points for the LLM baseline and 4.4 for the LangGraph agent. The agentic approach shows a higher performance than the LLM baseline, achieving an improvement of 10.2 percentage points. The human baseline achieved the highest overall accuracies.

6. Discussion on Linguistic Approaches

This section discusses the results for the agent, the human baseline and the LLM baseline. The LLM baseline shows results around 50%, suggesting it does not have adequate previous knowledge for this task. The reported performance is the average of two test runs. For the human linguist baseline, the analysis of the 80 SwissDial segments provided enough evidence for a decision in 58 cases. For these cases, the human accuracy was 81%. There were 12 ambiguous segments for Highest Alemannic and 10 for High Alemannic. As a result, for the calculation of these predictions as neutral (to avoid any biases), we assigned half of each with the correct and half of each with the incorrect value. This led to an overall accuracy of 72.5%. One could also allow the LLM to have a third option to abstain from choosing either class. However, early tests showed that this option may lead to a worse performance.

This result serves as our extrinsic evaluation for the ASR model, showing that even with the sub-optimal phone error rate, feeding only one sentence at a time can still allow for a classification of High and Highest Alemannic. Improving the ASR quality would likely further enhance the human accuracy.

6.1. Comparative Analysis with GPT-5

ChatGPT showed more robustness concerning the work with phonetic transcription, as we will show using an example. The alignment of phones to Standard German is an important step for the linguistic analysis since the analysis of the sound changes and morphological information depends on the phones being aligned to the correct etymons. It can be summarized that GPT-4o mini fails in ways that are difficult to conceive for a human, even in

normal cases, while ChatGPT showed surprising skill even in challenging situations (see Figure 5). Here we can see that GPT-5 aligned the words almost perfectly. It was able to handle this challenging case with multiple overlapping words quite well. However, GPT-4o mini hallucinated the transcriptions.

6.2. Other Challenges

The dialect recognition with linguistic resources was likely affected by ASR performance, however, the results of the extrinsic evaluation by a human surpass the HuBERT baseline. Additionally, Standard German loan words into Swiss German made the classification more difficult since they don't fit the dialect's sound changes. Lastly, some sentences may not contain any words with the potential to identify a dialect without ambiguity.

7. Conclusion

We implemented an LLM agent to see if it can utilize linguist information about dialects as context to classify dialects (of Swiss German) given phonetic transcriptions and translation into the literary language (Standard German). We focused on dialect regions, instead of towns' dialects, and excluded dialect transition areas. Implementing a LangGraph and GPT-4o-mini-based agent showed a considerable improvement over a baseline LLM. As a result, we conclude that LLM agents can, to some extent, use linguistic information and phonetic transcriptions to improve their dialect understanding. However, given the difficulty of the task as well as the quality of the ASR models, they would need more supervision to show human-level capabilities in dialect analysis.

8. Limitations

Given our limited budget, we used GPT-4o mini in our experiments. While it shows considerable capacity for our task, one could also use more capable models such as GPT-4 or GPT-5 as well as other model families such as Gemini (Team et al., 2023) and Llama (Touvron et al., 2023). In addition, ineffective tool use of the agent necessitated more focus on prompt engineering, something that can be investigated more for an improved framework. Finally, with the availability of more domain experts, one could increase the number of annotated test samples to perform a more comprehensive analysis of our system.

9. Ethical Considerations

We ensured that the data is used responsibly. Speaker identities were anonymized in the released datasets. Potential misuse of speaker identification technology could threaten privacy; we recommend deployment only in contexts with explicit user consent and appropriate legal frameworks.

Acknowledgments

This research is supported by the Academy of Science and Literature Mainz (grant REDE 0404), the German Federal Ministry of Education and Research (BMFTR) (grant AnDy 16DKWN007) and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B and the Research Center Deutscher Sprachatlas Marburg.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Werner Besch, Ulrich Knoop, Wolfgang Putschke, and Herbert E. Wiegand. 2008. *Dialektologie. 2. Halbband*. De Gruyter.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model.

Tobias Bystrich. 2025. Multilingual Automatic Phonetic Transcription – a Linguistic Investigation of its Performance on German and Approaches to Improving the State of the Art. Master's thesis, University of Bonn.

Nathan Andrew Chi, Teodor Malchev, Riley Kong, Ryan Andrew Chi, Lucas Huang, Ethan A Chi, R Thomas McCoy, and Dragomir Radev. 2025. ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 105–114.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Pelin Dogan-Schönberger and Julian Mäder and Thomas Hofmann. 2021. *SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German*.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli. 2024. Does Whisper Understand Swiss German? An Automatic, Qualitative, and Human Evaluation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 28–40.

Zhihua Duan and Jialin Wang. 2024. Exploration of LLM Multi-Agent Application Implementation Based on LangGraph+CrewAI.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025a. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Proc. Interspeech 2025*, pages 2780–2784.

Lea Fischbach, Caroline Kleen, Lucie Flek, and Alfred Lameli. 2025b. Does Preprocessing Matter? An Analysis of Acoustic Feature Importance in Deep Learning for Dialect Classification.

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.

Alfred Lameli. 2022. Syllable Structure Spatially Distributed: Patterns of Monosyllables in German Dialects. *Journal of Germanic Linguistics*, 34(3):241–287.

Alfred Lameli, Elvira Glaser, and Philipp Stöckle. 2020. Drawing Areal Information from A Corpus of Noisy Dialect Data. *Journal of Linguistic Geography*, 8(1):31–48.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, and Florian Metze Alan W Black. 2020. Universal Phone Recognition with a Multilingual Allophone System.

OpenAI. 2023. GPT-4 Technical Report.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! Recognizing Named Entities in Bavarian Dialectal Data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493.

Michel Plüss, Jan Milan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardzic, Manfred Vogel, et al. 2023. Stt4sg-350: A speech corpus for all swiss german dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the Robustness of Arabic Speech Dialect Identification.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal Automatic Phonetic Transcription into the International Phonetic Alphabet.

Shahin Tavakoli, Davide Pigoli, John AD Aston, and John S Coleman. 2019. A Spatial Modeling Approach for Linguistic Object Data: Analyzing Dialect Sound Variations Across Great Britain. *Journal of the American Statistical Association*, 114(527):1081–1096.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.

Christoph Wolk and Benedikt Szmrecsanyi. 2018. Probabilistic Corpus-based Dialectometry. *Journal of Linguistic Geography*, 6(1):56–75.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I. Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech processing Universal PERformance Benchmark.