VarDial Evaluation Campaign 2024: Commonsense Reasoning in Dialects and Multi-Label Similar Language Identification

Adrian-Gabriel Chifu[1], **Goran Glavaš**[2], **Radu Tudor Ionescu**[3], **Nikola Ljubešić**[4,5],
Aleksandra Miletić[6], **Filip Miletić**[7], **Yves Scherrer**[6,8], **Ivan Vulić**[9]
[1]Aix-Marseille University, [2]University of Würzburg, [3]University of Bucharest,
[4]Jožef Stefan Institute, [5]University of Ljubljana, [6]University of Helsinki,
[7]University of Stuttgart, [8]University of Oslo, [9]University of Cambridge

Abstract

This report presents the results of the shared tasks organized as part of the VarDial Evaluation Campaign 2024. The campaign is part of the eleventh workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with NAACL 2024. Two shared tasks were included this year: dialectal causal commonsense reasoning (DIALECT-COPA), and Multi-label classification of similar languages (DSL-ML). Both tasks were organized for the first time this year, but DSL-ML partially overlaps with the DSL-TL task organized in 2023.

1 Introduction

The workshop series on *NLP for Similar Languages, Varieties and Dialects* (VarDial), traditionally co-located with international conferences, has reached its eleventh edition. Since the first edition, VarDial has hosted shared tasks on various topics such as language and dialect identification, morphosyntactic tagging, question answering, and cross-lingual dependency parsing. The shared tasks have featured many languages and dialects from different families and data from various sources, genres, and domains (Aepli et al., 2023, 2022; Chakravarthi et al., 2021; Gaman et al., 2020; Zampieri et al., 2019, 2018, 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014).

As part of the VarDial Evaluation Campaign 2024, we offered two shared tasks which we present in this paper:

- **DIALECT-COPA:** Dialectal causal commonsense reasoning[1]
- **DSL-ML:** Multi-label classification of similar languages[2]

DSL-ML continues the long line of language and dialect identification (Jauhiainen et al., 2019) shared tasks at VarDial, whereas DIALECT-COPA features a task novel to the evaluation campaigns.

The evaluation campaign took place in January – March 2024. The call for participation and the training data sets for the shared tasks were published in the second half of January, and the results were due to be submitted on March 11th.[3]

In the following sections, the two tasks are discussed in detail, focusing on the data, the participants' approaches, and the obtained results. Section 2 is dedicated to DIALECT-COPA and Section 3 to DSL-ML.

2 The DIALECT-COPA Task on Causal Commonsense Reasoning

2.1 Motivation

The causal commonsense reasoning (CCR) task has been established as an important task in evaluation of natural language understanding (NLU) capabilities of pretrained language models, including the latest family of the so-called Large Language Models (LLMs). The original English dataset, Choice Of Plausible Alternatives (COPA) (Roemmele et al., 2011) has been used as the standard evaluation benchmark for the English CCR task since its release, and it is also included in the English SuperGLUE benchmark (Wang et al., 2019).

Language-specific variants of COPA have also been created, where the bulk of the data is covered in the multilingual XCOPA dataset (Ponti et al., 2020). The original XCOPA covers 11 standard language varieties from 11 language families, including some lower-resource languages such as Haitian Creole, Tamil, and Southern Quechua. It has been included into the established XTREME-R benchmark (Ruder et al., 2021) for the evalua-

[1] Task organizers: Nikola Ljubešić, Ivan Vulić, Goran Glavaš.

[2] Task organizers: Adrian Chifu, Radu Ionescu, Aleksandra Miletić, Filip Miletić, Yves Scherrer.

[3] https://sites.google.com/view/vardial-2024/shared-tasks

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 1–15
June 20, 2024

tion of cross-lingual transfer, and has consequently been used as a de facto evaluation benchmark for CCR in cross-lingual and multilingual scenarios. Besides XCOPA, there also exist single-language translations of adaptations of COPA into other languages such as Slovenian (Žagar and Robnik-Šikonja, 2022), Russian (Shavrina et al., 2020), and Catalan,[4] among others.

While COPA and XCOPA were considered challenging benchmarks for previous encoder-style models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and XLM-R (Conneau et al., 2020), current state-of-the-art LLMs now provide impressive performance on these datasets (Chowdhery et al., 2023; Zhong et al., 2022; Shi et al., 2023): they are able to reach $\geq$ 90% accuracy for diverse languages such as Thai, Estonian, Indonesian, Tamil, Vietnamese or Turkish (Shi et al., 2023). Whereas LLMs have been proven to perform extremely well on high-resource and even moderately resourced standard languages, their ability to conduct CCR for truly low-resource languages (Senel et al., 2024) and *especially dialects* (Joshi et al., 2024) has been much less investigated and empirically measured. For instance, lower performance on the standard lower-resource languages of the XCOPA dataset (e.g., Haitian Creole, Quechua, Swahili) already indicates additional difficulty for and reduced capability of current LLMs.

All COPA datasets to date comprise the same set of instances covering the same or similar set of topics. The only core difference between different datasets is the actual, target language variety of a particular dataset. Another property of COPA and its derivatives is its simple and easy-to-evaluate data format. In a nutshell, each data instance consists of three sentences: a statement (*premise*) and two possible *effects* or *causes* (termed *alternatives*) for the premise. Given an English example, a premise *'The man turned on the faucet.'* is combined with two alternatives *'The toilet filled with water.'* and *'Water flowed from the spout'*. The task is then to select the alternative that more plausibly has a causal relation with the premise, where each instance is manually annotated with a correct answer. The standard evaluation measure is accuracy, where the random baseline is therefore at 50% accuracy, and errors made by the systems could be due to subtle details related to understanding causality relationships.

The above background related to CCR in general and COPA-style datasets in particular has motivated us to create a first shared task on CCR for *dialectal data*, DIALECT-COPA, which we discuss next. In summary, the selection of the task has been guided by the following observations and criteria:

- CCR is an established and important NLU task for the evaluation of language models in monolingual, multilingual, and cross-lingual setups;

- CCR has never been in focus of VarDial evaluation campaigns and, vice versa, there have been no attempts to date to extend the CCR task and the corresponding COPA-style data to non-standard language varieties and dialects;

- CCR based on the standard COPA data format offers an excellent balance between the structural simplicity and semantic complexity of the task, with clear and straightforward evaluation protocols and measures.

- The standardized COPA format and the multi-parallel nature of COPA-based datasets in different standard language varieties combined with newly created dialectal COPA variants offer ample opportunity for cross-linguistic and cross-dialectal analyses and studies of model behavior and performance, as part of the shared task as well as for future research.

- For dialects chosen for DIALECT-COPA, obtaining large quantities of raw text is typically not possible, which renders good out-of-the-box performance of LLMs for them difficult and unlikely; this calls for new and creative approaches in order to mitigate the current gaps of LLMs when faced with CCR on dialectal data.

2.2 Data

The focus of the first DIALECT-COPA shared task has been on *micro-dialects* of several South-Slavic languages. This choice has been partially motivated by the recent creation of COPA datasets for standard language varieties of several, moderately resourced in NLP terms, South-Slavic languages: Slovenian COPA-SL (Žagar and Robnik-Šikonja, 2022), Croatian COPA-HR (Ljubešić, 2021), Serbian COPA-SR (Ljubešić et al., 2022b) and Macedonian COPA-MK (Ljubešić et al., 2022a). All the datasets were translated by human translators,

[4] https://huggingface.co/datasets/projecte-aina/COPA-ca

native speakers of the target languages, from the English COPA dataset (Roemmele et al., 2011), with all the datasets, except for COPA-SL, following the XCOPA translation and adaptation methodology (Ponti et al., 2020). COPA-SL was translated without any additional adaptation as part of the Slovenian SuperGLUE benchmark (Žagar and Robnik-Šikonja, 2022). Serbian and Macedonian datasets are written in Cyrillic, while the other data are in the Latin script.

For the shared task, the COPA-* datasets in the standard South-Slavic languages were then extended to three micro-dialects that are spoken in narrow micro-geographical areas (Ljubešić et al., 2024a): **1)** the Cerkno dialect of Slovenian (COPA-SL-CER), spoken in the Slovenian Littoral region, specifically from the town of Idrija; **2)** the Chakavian dialect of Croatian from northern Adriatic (COPA-HR-CKM), specifically from the town of Žminj, and **3)** the Torlak dialect from southeastern Serbia (COPA-SR-TOR), specifically from the town of Lebane in Serbia.

The three dialectal datasets featuring in the DIALECT-COPA task were again created following the established translation and adaptation methodology of XCOPA. All data instances were translated and adapted from the closest standard language COPA (e.g., COPA-HR was used to derive COPA-HR-CKM), allowing the human translators to also consult the original English COPA as the additional source. Following the original COPA data split, COPA-SL-CER and COPA-SR-TOR contain 400 instances for training, 100 for development and 500 test instances. COPA-HR-CKM was treated as a surprise dialect, and it comprises only the 500 translated and adapted test instances. We allowed the use of any external data except the 500 test instances in any language for which a COPA dataset variant exists,[5] given the multi-parallel nature of the COPA datasets.

While the contamination of todays' LLMs with the English COPA dataset is very likely, we are rather sure that there is a minimum danger of the results of this shared task to be contaminated, and this is for the following reasons: (1) the dialectal datasets were not published before this shared task, (2) inspections of performance of various recent LLMs has shown not-perfect results on the English dataset, and (3) comparable results to the English ones were achieved on the non-English datasets, that are available for a short period of time. Finally, to ensure future validity of the measurements on this shared task's data, the test data of the DIALECT-COPA dataset are not published publicly, but are available only upon request of fellow researchers.

The evaluation metric regularly used in the COPA datasets, as well as inside this shared task, is accuracy, which puts the random baseline, given the binary nature of the task, at 50%. Ljubešić et al. (2024a) propose already competitive baselines, with Mixtral 8x7B Instruct (Jiang et al., 2024) zero-shotting achieving results around 70% accuracy on standard South Slavic datasets, but random to 63% accuracy on the dialectal datasets. Similarly, with zero-shotting the GPT-4 model (OpenAI et al., 2024), results of around 95% accuracy are reported for the standard South Slavic datasets, while the dialectal datasets achieve results between 60% and 93%. The significantly lower results on dialectal datasets, regardless of the model applied, show for the DIALECT-COPA dataset to be a very much open challenge and therefore a great fit for this evaluation campaign.

2.3 Participants

gmu-nlp. The team from the George Mason University submitted 10 runs, which is the maximum number of allowed runs in the shared task. Their approach (Faisal and Anastasopoulos, 2024) primarily focused on adaptation to dialects through various techniques of data augmentation: namely transforming *cause* instances into *effect* instances (and vice versa) by switching the place of the premise and the correct hypothesis, generating the non-available Chakavian training data by translating the standard data into the dialect via the the Claude 3 (Anthropic, 2024) and GPT-4 (OpenAI et al., 2024) models prompted with dialectal translation examples and rules, and fine-tuning a model on a combination of training data from specific languages and dialects. They inspected two models: the smaller Electra-style BERTić model (Ljubešić and Lauc, 2021), and the mT5-based aya-101 model (Üstün et al., 2024). The authors also used the 'trick' of independently fine-tuning a *cause* and an *effect* model.

JSI. The team from the Jožef Stefan Insitute submitted six runs, all based on zero- and few-shotting the Mixtral 8×7B Instruct model (Jiang et al., 2024)

[5] This of course refers to all the other 'COPA languages' beyond the South Slavic languages, e.g., all the XCOPA languages, Russian, and Catalan

team	run	name	API-only	adapt	sl-cer	hr-ckm	sr-tor	mean
gmu-nlp	1	orgl_hr_ckm_test	N	FT	0.700	0.750	0.824	0.758
gmu-nlp	2	aya	N	FS	0.694	0.756	0.84	0.763
gmu-nlp	3	orglc_omix_mk_hr_ckm_test	N	FT	0.690	0.756	0.836	0.761
gmu-nlp	4	orgl_sl_cer_test	N	FT	0.686	0.718	0.836	0.747
gmu-nlp	5	orgl_test	N	FT	0.682	0.760	0.824	0.755
gmu-nlp	6	orgl_mk_hr_ckm_test	N	FT	0.660	0.742	0.848	0.750
gmu-nlp	7	orgl_mk_hr_ckm	N	FT	0.582	0.634	0.682	0.633
gmu-nlp	8	all_train_rev_genx_omixmatch_select	N	FT	0.576	0.622	0.692	0.630
gmu-nlp	9	orgl_mk_hr_ckm_10	N	FT	0.572	0.626	0.722	0.640
gmu-nlp	10	orgl_10	N	FT	0.540	0.622	0.700	0.621
JSI	1	gpt4-zero	Y	ZS	0.594	0.754	0.908	0.752
JSI	2	gpt4-task	Y	FS	0.734	0.890	0.974	0.866
JSI	3	gpt4-list	Y	FS	0.696	0.846	0.946	0.829
JSI	4	mixtral-zero	N	ZS	0.518	0.576	0.706	0.600
JSI	5	mixtral-task	N	FS	0.542	0.640	0.724	0.635
JSI	6	mixtral-list	N	FS	0.578	0.618	0.722	0.639
WueNLP	1	MixtralLoRA-en-last	N	FT	0.562	0.626	0.714	0.634
WueNLP	2	MixtralLoRA-en-val	N	FT	0.574	0.620	0.706	0.633
WueNLP	3	MixtralLoRA-x-last	N	FT	0.556	0.606	0.738	0.633
WueNLP	4	MixtralLoRA-x-val	N	FT	0.550	0.608	0.738	0.632
UNIRI	1	RAG_simple_1	Y	ZS	0.688	0.760	-	-
UNIRI	2	simple_1	Y	ZS	0.664	0.774	0.894	0.777
UNIRI	3	RAG_with_reasoning_1	Y	ZS	0.708	0.764	-	-
UNIRI	4	with_reasoning_1	Y	ZS	0.608	0.664	0.806	0.693

Table 1: Official results on the DIALECT-COPA shared task. The evaluation metric is accuracy, with a random baseline of 0.5. The *API-only* column encodes whether the system is based on a closed model, available only through API calls or not. The *adapt* column categorizes the system adaptations whether they are based on fine-tuning (FT), few-shot (FS) or zero-shot (ZS) approaches.

and the GPT-4 model (OpenAI et al., 2024), the few-shotting approach exploiting their finding that correct answers are not crucial for the in-context learning of the dialect, and that the first N test instances, where correct answers are not given, can easily be exploited for that task, with great enhancements in results (Ljubešić et al., 2024b). The team also investigated a plethora of other models, the two selected models being by far the best performing in the group of open-source models (Mixtral 8x7B) and closed-source models (GPT-4).

WueNLP. The team from the University of Würzburg submitted four runs, all being focused on LoRA-fine-tuning the Mixtral 8×7B Instruct model (Jiang et al., 2024) either on English or on standard language data, following upon the logic that dialectal data might not be available for fine-tuning the model (Ljubešić et al., 2024b). The team regularly fine-tuned the model on the training subset only, keeping the development data for selecting the checkpoint with the best results.

UNIRI. The team from the University of Rijeka submitted four runs, all exploiting the GPT-4 model, the basic zero-shot approach being extended with a step-by-step-reasoning prompt and a retrieval-augmented-generation-based use of dialectal lexicons (Perak et al., 2024). The dialectal lexicons, available for two out of the three dialects in question, have previously been extended with examples generated by GPT-4.

2.4 Results

The official results of the four teams that have submitted their system descriptions are given in Table 1. The first observation to be made is that all of the runs on all of the systems have beaten the random baseline of 50% accuracy.

Starting with the *gmu-nlp* team, their results show an expected improvement in results when the aya-101 model is employed (runs 1-6) in comparison to the smaller BERTić model (runs 7-10). While the team provides very interesting approaches to data augmentation, the second run, based only on few-shotting the aya model, achieves very competitive results to the remaining runs employing the same model, but relying on LoRA-fine-tuning on various combinations and enhancements of the training data. Important to note is that the *gmu-nlp* team provided the best results overall when an open-source backbone LLM is used.

Moving on to the *JSI* team, they have reached the best results overall, but with the API-only closed-source GPT-4 model. They propose a simple zero-shot prompt, and two improvements of that prompt, both exploiting the first 10 instances from the test set. While the *list* prompt only gives exemplary sentences of the target dialect, the *task* prompt contains the structure and the goal of the task, but without an answer given. Both 10-shot prompts improve the zero-shot approach significantly, the *list* prompt being inferior to the *task* prompt, showing that, while learning about the dialect in-context is the biggest source of improvement, learning about the task itself does help further.

The *WueNLP* team, exploiting LoRA-based fine-tuning of Mixtral, obtained very similar results to those few-shot results of the *JSI* team. This shows that fine-tuning an LLM on 400 training instances on the specific task, either on English data (runs 1 and 2), or on the standard language data closest to the target dialect (runs 3 and 4), is equivalent to in-context learning from 10 instances in the target dialect (*JSI* team runs 5 and 6), even if the task itself (*JSI* team run 6), or an answer (*JSI* team run 5), are not provided. Interestingly, there is no difference in the results regardless of whether the English or the standard-variety training data are used for fine-tuning, showing that fine-tuning successfully informs the model of the task (the results are three points better than *JSI* team run 4 - Mixtral zero-shot results), but not of the final dialect.

Finally, the *UNIRI* team exploits, similarly to the *JSI* team, the GPT-4 model, but obtains better results on simple zero-shotting (*UNIRI* team run 2 vs. *JSI* team run 1), quite likely due to a better stated prompt, starting with *This is a reasoning task*. Where *UNIRI* do not improve is with the step-by-step-reasoning prompt, which lowers all their results (run 4). Interestingly enough, the step-by-step-reasoning prompt improves their results on standard languages (reported in their paper), showing that even GPT-4 is challenged by reasoning in a dialect to a level where the step-by-step-reasoning requirement hurts the performance. Interestingly, the retrieval-augmented-generation approach of *UNIRI* does help on the Slovenian Cerkno dialect, but slightly hurts the performance on the Chakavian dialect. A potential reason is that the overall performance on the Cerkno dialect is lower: therefore, the additional lexical information is more helpful than in the case with the Chakavian dialect.

2.5 Conclusions

The overall conclusions that can be drawn from the results of the DIALECT-COPA task are the following. First, there is a large dialectal gap present, given the difference between the results reported on the standard datasets and the dialectal datasets. Second, open-source models do not perform as well as the closed API-based models; however, few-shot or fine-tuned open models achieve the level of performance of zero-shot closed models. Third, data augmentation or retrieval-augmented-generation through dialectal lexicons seems to be as efficient as simply in-context learning from a few dialectal examples. Finally, the highly-efficient in-context learning seems to benefit mostly from the additional information on the dialect to be processed, rather than on the task itself.

3 The DSL-ML Task on Multi-Label Similar Language Identification

3.1 Motivation

VarDial has run shared tasks on the topic of discriminating between similar languages and varieties since its first edition. The DSL shared tasks organized from 2014 to 2017 focused on languages with several varieties like English, Spanish, Portuguese, and BCMS (Bosnian, Croatian, Montenegrin, Serbian) (Zampieri et al., 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014). These tasks were based on the DSL Corpus Collection (DSLCC Tan et al., 2014),[6] a collection of journalistic texts compiled assuming that each instance's variety label is determined by where the text is retrieved from. Previous research (e.g. Goutte et al., 2016) has shown the limitations of this problem formulation, as some texts (especially short texts such as single sentences) may not contain any linguistic marker that would allow systems, or even native speakers, to discriminate between two similar language varieties. In the past years, several proposals were made to address this issue:

- The DSL-TL dataset (Zampieri et al., 2023), introduced in conjunction with a shared task

[6] http://ttg.uni-saarland.de/resources/DSLCC/

	English	Portuguese	Spanish	French	BCMS
Number of varieties	2 (UK, US)	2 (PT, BR)	2 (AR, ES)	4 (BE, CA, CH, FR)	4 (BS, HR, ME, SR)
Annotation	Human	Human	Human	Automatic	Human
Train labeling	Multi-label	Multi-label	Multi-label	Multi-label	Single-label
Dev labeling	Multi-label	Multi-label	Multi-label	Multi-label	Multi-label
Test labeling	Multi-label	Multi-label	Multi-label	Single-label	Multi-label
Named entities	Present	Present	Present	Masked	Present
Avg. tokens/instance in *train*	33	38	52	64	5548
Training instances	2097	3467	3467	340,363	368
Multi-label instances in *dev*	13%	14%	32%	0.7%	20%

Table 2: Key properties of the datasets used in the DSL-ML task.

at VarDial 2023 (Aepli et al., 2023), contains Spanish, Portuguese and English sentences that were manually annotated using crowdsourcing. The annotation setup is restricted to two varieties per language (e.g. Peninsular and Argentinian Spanish), but allows a third option "Both or neither" if the instance does not provide sufficient grounds for reliable classification.

- Bernier-Colborne et al. (2023) argue that language variety identification is best framed as a multi-label classification problem. They analyze the FreCDo corpus (Găman et al., 2023) used in the VarDial 2022 FDI shared task (Aepli et al., 2022) and find substantial amounts of near-duplicate sentences associated with different labels in FreCDo. This near-duplicate analysis allows them to automatically derive a variant of FreCDo where ambiguous instances are annotated with multiple labels.

- Keleg and Magdy (2023) analyze different datasets used for Arabic dialect identification and find that many of the analyzed samples are valid in multiple dialects. As a result, the performance of dialect identification models is underestimated, as about two thirds of false positives are actually not true errors. Like Bernier-Colborne et al. (2023), they recommend multi-label annotations as a solution for future dialect identification tasks.

- Miletić and Miletić (2024) propose a reannotation of a single-annotator, single-label dataset for BCMS based on Twitter data (Rupnik et al., 2023). They explicitly introduce multi-label annotation based on labels produced by multiple annotators from all target regions. A re-evaluation of a previously proposed DSL system (Rupnik et al., 2023) against the multi-label annotation shows an improvement of the accuracy assessment (+4.1 points), indicating that some of the model predictions that were considered as wrong in the single-label setting are not necessarily errors. These results further support the multi-label annotation for the DSL task.

3.2 Data

The DSL-ML task is based on three data sources from five different languages. The choice of languages was mainly motivated by the availability of existing multi-label-annotated datasets. The five datasets have rather distinct properties in terms of size, instance lengths, genre, annotation and preprocessing. Table 2 summarizes these differences across the datasets (detailed statistics are provided in Table 3 in the appendix). For this reason, we provide distinct datasets for the five languages and evaluate the participants' submissions separately on each of them.

English, Portuguese, Spanish. For these languages, we re-use the DSL-TL dataset with the same split as in the VarDial 2023 task. We merely transform the "neither/both" labels to a comma-separated list of variant annotations. For example, the generic label `ES` becomes `ES-ES,ES-AR`.

French. The French training and development sets are obtained by combining the FreCDo (Găman et al., 2023) and DSLCC v4 (Tan et al., 2014) datasets, which comprise French (`FR-FR`), Swiss (`FR-CH`), Belgian (`FR-BE`), and Canadian (`FR-CA`) samples of text collected from the news domain. The topics used to collect most of the training and development data are available in the FreCDo paper. For the test data, we choose a new set of

		Training		Development		Test	
Language	**Label**	**# Samples**	**# Tokens**	**# Samples**	**# Tokens**	**# Samples**	**# Tokens**
English	EN-GB	755	21,011	211	5,767	114	3068
	EN-GB, EN-US	273	8,686	76	2,409	30	978
	EN-US	1,069	49,761	312	12,380	156	6352
	Total	2097	79,458	599	20,556	300	10,398
	Multi-label	13.0%		12.7%		10.0%	
Spanish	ES-AR	851	49,009	227	12,725	133	8,034
	ES-AR, ES-ES	1,131	61,559	318	17,421	156	8,528
	ES-ES	1,485	93,584	444	28,021	206	13,290
	Total	3,467	204,152	989	58,167	495	29,852
	Multi-label	32.6%		32.2%		31.5%	
Portuguese	PT-BR	2,136	98,061	588	26,848	299	13,605
	PT-BR, PT-PT	420	17,684	134	5,562	59	2,232
	PT-PT	911	38,524	269	11,379	137	5,887
	Total	3,467	154,269	991	43,789	495	21,724
	Multi-label	12.1%		13.5%		11.9%	
French	FR-BE	120,653	8,147,415	7,444	508,853	3,000	333,001
	FR-BE, FR-CA			2	108		
	FR-BE, FR-CH	603	44,991	31	1,920		
	FR-BE, FR-CH, FR-FR	61	2,681				
	FR-BE, FR-FR	1,052	81,602	82	5,295		
	FR-CA	19,041	557,468	2,167	148,669	3,000	334,755
	FR-CA, FR-FR			2	161		
	FR-CH	115,664	7,530,080	1,021	70,245	3,000	317,727
	FR-CH, FR-FR	162	12,218	3	186		
	FR-FR	83,127	5,280,740	6,338	432,269	3,000	323,485
	Total	339,537	21,657,195	17,090	1,167,706	12,000	1,308,959
	Multi-label	0.6%		0.7%		0.0%	
BCMS	BS	45	257,856	7	66,186	10	65,660
	BS, HR			4	29,596	3	9,661
	BS, HR, ME					1	1,634
	BS, HR, ME, SR			1	7,294		
	BS, ME			5	24,791	4	42,262
	BS, ME, SR					2	26,958
	BS, SR			4	23,398	1	2,015
	HR	53	385,385	16	128,760	16	131,821
	HR, SR			6	25,496	2	10,247
	ME	34	242,084	4	20,385	8	66,157
	ME, SR			5	45,738	3	17,340
	SR	236	1,489,997	70	434,136	73	479,606
	Total	368	2,375,322	122	805,780	123	853,361
	Multi-label	0.0%		13.0%		13.0%	

Table 3: Distribution of samples and tokens in the DSL-ML datasets.

topics, namely “inflation” (En.: “inflation”), “jeux olympiques” (En.: “olympic games”), and “reine d’angleterre” (En.: “queen of england”). Each topic was used to query two sources per country. We underline that the training and test topics and sources are disjoint, which generates a cross-domain evaluation setting. Multi-label annotations are inferred using the approach of Bernier-Colborne et al. (2023), which converts near duplicates into multi-label samples. After applying this data cleaning procedure, the training set remains with 340,363 samples, while the development and test sets consist of 17,090 and 12,000 samples, respectively. The training and development data are multi-label, meaning that samples may belong to more than one class, while the testing samples are single-label.[7] In contrast to the datasets of the other languages, named entities are

[7] Running the code of Bernier-Colborne et al. (2023) on the test data did not result in finding near duplicates.

replaced with the `$NE$` tag to prevent systems from learning named-entity-related shortcuts. The complete dataset contains approximately 370K samples and 33M tokens.

BCMS. The training set is the same as the BENCHIC-langTwitter training set (Rupnik et al., 2023) (except that retweets were removed from the data for the shared task) and thus only contains single-label annotations. The development and test sets come from the same collection, but were manually reannotated with multiple labels (Miletić and Miletić, 2024). The instances in this dataset cover the entire tweet production of a user and are thus much longer than the single-sentence instances of the other datasets.

Table 3 shows the number of samples and tokens per label and split for all DSL-ML languages, as well as the corresponding percentages of multi-label samples.

3.3 Baseline

The baseline proposed by the shared task organizers is based on an SVM classifier applied on a combination of TF-IDF-weighted character and word n-grams.[8] The classifier follows a multi-class (but not multi-label) setup where label combinations are added as distinct atomic labels. For example, the English task would have three distinct labels: the two single-variety labels `EN-GB` and `EN-US` as well as the multi-variety-label `EN-GB,EN-US`. This setup is equivalent to the one used in DSL-TL, except that the `EN` label is renamed to `EN-GB,EN-US`.

3.4 Participants

Brandeis. The Brandeis team (Sälevä and Palen-Michel, 2024) submitted 3 runs for each of the five languages. Their first run is based on a simple classifier applied to bag-of-n-gram features, where the n-grams are considered at both word and character levels. Aside from count n-gram-based statistics, they also employ the TF-IDF scheme as an alternative representation. For the classification, they alternatively consider logistic regression models, linear-kernel SVMs and random forest models.

For their second run, Sälevä and Palen-Michel (2024) employ a pre-trained multi-lingual BERT (mBERT) (Devlin et al., 2019) and independently fine-tune it on each subset of languages. To address the multi-label classification task, the authors attach a linear classification layer with a sigmoid activation for each unit, and use a threshold of 0.5 for the label to be included in the set of predicted labels. However, if there is no label surpassing the initial threshold, they gradually lower the threshold to 0.25 and 0.05, respectively.

The third run submitted by Brandeis is a variation of the second run, where the fine-tuning of mBERT is jointly performed on all languages (from all sub-tasks) at once.

Jelly. The Jelly team (Gillin, 2024) submitted 3 runs for English, Spanish and Portuguese and 1 run for French; they did not participate in the BCMS subtask. All submitted runs except one are based on one-shot prompting a large language model (LLM). The authors choose the open-source Mistral-7B model (Jiang et al., 2023). For each test sample, the authors provide a prompt containing one training example per language variety and expect the model to produce the multi-label prediction for the given test sample. The different runs differ in the postprocessing of the model output and the back-off strategy chosen if the model output did not contain any valid label.

For the English sub-task, run 2 refers to a variant of in-context learning where the prompt also contains instructions for the labeling task, and run 3 is an ensemble of runs 1 and 2. This team also submitted the raw outputs of Mistral-7B without postprocessing and backoff for comparison - these runs are marked as *open*.

VLP. The VLP team (Ngo et al., 2024) submitted one or two runs for each language. Their first run is based on a bidirectional long short-term memory network (BiLSTM) (Graves et al., 2013). It comprises an embedding layer, several BiLSTM layers and two dense layers, where the last one performs the classification of samples via softmax.

The second run employs the same architecture, but the input is based on ConceptNet embeddings (Speer et al., 2017). More specifically, the authors use ConceptNet Numberbatch semantic vectors, which provide a representation of word meanings extracted from ConceptNet. The ConceptNet embeddings are not available for BCMS, therefore only run 1 is submitted for that subtask. The VLP submissions consider all target labels as atomic, in the same way as the baseline.

[8] The code for the baseline system is available at `https://github.com/yvesscherrer/DSL-ML-2024/tree/main/baseline`. The system described here corresponds to the *atomic* option in the provided script.

English

Rank	Team	Run	Macro-F1	Multi-label EM
1	Brandeis	3	0.855	0.267
2	Brandeis	2	0.853	0.267
3	Brandeis	1	0.806	0.267
4	VLP	2	0.770	0.167
5	VLP	1	0.759	0.267
6	Jelly	2	0.755	0.133
7	Jelly	2-open	0.752	0.367
8	*Baseline*		0.751	0.100
9	Jelly	1	0.751	0.300
10	Jelly	3	0.750	0.367
11	Jelly	1-open	0.717	0.233

Spanish

Rank	Team	Run	Macro-F1	Multi-label EM
1	Brandeis	2	0.823	0.500
2	Brandeis	3	0.821	0.551
3	*Baseline*		0.770	0.391
4	VLP	1	0.754	0.455
5	Brandeis	1	0.746	0.455
6	VLP	2	0.741	0.423
7	Jelly	1	0.663	0.333
8	Jelly	2	0.655	0.289
9	Jelly	3	0.649	0.289
10	Jelly	1-open	0.601	0.199

Portuguese

Rank	Team	Run	Macro-F1	Multi-label EM
1	Brandeis	3	0.752	0.424
2	Brandeis	1	0.724	0.220
3	Brandeis	2	0.714	0.136
4	*Baseline*		0.683	0.068
5	VLP	1	0.664	0.136
6	Jelly	1	0.629	0.356
7	Jelly	2	0.593	0.136
8	Jelly	3	0.586	0.136
9	VLP	2	0.566	0.000
10	Jelly	1-open	0.388	0.034

French

Rank	Team	Run	Macro-F1
1	Brandeis	3	0.385
2	*Baseline*		0.372
3	Jelly	1	0.313
4	Brandeis	1	0.270
5	Brandeis	2	0.265
6	VLP	2	0.260
7	VLP	1	0.257

BCMS

Rank	Team	Run	Macro-F1	Weighted F1	Multi-label EM
1	Brandeis	1	0.762	0.843	0.000
2	Brandeis	2	0.719	0.756	0.125
3	*Baseline*		0.606	0.737	0.000
4	VLP	1	0.272	0.370	0.000
5	Brandeis	3	0.199	0.453	0.000

Table 4: Results of the DSL-ML shared task. The official metric is macro F1 score. We do not report weighted F1 score for English, Spanish, Portuguese and French since their test sets are (relatively) balanced and produce the same ranking. For BCMS, we report both macro-averaged and weighted F1-scores. *Multi-label exact match (EM)* refers to the proportion of correctly predicted instances with multiple labels. The French test set does not have multiple labels.

3.5 Results

We evaluate each subtask separately, using macro-averaged F1-score as the main metric. We additionally report weighted-average F1-score for the BCMS task since the class distribution in the test set is much less balanced than in the other tasks.

Furthermore, we measure the models' ability to perform multi-label classification by measuring *multi-label exact match*, i.e., the proportion of gold instances containing two or more labels for which the same set of labels was predicted. The results are presented per language in Table 4.

In general, we see that *Brandeis* is the only team that consistently beats the baseline on all subtasks. While their traditional machine learning submission (run 1) obtained first rank for BCMS, the BERT-based submissions (runs 2 and 3) are ranked highest on the other subtasks. *VLP* beats the baseline for English, is slightly below the baseline for Spanish and Portuguese, and considerably lower for French and BCMS. Their two runs perform roughly on par. Finally, *Jelly* narrowly outperforms the baseline for English, but remains several points below it for the other subtasks.

It can also be seen that the baseline does a comparatively poor job in correctly predicting the multi-labeled instances. While all three participating teams outperform the baseline in terms of multi-label exact match, team *Brandeis* again shows the most consistent performance.

Multi-Label Classification of the DSL-TL Data. Among all languages of this subtask, the overall results are the most encouraging for English. Seven out of ten submitted runs scored above the baseline based on the macro-F1 score (all three runs from *Brandeis*, runs 1 and 2 from *VLP*, and runs 2 and 2-open from *Jelly*), with the top-ranked system achieving a 10% improvement over the baseline. All systems also outperform the baseline on the multi-label exact match score. However, the multi-label exact match score remains relatively low, with the best score at 36.67%, achieved by runs 2-open and 3 submitted by *Jelly*, which are based on the Mistral-7B model. These runs ranked 7th and 10th, respectively.

For Spanish, only runs 2 and 3 by *Brandeis* score above the baseline, with the best VLP system scoring 3% below the baseline, and the *Jelly* runs lagging by 10 or more points. On this language, highly ranked systems also achieve solid results on the multi-label exact match score compared to other languages. In particular, run 3 from *Brandeis* reaches 55.13%.

For Portuguese, the three runs from *Brandeis* are the only systems that outperform the baseline on the macro-F1 score. Overall, the results on the multi-label exact match score are lower for this dataset than for other languages except BCMS. However, the top-ranked system does achieve 42.37%, and the second-best system on this metric is run 1 from *Jelly*, with 35.59%. This is another example of a system that lags behind the baseline based on the macro-F1 score (in this case, by 6 points), but which has a solid performance compared to other systems when it comes to labelling multi-label instances.

Multi-Label French Dialect Identification. For French, two models, one proposed by the Brandeis team and the other by the organizers, stand out from the rest. The top scoring model is based on jointly fine-tuning the mBERT model on all languages. Interestingly, this model is significantly better than the mBERT version fine-tuned on French data (run 2 of Brandeis team), indicating a large benefit from training on multiple languages.

The baseline is a shallow approach (linear SVM) based on basic features, which generalizes fairly well to the cross-domain setup of the French subtask. It is able to compete with the deep model based on multi-lingual fine-tuning submitted by the Brandeis team, being only 1.3% behind.

The third best model, submitted by the Jelly team, uses the Mistral-7B LLM based on in-context learning. Although in-context learning seems to work fairly well, the approach is clearly below the system based on multi-lingual fine-tuning proposed by Brandeis. The Jelly team (Gillin, 2024) obtained much better results on the English sub-task, likely because Mistral-7B is mostly trained on English text. Therefore, in the future, it would be interesting to explore approaches that combine fine-tuning and in-context learning.

The other models submitted by the participants are barely able to surpass the random chance baseline (with an F1 score of 0.25). The last three models are based on deep architectures, and their poor results are likely to be attributed to overfitting. In summary, we conclude that the French sub-task proposed for the 2024 edition of VarDial is very challenging, particularly because of the domain-shift between training and test data, as well as the generally short text samples which may not always contain dialectal patterns.

Multi-Label BCMS Variety Identification. Only Brandeis and VLP submitted runs on the BCMS data. Runs 1 and 3 by the Brandeis team score above the baseline, whereas the remaining submissions score significantly lower on both reported F1 scores. The top two systems achieve solid F1 results, on par with the ones they achieve on Portuguese, although lagging somewhat behind the top scores on English and Spanish.

As noted above, the Brandeis run 1, based on traditional machine learning approaches, achieves the best overall scores on BCMS. However, it is notable that only the Brandeis run 2, based on mBERT, scores above zero on the Multi-label Exact Match score. In other words, this is the only system that manages to correctly label any multi-label instances in the test set. Overall, the multi-label EM is the lowest on BCMS out of all of the languages of this subtask. These results indicate that, while the general task of distinguishing between the varieties of BCMS may be less difficult than it is for French, correctly labelling multi-label instances remains very challenging.

3.6 Conclusions

For the first time at VarDial, we proposed a language and dialect identification task that accepts multi-label scenarios with any number of classes. It includes three two-country settings (with three possible labels, for English, Spanish and Portuguese) as well as two four-country settings (with up to fifteen possible labels, for French and BCMS).

Among the five languages, French turned out to be the most challenging one in terms of obtained macro F1-scores. There are several possible explanations for this. The French data distinguishes itself from the other datasets by a domain shift between training and test data, by its reliance on automatic labeling (both for the initial single-label annotations and the inference of multi-label annotations), and by the masking of named entities. The relative impact of these properties is hard to quantify at the moment and will require additional experiments.

The BCMS task has also been found difficult, especially in terms of multi-label exact match. Eleven labels (country combinations) occur in the test set, but only four of them were observed in the training data, and nine of them in the development set. In such scenarios, it is crucial to use specific multi-label classifiers that can produce combinations of labels unseen at training time.

In terms of methods, both traditional classifiers and embedding-based models were proposed, but none of the two approaches clearly outperforms the other across languages. The *Jelly* submission introduces few-shot prompting as a potentially appealing training-free approach, but the results are not competitive yet with task-specific models. The used large language model often fails to provide the output labels in the correct format, and therefore heavy post-processing is required.

The five datasets used in the DSL-ML task differ widely in size and annotation procedures, and it can be seen that the different submissions are sensitive to different aspects of multi-label classification of similar varieties. We hope to have paved the way for further tasks that embrace the multi-label scenario.

4 Conclusion

This paper presented an overview of the two shared tasks organized as part of the VarDial Evaluation Campaign 2024: Dialectal causal commonsense reasoning (DIALECT-COPA) and Multi-label classification of similar languages (DSL-ML).

Among all the conclusions from the results on the DIALECT-COPA shared task presented in Section 2.5, the most interesting one is that in-context learning on dialectal examples seems to be a highly potent method of adapting an LLM to dialectal tasks. The intuition we have developed through this shared task is that it is all about managing expectations of LLMs, and that letting the LLM simply know about the modified language variant it will be tested on improves its performance significantly.

When it comes to the DSL-ML task, the observations stemming from this iteration further justify the multi-label approach to this task. This is supported both by the proportion of multi-label instances found in the data and by the multi-label exact match scores, which point to the difficulty of the task. We also noted that there were no clear winners in terms of methods between traditional classifiers and embedding-based models. However, as indicated above, the level of disparity between the five datasets used in this year's shared task makes it challenging to identify the impact of different factors on model performance. One possible way forward for this task would consist in creating a homogeneous dataset, taking advantage of best practices from the existing datasets.

Both tasks were shown to be rather challenging, opening up opportunities for future evaluation campaigns.

Acknowledgements

We thank all the participants for their interest in the shared tasks.

The work related to the DIALECT-COPA shared task has received partial funding from the Slovenian Research and Innovation Agency (J7-4642, P6-0411).

The work related to the DSL-ML shared task has received partial funding from the Academy of Finland (funding decision no. 341798).

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2024. Data-Augmentation based Dialectal Adaptation for LLMs. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. FreCDo: A large corpus for French cross-domain dialect identification. *Procedia Computer Science*, 225:366–373.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Nat Gillin. 2024. One-shot Prompt for Language Variety Identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Alex Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Nikola Ljubešić. 2021. Choice of plausible alternatives dataset in Croatian COPA-HR. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Boshko Koloski, Kristina Zdravkovska, and Taja Kuzman. 2022a. Choice of plausible alternatives dataset in Macedonian COPA-MK. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022b. Choice of plausible alternatives dataset in Serbian COPA-SR. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024a. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to South Slavic dialects. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Goran Glavaš, Fabian David Schmidt, and Ivan Vulić. 2024b. JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far . In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. European Language Resources Association.

The Quyen Ngo, Thi Anh Phuong Nguyen, My Linh Ha, Thi Minh Huyen Nguyen, and Phuong Le-Hong. 2024. Improving Multi-label Classification of Similar Languages by Semantics-Aware Word Embeddings. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating Dialect Understanding into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonne Sälevä and Chester Palen-Michel. 2024. Brandeis at VarDial 2024 DSL-ML Shared Task: Multilingual models, simple baselines and data augmentation. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh*

International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451. AAAI Press.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE Benchmark: Translation and Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels. *arXiv preprint arXiv:2303.01490*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue. *arXiv preprint arXiv:2212.01853*.