A Appendix

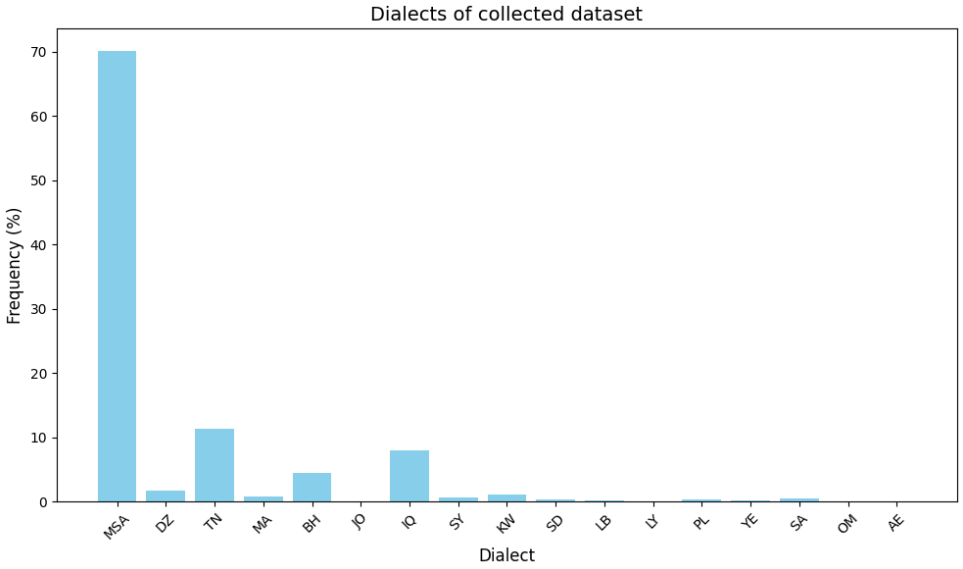

Figure 4: Sentences classified as dialects in the collected dataset.

Gemini-flash filter prompt

Alege id-urile propozițiilor care au o formulare ciudată sau care sunt negramaticale. Returnează rezultatul sub forma de text conținând doar id-urile separate prin ',' și nimic altceva. {Urmat de lista de propoziții din dataset}

Jais Translation Prompt

Instruction: You can answer in English only. Translate the Arabic sentence into English.

Input: [—Human—] {Question}

GPT-4 and Gemini translation prompt

Role: Developer: You are a helpful assistant.

Role: User: Translate the following arabic text: {Arabic text} into Romanian. Write only the translation and nothing else.

RoLLama3.1-8b and RoMistral-7b-Instruct training and inference prompt

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
Instruction: Given the following arabic sentence, translate it into romanian
Input: {Arabic sentence}
Response: {Romanian sentence}

Dataset	Model	BLEU	chrF++	wmt22-comet-da	BERTScore
	Gemini-flash-002	18.89	54.98	0.877	0.832
	GPT-4	19.79	55.34	0.871	0.832
	Fine-tuned NLLB-600M	14.37	47.43	0.824	0.822
	Pre-trained NLLB-600M	13.16	47.32	0.821	0.821
Out-of-Domain	Fine-tuned RoMistral-7b	12.32	43.23	0.817	0.799
	Pre-trained RoMistral-7b	7.82	41.63	0.793	0.788
	Pre-trained RoLLama3.1-8b	5.35	36.47	0.788	0.764
	Transf. (Dialects removed)	16.9	48.03	0.786	0.807
	Transf. (Duplicates removed)	17.86	49.03	0.781	0.803
	Fine-tuned RoLLama3.1-8b	9.08	41.37	0.780	0.762
	Transf. (Gemini filter)	16.41	47.25	0.770	0.804
	Jais+Transf.	12.83	43.8	0.764	0.767
	Transf. (Lealla filter)	17.10	48.04	0.769	0.799
Dataset	**Model**	**BLEU**	**chrF++**	**wmt22-comet-da**	**BERTScore**
	Gemini-flash-002	30.54	57.81	0.883	0.866
	GPT-4	30.58	58.25	0.881	0.865
	Fine-tuned RoLLama3.1-8b	23.25	50.71	0.846	0.839
	Fine-tuned RoMistral-7b	24.56	51.45	0.840	0.843
FLORES+	Transf. (Duplicates removed)	25.12	52.38	0.834	0.851
	Transf. (Lealla filter)	24.88	52.02	0.831	0.850
	Pre-trained RoLLama3.1-8b	17.02	49.10	0.829	0.818
	Jais+Transf.	25.55	52.61	0.826	0.846
	Transf. (Gemini filter)	24.19	51.21	0.825	0.848
	Fine-tuned NLLB-600M	23.56	51.12	0.825	0.846
	Transf. (Dialects removed)	24.15	51.52	0.822	0.846
	Pre-trained NLLB-600M	21.81	49.34	0.819	0.843
	Pre-trained RoMistral-7b	17.47	46.16	0.783	0.799
Dataset	**Model**	**BLEU**	**chrF++**	**wmt22-comet-da**	**BERTScore**
	GPT-4	23.62	52.48	0.842	0.825
	Gemini-flash-002	23.62	51.42	0.838	0.824
	Transf. (Duplicates removed)	29.71	54.18	0.834	0.843
	Transf. (Lealla filter)	29.32	53.81	0.833	0.842
In-domain	Transf. (Gemini filter)	29.02	53.47	0.830	0.842
	Transf. (Dialects removed)	29.21	53.58	0.830	0.840
	Fine-tuned NLLB-600M	25.65	50.88	0.820	0.832
	Fine-tuned RoMistral-7b	22.02	49.29	0.814	0.820
	Pre-trained NLLB-600M	23.34	49.08	0.811	0.829
	Fine-tuned RoLLama3.1-8b	8.11	41.94	0.778	0.788
	Pre-trained RoLLama3.1-8b	8.63	40.47	0.778	0.773
	Jais+Transf.	19.47	42.12	0.776	0.782
	Pre-trained RoMistral-7b	12.04	40.49	0.753	0.767

Table 4: Aggregated results for all models. Saudi-Romania Econ., Geneva Convention, Children Rights, Economic Rights, Human Rights, Saudi Embassy, Refugee Status subsets represent Out-of-Domain data (OOD), results are averaged. Values have been rounded - BLEU to two decimals, `Unbabel/wmt22-comet-da` and BERTScore to three decimals. The models are sorted by wmt22-comet-da score.

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	14.6	45.92	0.7517	0.7978
Geneva Convention	10.2	43.00	0.7546	0.7846
Children Rights	17.2	50.28	0.8400	0.8360
Economic Rights	16.5	51.36	0.8137	0.7858
Human Rights	17.0	48.70	0.8110	0.7781
Saudi Embassy	12.3	43.24	0.6679	0.8107
Refugee Status	11.3	44.74	0.7907	0.7858
FLORES+	24.1	51.52	0.8224	0.8467
In-Domain dataset	29.2	53.50	0.8300	0.8403

Table 5: Transformer trained on the dataset with pruned sentence pairs identified as part of dialects.

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	14.13	44.52	0.7476	0.7961
Geneva Convention	10.62	42.98	0.7336	0.7868
Children Rights	16.81	49.49	0.7547	0.7831
Economic Rights	15.73	50.45	0.7924	0.7941
Human Rights	14.96	46.97	0.7704	0.7833
Saudi Embassy	13.10	42.23	0.6918	0.8242
Refugee Status	9.12	43.94	0.7901	0.7826
FLORES+	24.19	51.21	0.8252	0.8481
In-Domain dataset	29.00	53.47	0.8302	0.8423

Table 6: Transformer trained on the dataset filtered using Gemini.

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	14.4	45.71	0.7476	0.7909
Geneva Convention	10.6	43.42	0.7257	0.7785
Children Rights	15.6	48.23	0.7523	0.7747
Economic Rights	18.1	51.96	0.7976	0.7939
Human Rights	16.8	49.00	0.7797	0.7859
Saudi Embassy	14.2	43.63	0.6863	0.8188
Refugee Status	10.0	44.60	0.7762	0.7592
FLORES+	24.8	52.02	0.8311	0.8503
In-Domain dataset	29.3	53.81	0.8330	0.8420

Table 7: Transformer trained on the dataset filtered using Lealla.

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	15.34	47.25	0.7535	0.7932
Geneva Convention	10.11	43.41	0.7754	0.7901
Children Rights	17.64	50.53	0.7544	0.7737
Economic Rights	19.75	53.87	0.8006	0.7990
Human Rights	17.61	49.70	0.7954	0.7950
Saudi Embassy	14.96	44.60	0.6925	0.8216
Refugee Status	10.53	45.39	0.7881	0.7687
FLORES+	25.12	52.38	0.8344	0.8510
In-Domain dataset	29.71	54.18	0.8347	0.8431

Table 8: Transformer trained dataset with duplicates removed.

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	0.2	6.169	0.2825	0.6084
Geneva Convention	0.4	10.66	0.4063	0.6467
Children Rights	0.3	9.45	0.3835	0.6212
Economic Rights	0.0	9.24	0.4071	0.6404
Human Rights	0.0	10.02	0.4246	0.6578
Saudi Embassy	0.0	5.24	0.2280	0.5842
Refugee Status	0.0	8.65	0.3400	0.6081
FLORES+	1.5	13.38	0.3584	0.6542
In-Domain dataset	4.5	17.93	0.5063	0.7148

Table 9: Baseline Transformer trained on unclean data.

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	13.1	47.26	0.7650	0.7886
Geneva Convention	10.3	44.16	0.8227	0.8450
Children Rights	14.5	49.59	0.8441	0.7894
Economic Rights	14.0	51.64	0.8795	0.8550
Human Rights	15.3	48.68	0.9120	0.8967
Saudi Embassy	16.4	47.08	0.7200	0.8176
Refugee Status	8.5	42.72	0.8007	0.7521
FLORES+	21.8	49.34	0.8189	0.8429
In-Domain dataset	23.3	49.08	0.8110	0.8291

Table 10: Pre-trained NLLB-600M

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	14.9	48.23	0.7759	0.7980
Geneva Convention	10.5	43.58	0.8195	0.8437
Children Rights	19.4	51.87	0.8567	0.8024
Economic Rights	15.4	51.27	0.8872	0.8556
Human Rights	16.7	48.85	0.9050	0.8961
Saudi Embassy	15.3	45.54	0.7291	0.8109
Refugee Status	8.4	42.61	0.8014	0.7553
FLORES+	23.5	51.12	0.8252	0.8464
In-Domain dataset	25.6	50.88	0.8201	0.8325

Table 11: Fine-tuned NLLB-600M

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	15.4	46.77	0.7634	0.7775
Geneva Convention	8.3	40.53	0.7700	0.7724
Children Rights	12.7	44.87	0.7422	0.7557
Economic Rights	10.2	44.29	0.7537	0.7575
Human Rights	20.4	47.50	0.7765	0.7932
Saudi Embassy	15.8	43.26	0.7579	0.7803
Refugee Status	7.0	39.38	0.7847	0.7343
FLORES+	25.5	52.61	0.8255	0.8464
In-Domain dataset	19.4	42.12	0.7762	0.7821

Table 12: Jais+Transformer results

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	21.1	56.05	0.8212	0.8067
Geneva Convention	14.4	51.09	0.8483	0.8135
Children Rights	18.3	53.52	0.8753	0.7948
Economic Rights	15.7	54.10	0.9090	0.8637
Human Rights	24.5	58.18	0.9412	0.9119
Saudi Embassy	33.6	67.54	0.8652	0.8719
Refugee Status	10.9	46.87	0.8399	0.7627
FLORES+	30.5	58.25	0.8807	0.8648
In-Domain dataset	23.6	52.48	0.8428	0.8252

Table 13: GPT4 results

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	17.7	53.73	0.8186	0.7954
Geneva Convention	13.6	50.48	0.8703	0.8317
Children Rights	15.4	52.69	0.8756	0.7882
Economic Rights	15.6	54.89	0.9088	0.8620
Human Rights	20.5	55.66	0.9380	0.9027
Saudi Embassy	38.8	70.61	0.8844	0.8822
Refugee Status	10.6	46.82	0.8413	0.7611
FLORES+	30.5	57.81	0.8832	0.8655
In-Domain dataset	23.62	51.42	0.8385	0.8243

Table 14: Gemini-flash-002 results

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	6.5	40.23	0.7327	0.7375
Geneva Convention	1.5	25.92	0.7045	0.7048
Children Rights	4.4	35.61	0.7941	0.7446
Economic Rights	3.4	39.22	0.8311	0.7916
Human Rights	2.9	30.04	0.8742	0.8275
Saudi Embassy	15.9	50.76	0.7914	0.8213
Refugee Status	2.9	33.51	0.7915	0.7274
FLORES+	17.0	49.10	0.8299	0.8181
In-Domain dataset	8.6	40.47	0.7785	0.7735

Table 15: Pre-trained RoLlama3.1-8b-Instruct results

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	8.5	41.39	0.7695	0.7714
Geneva Convention	6.4	37.36	0.7418	0.7246
Children Rights	11.3	44.94	0.7879	0.7641
Economic Rights	9.1	44.25	0.7988	0.7640
Human Rights	5.9	37.11	0.7604	0.7454
Saudi Embassy	14.2	42.07	0.7910	0.8113
Refugee Status	8.2	42.53	0.8121	0.7551
FLORES+	23.2	50.71	0.8469	0.8399
In-Domain dataset	8.1	41.94	0.7784	0.7881

Table 16: Fine-tuned RoLlama3.1-8b-Instruct results

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	4.7	36.34	0.7029	0.7393
Geneva Convention	6.6	40.39	0.8025	0.8175
Children Rights	9.2	42.97	0.8173	0.7562
Economic Rights	6.0	43.79	0.8462	0.8223
Human Rights	8.3	43.81	0.8713	0.8566
Saudi Embassy	16.5	50.08	0.7654	0.8080
Refugee Status	3.5	34.03	0.7500	0.7204
FLORES+	17.4	46.16	0.7833	0.7992
In-Domain dataset	12.0	40.49	0.7533	0.7676

Table 17: Pre-trained RoMistral-7b-Instruct

Dataset	BLEU	chrF++	wmt22-comet-da	BERTScore
Saudi-Romania Econ	12.8	43.06	0.7815	0.7874
Geneva Convention	7.7	37.37	0.6947	0.7125
Children Rights	14.7	46.71	0.8547	0.7898
Economic Rights	12.0	46.61	0.8844	0.8525
Human Rights	18.3	48.95	0.9174	0.8971
Saudi Embassy	12.4	38.90	0.7845	0.8112
Refugee Status	8.4	41.04	0.8091	0.7452
FLORES+	24.5	51.45	0.8409	0.8436
In-Domain dataset	22.0	49.29	0.8144	0.8205

Table 18: Fine-tuned RoMistral-7b-Instruct

Model	Severity	MT Output (Ro)	Reference (Ro)
Gemini	critical	Alteța Sa Regală Prințul Mohammed bin Salman bin Abdulaziz Al Saud, Prinț Moștenitor și Prim-Ministru, a trimis o scrisoare oficială către Prim-Ministrul României, domnul Nicolae Ciucă. **Scrisoarea a fost remisă de către ambasadorul Custodelui celor Două Sfinte Moschei în România, domnul dr. Mohammed bin Abdulghani Khayat, consilierului Prim-Ministrului României, domnul Iulian Chifu, în numele Prim-Ministrului român.**	Alteța Sa Regală Prințul Mohammed bin Salman bin Abdulaziz Al Saud, Prinț Moștenitor și Prim-ministru al Regatului Arabiei Saudite, a trimis un mesaj scris Prim-ministrului #României, domnului Nicolae Ciucă. Excelența Sa Dl. Iulian Chivu, Consilierul Prim-ministrului #României, a primit în numele premierului, scrisoarea de la Excelența Sa Dl. Dr. Mohammed bin Abdulghani Khayat, Ambasadorul Regatului Arabiei Saudite la #București.
LLAMA_PRE	major	Se pare că districtul Al-Qaim, situat în partea de vest a **guvernatoratului Anbar**, care se întinde până la granița cu Bagdad, a fost o zonă de contrabandă timp de mulți ani**, unde triburile trăiau** de ambele părți ale frontierei.	Al-Qaim, situat în vestul extrem al provinciei deşertice irakiene Anbar, care se întinde până la marginea Bagdadului, este de mult timp o zonă de contrabandă, în care trăiesc triburi de o parte şi de alta a frontierei.
GPT4	minor	**Întoarceți-vă la mijlocul secolului al XIX-lea.** Cine vrea să înceapă licitația?	Cine vrea să înceapă licitaţia? Două lire careva?

Table 19: Representative translations per severity level, comparing MT output to reference. Highlighted with bold are annotated MT mistakes.