

hüküm1 and hüküm1nlp: Turkish Decision Models from Turkish-Gemma-4b-T1-Scout with LoRA and a Clef Decision Head

Abstract

We describe how two Turkish *decision models*, hüküm1 and hüküm1nlp, were built from YTÜ COSMOS’s Turkish-Gemma-4b-T1-Scout [1], a Turkish adaptation of Gemma 3 4B [2]. A decision model does not generate text: it reads a state and a set of typed questions and returns a calibrated probability distribution over the options of every question in a single forward pass. We attach Unsloth’s [5] implementation of the Clef joint-schema decision head [6] to the backbone, adapt the backbone with LoRA [4], and train both with a soft-target cross-entropy objective, followed by per-question-type temperature calibration. hüküm1 is trained for one epoch on a general Turkish/English decision mix of 101,916 records (205,039 decisions). hüküm1nlp continues from hüküm1 on 17 Turkish NLP datasets recast as decisions, with 25% replay. hüküm1 scores 67.7 on the 260 human-gold questions of HakemBench and 69.3 on average over 29 closed-form Turkish benchmarks; hüküm1nlp improves the Turkish NLP sets (TREMO 82.2→94.5, XNLI-tr 74.4→82.6) and loses 1–5 points on the general decision sets. Both models do well on inference and reading tasks and poorly on knowledge tests such as TR-MMLU (54.6), where the 4B backbone is the limit.

1 Introduction

Large language models are often used as classifiers by prompting them and parsing the answer they generate. A decision model (a “System One” model in the terminology of Jev [7]) skips generation: the input and the candidate options are encoded once, and a small head returns a probability for each option. The output is always one of the defined options, and the probabilities can be thresholded so that uncertain cases go to a person.

Most open decision models are trained on English data. We found four open Turkish ones and compare against them in §7.4. This report describes two models built on Turkish-Gemma-4b-T1-Scout with LoRA and a Clef head. hüküm1 is trained on a general decision mix (§5), and hüküm1nlp adds a second stage on Turkish NLP tasks (§6). The weights are at <https://huggingface.co/vinnipukh/hukum1> and <https://huggingface.co/vinnipukh/hukum1-nlp>.

2 Decision format

A request consists of a *state* s (free text or JSON) and a set of named questions Q (up to six per request in training). Each question $q \in Q$ has one of three types:

- **choice**: one of K given options, $p \in \Delta^{K-1}$;
- **noul**: a yes/no question, $p(\text{yes}) \in [0, 1]$;
- **score**: a level on an ordered scale of K levels, $p \in \Delta^{K-1}$.

Options carry a key and an optional natural-language description. Gold targets are always distributions: hard labels become one-hot vectors; where several annotators or a teacher model provide votes or probabilities, the target is the soft distribution. The training loss is the cross-entropy between the predicted distribution and the soft target, summed over the questions of a record.

3 Model

3.1 Backbone

Turkish-Gemma-4b-T1-Scout [1] is a Turkish fine-tune of Gemma 3 4B [2]. Its text tower has 34 decoder layers, hidden size 2,560, MLP size 10,240, 8 query and 4 key/value heads of dimension 256, interleaved sliding-window (1,024) and global attention, and a 262k vocabulary; the full model, including the vision tower, has 4.3B parameters. Only text inputs were used; the vision tower is loaded but not exercised by the decision forward pass.

3.2 LoRA

LoRA adapters [4] with rank $r = 64$ and $\alpha = 64$ (no dropout) are attached to all attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and all MLP projections (`gate_proj`, `up_proj`, `down_proj`) of every text layer. This adds 3.51M parameters per layer, 119.2M in total.

3.3 Clef decision head

The head follows the Clef joint-schema design [6] as implemented in Unsloth 2026.10.3. The state, every question and every option (key and description) are laid out in one token sequence; the backbone processes it once, and the head reads the final-layer hidden states at the positions of the question and option spans. All questions of a request share this single prefill.

The released head (`joint_head.safetensors`) has 32.0M parameters, configured with width 512, 2 routing layers, 4 joint layers, 8 attention heads and feed-forward size 2,048. The backbone states pass through a layer norm and six linear projections from 2,560 to 512 dimensions. These give the question vectors, two option vectors (lexical and contextual), a question-conditioned option vector, a global summary and a “memory” of the whole prompt. A learned embedding marks the question type. In the two routing (evidence) layers, question and option fields attend to the prompt memory. The four joint transformer layers then run over all fields of the request, so an option can be compared with the other options of its question and with the other questions. The final scorer adds a joint logit, a lexical-prior logit and a gated residual MLP score, each with a learned scale, into one logit per option. For `noul` questions the two outcomes are scored as options; for `score` questions the levels are. A softmax over the options of each question gives the predicted distribution.

The LoRA adapters and the head together give 151.2M trainable parameters, 3.5% of the 4.3B total.

3.4 Calibration

After training, one temperature per question type (`choice`, `noul`, `score`) is fitted on a held-out calibration set and stored with the model; at inference the logits of each question are divided by the temperature of its type before the softmax (temperature scaling). Table 1 lists the fitted values.

Table 1: Calibration temperatures and held-out calibration-set results.

Model	T_{choice}	T_{noul}	T_{score}	Accuracy	ECE
hüküm1	1.074	0.732	1.016	81.0	3.5
hüküm1nlp	1.002	0.500	1.110	88.8	3.4

4 Backbone selection

Before full training, four backbones were trained under identical conditions: 4,800 records sampled from the stage-1 mix, 300 steps, effective batch 16, LoRA $r = 64$, the same Clef head and learning rates. Each was evaluated on the same 300 unseen examples of four evaluation sets (Table 2; about ± 5 points of sampling uncertainty). Kumru-2B memorised small subsets but did not generalise and was dropped. Turkish-Gemma-4b had the best results on the Turkish sets and the lowest Turkish evaluation loss, and was chosen.

Table 2: Backbone control experiment: accuracy before $\rightarrow$ after 300 steps on 300 unseen examples per set.

Backbone	cisimcik (tr)	typed (en)	jev-bench-tr	turkish_noul
Kumru-2B	33.1 $\rightarrow$ 44.2	29.6 $\rightarrow$ 35.4	25.3 $\rightarrow$ 31.7	48.7 $\rightarrow$ 52.3
Qwen3.5-2B	39.0 $\rightarrow$ 76.3	32.7 $\rightarrow$ 66.7	34.7 $\rightarrow$ 56.7	52.0 $\rightarrow$ 75.7
Nanbeige4.2-3B	38.1 $\rightarrow$ 74.2	37.1 $\rightarrow$ 68.7	31.0 $\rightarrow$ 55.3	51.7 $\rightarrow$ 74.7
Turkish-Gemma-4b	32.0 $\rightarrow$ 77.6	38.1 $\rightarrow$ 65.0	24.0 $\rightarrow$ 55.3	48.7 $\rightarrow$ 80.7

5 Stage 1: hüküm1

5.1 Data

The general decision mix combines open decision datasets in a single canonical format (Table 3). Labels are stored positionally, one probability per option in order. Long question sets are split into records of at most six questions over the same state, and states are capped at 10,000 characters. Large sources are subsampled with stratification over their sub-tasks.

Table 3: Stage-1 decision mix: 101,916 records, 205,039 decisions, ~ 47.7 M tokens.

Source	Records	Language
tasksource typed decisions [8]	30,193	en (+193 tr)
BEV decision mix [9]	20,000	en
tasksource procedural [8]	10,000	en
jev-bench train [10]	10,000	en
turkish_jev_noul [11]	8,100	tr
tasksource synthetic [8]	7,045	en
hmm synthetic [12]	6,684	en
Turkish Decisions Max [13] ($\times 2$)	6,490	tr
SemEval-2016 Turkish restaurant ABSA, relabelled	2,204	tr
typed-decisions [14]	1,200	en
Total	101,916	$\sim 17\%$ tr

Of the 205,039 decisions, 120 did not fit the 6,144-token limit and were dropped and 35 were truncated. 193 records were held out for calibration, leaving 101,701 training records.

5.2 Training

The loss is soft-target cross-entropy per question. We use AdamW with weight decay 0.01, a learning rate of 10^{-4} for LoRA and 3×10^{-4} for the head, a cosine schedule and 3% warm-up; the head is trained together with LoRA from the first step. Each step takes 8 records with 2 gradient-accumulation steps, sequences of up to 6,144 tokens, in bf16 without `torch.compile`.

One epoch is 6,357 steps and took about 6.7 hours on one A100 40GB (3.7–3.9 s per step), with Unsloth 2026.10.3, Transformers 5.17.0, PEFT 0.21.1, TRL 1.13.0 and PyTorch 2.11.

The run was split into three resumed segments; adapter, head, optimiser, scheduler and RNG state were restored at each boundary, and the learning-rate schedule spans the whole run, so the segmentation does not change the result. The 500-step mean training loss fell from 0.911 (steps 0–500) to 0.685 (step 1,800), 0.588 (step 4,080) and 0.535 at the end.

5.3 Readout ablation

To check whether the Clef head limits the backbone, three readouts were compared on the same 300-step budget: (a) the Clef head, (b) reading the option-letter rows of the backbone’s own LM head, and (c) (b) plus a small residual head initialised at zero (Table 4). The letter readout was at least as good as Clef at this budget and better calibrated on some sets. The full runs kept the Clef head. We did not test whether the letter readout keeps its lead over a full epoch.

Table 4: Readout ablation, 300 steps, one seed. Accuracy / ECE (%).

Set (questions)	(a) Clef	(b) LM-head letters	(c) (b) + residual head
cisimcik holdout (1,566)	78.9 / 10.0	79.2 / 4.4	79.6 / 5.0
typed-decisions (1,500)	66.1 / 9.5	70.9 / 10.0	69.1 / 8.8
jev-bench-tr (243)	59.7 / 10.6	60.9 / 9.1	57.6 / 14.2
turkish_noul (300)	79.7 / 4.0	82.3 / 3.8	82.3 / 3.9
Tachyon100 (100)	78.0 / 11.1	82.0 / 8.2	84.0 / 4.9

6 Stage 2: hüküm1nlp

6.1 Data

Eighteen Turkish NLP datasets were converted into the decision format; 17 contribute training data and XQuAD-tr [22] is used only for evaluation (Table 5). Conversion rules:

- Training questions use two phrasing templates and short option keys; test questions use three templates (one never seen in training) and options with descriptions. Every test set is also evaluated without option descriptions.
- Training sets are lightly balanced; test sets keep their natural label distribution.
- Where annotator vote shares exist (e.g. TREMO), they are used as soft targets.
- Reading comprehension uses distractor answers drawn from the same paragraph.

The stage-2 mix has 66,872 records, 114,679 questions and about 17M tokens. Of these, 50,154 records come from the NLP sets, capped at 8k questions per dataset (3k for BilCat, wikiNER, word sense and semantic roles). To 15% of the questions we added a sibling question (10,602 in total): a **choice** question gets “Is this answer correct: X ?” or “Is this answer wrong: X ?”, and a **noul** question gets its negated form. The remaining 16,718 records are replayed from the stage-1 mix, 25% of the records and about 44% of the tokens. A text is removed when 50% or more of its 8-grams occur on the other side of a split; train rows that copy evaluation or test text and test rows that copy training text were removed, comparing against all NLP splits, the Turkish stage-1 rows, the stage-1 evaluation sets and the external benchmarks.

Table 5: Turkish NLP datasets used in stage 2.

Task	Datasets
Aspect sentiment	YTÜ Cosmos absa-tr [15] (LLM-consensus labels)
Emotion / irony	TREMO; IronyTR [16]
Fact checking	FACTurk (verdict words removed from the evidence)
Inference / commonsense	XNLI-tr [17] (human-translated dev split); TrGLUE MNLI and RTE [18]; WinoGrande-tr
Classification	MASSIVE tr-TR intents [19]; BilCat news topics; complaint topics; Turkish SMS spam [20]
Reading comprehension	TQuAD-2 [21] (XQuAD-tr evaluation only)
Structure	turkish-wikiNER entity types; Starlang KeNet word sense, PropBank semantic roles and FrameNet frames; CorefUD Turkish-ITCC coreference [23]

6.2 Training

hüküm1nlp starts from hüküm1’s LoRA adapter and head at step 6,357 with a fresh optimiser and the same hyper-parameters (LoRA 10^{-4} , head 3×10^{-4} , cosine, 3% warm-up, batch 8×2 , length 6,144). One epoch is 4,164 steps: 2.73 hours on one A100 40GB (~ 2.4 s per step), mean training loss 0.352. The temperatures were refitted afterwards (Table 1).

7 Evaluation

Protocol. Accuracy is per question, at the calibrated temperatures; ECE uses equal-width bins; macro-F1 is computed over option positions. All test sets were screened against the training data with the 8-gram rule above. Sets of 100–400 questions carry ± 3 –5 points of sampling noise; all numbers are from a single training run.

Benchmarks. The decision sets are the held-out splits of the stage-1 sources. HakemBench v1.0 has 4,275 Turkish questions in 7 tracks, and we report its 260 human-gold questions separately. For MMLU-Pro-TR we use a fixed 1,400-question subset, after removing 141 questions that share a 10-gram with the training data. Tachyon100 is a 100-question Turkish sentiment set.

hukum-bench (not released) recasts 29 closed-form Turkish benchmarks as decisions. The question becomes the state and the answer choices become letter keys, with the choice text as their description; “is this correct?” tasks become nou1 questions, and NLI uses a three-way template. Choices keep their order from the source. Large sets are capped at 2,000 questions, while XNLI, TR-MMLU and TrClaim are used in full, and questions that share 50% or more of their 8-grams with training text were dropped. Because a decision model sees every option, hukum-bench numbers are not directly comparable with log-likelihood results from lm-evaluation-harness, in which the model does not see the alternatives.

7.1 Decision sets and external benchmarks

7.2 Turkish NLP sets: hüküm1 $\rightarrow$ hüküm1nlp

Before stage 2, option descriptions changed accuracy in both directions (FACTurk +17 and absa-tr +7 without them; XNLI -4 and wikiNER -5); after stage 2 the two conditions are within 2.5 points on every set. A character 2–5-gram TF-IDF + logistic-regression baseline trained on each set’s own training split scores 46.8 on XNLI-tr, 88.8 on TREMO, 89.8 on BilCat, 86.8 on MASSIVE and 47.7 macro-F1 on absa-tr; stage 2 closes the gap on the topic-style sets where hüküm1 trailed this baseline.

Table 6: Decision sets and external sets. Accuracy (ECE) in %. The train splits of the first four sets are in the stage-1 data; their test rows are not.

Set	Questions	hüküm1	hüküm1nlp
cisimcik holdout (tr)	1,946	84.5 (5.0)	82.6 (3.9)
turkish_noul test (tr)	1,900	88.2 (2.5)	87.3 (4.5)
jev-bench-tr test (tr)	3,000	69.1 (2.8)	68.2 (6.3)
typed-decisions test (en)	2,000	77.4 (11.9)	75.4 (9.5)
HakemBench human-gold	260	67.7 (9.3)	63.1 (8.7)
HakemBench, mean of 7 tracks	4,275	78.0	79.1
MMLU-Pro-TR subset	1,400	27.6 (11.3)	24.5 (16.7)
Tachyon100	100	85.0 (13.3)	92.0 (8.0)

Table 7: Turkish NLP test sets before (hüküm1) and after (hüküm1nlp) stage 2. Accuracy in %; macro-F1 in parentheses where labels are skewed. “No desc.”: accuracy with option descriptions removed.

Set	Task	Questions	With descriptions		No desc.	
			hüküm1	hüküm1nlp	hüküm1	hüküm1nlp
TREMO	emotion	400	82.2	94.5	85.5	95.0
absa-tr	aspect sentiment	672	80.7 (70.8)	90.2 (77.4)	87.8	89.6
IronyTR	irony	120	70.8	75.0	71.7	75.0
FACTurk	fact checking	400	42.8 (28.0)	58.8 (50.5)	59.8	61.3
XNLI-tr (400 premises)	NLI	1,200	77.2	82.4	73.2	82.3
TrGLUE MNLI + RTE	NLI	400	86.5	87.8	82.8	87.0
WinoGrande-tr	Winograd	400	59.2	63.5	59.2	63.5
MASSIVE tr-TR	intent	765	85.4	91.0	85.4	91.0
BilCat	news topic	400	79.0	90.0	75.8	88.0
complaints	complaint topic	400	93.2	96.5	93.5	96.2
Turkish SMS	spam	400	99.5	100.0	96.8	100.0
TQuAD-2	reading	1,377	97.5	99.1	97.5	99.1
XQuAD-tr	reading	1,127	93.2	97.0	93.2	97.0
turkish-wikiNER	entity type	1,281	79.9	93.8	74.5	94.2
Starlang KeNet	word sense	917	59.5	65.9	59.5	65.9
Starlang PropBank	semantic role	486	67.5	92.0	67.5	92.0
Starlang FrameNet	frame	91	76.9	93.4	76.9	93.4
CorefUD Turkish-ITCC	coreference	236	75.4	88.1	75.4	88.1

Table 8: hukum-bench: 29 Turkish benchmarks recast as decisions. Accuracy in %.

Group	Task	Questions	hüküm1	hüküm1nlp
NLI	XNLI-tr (full test) [17]	5,010	74.4	82.6
	TrGLUE MNLI-m [18]	2,000	81.8	86.7
	TrGLUE MNLI-mm	2,000	88.3	92.3
	TrGLUE RTE	685	94.9	96.2
Pairs / classification	TrGLUE QNLI	2,000	89.8	88.3
	TrGLUE QQP	2,000	92.0	91.5
	TrGLUE MRPC	1,000	59.4	61.3
	TrGLUE SST-2	2,000	85.7	84.4
	TrGLUE CoLA	985	51.3	51.2
	news category	250	92.8	92.0
	TrClaim relevance	2,187	71.3	73.2
	TrClaim check-worthiness	2,187	54.7	46.2
	X-FACT-tr [27]	169	36.1	33.1
Commonsense / procedural	XCOPA-tr [24]	500	89.6	89.4
	Global PIQA, non-parallel [25]	100	75.0	73.0
	Global PIQA, parallel	103	44.7	47.6
	PLU goal inference [26]	837	76.1	74.2
	PLU next event	655	71.0	71.0
	PLU step inference	612	77.8	75.3
	PLU step ordering	1,021	61.8	65.7
Knowledge	TR-MMLU [28]	6,200	54.6	54.1
	TurkishMMLU [29]	899	48.8	44.9
	Global-MMLU-tr [30]	2,000	50.0	48.8
	INCLUDE-tr [31]	548	56.2	55.3
	EXAMS-tr [32]	393	58.8	58.5
Language	TurBLiMP [33]	2,000	65.1	69.3
	TokSuite-tr [34]	621	88.4	88.9
	riddles (BilmeceBench)	442	56.8	61.1
	circumflex	70	61.4	58.6
Mean (29 tasks)			69.3	69.5

7.3 hukum-bench

7.4 Comparison with other Turkish decision models

We ran four other open Turkish decision models on the same questions, each with its own inference code and its own calibration (Table 9). To check our adapters we reproduced numbers from their model cards: ufakzeka-karar scores 95.2 on HakemBench moderation (card: 95.2) and Karar-4B 58.2 on TurkishMMLU (card: 58.5).

Table 9: Same questions, six Turkish decision models. Knowledge = mean of TR-MMLU, TurkishMMLU, Global-MMLU-tr, INCLUDE-tr, EXAMS-tr and MMLU-Pro-TR. Accuracy in %.

Model	Backbone	hukum-bench (29)	Knowledge (6)	HakemBench 7 tracks	HakemBench human-gold
hüküm1	Turkish-Gemma-4b + LoRA	69.3	49.3	78.0	67.7
hüküm1nlp	hüküm1 + NLP LoRA	69.5	47.7	79.1	63.1
Karar-4B	Qwen3.5-4B based + LoRA	70.3	56.2	78.8	66.9
ufakzeka-karar	182M causal LM	47.9	23.7	75.8	64.2
laya-tr-karar	mmBERT-base encoder	51.0	22.9	49.5	41.5
saturday-labs/karar	modernbert-tr-base	51.1	26.2	52.1	47.3

The 4B models are within a point of each other on hukum-bench; Karar-4B, built on a Qwen3.5-based backbone, leads hüküm1 by 6.9 points on the knowledge tests. On NLI and yes/no tasks hüküm1 and hüküm1nlp lead, partly because TrGLUE-derived questions are in hüküm1’s training data and TrGLUE MNLI/RTE and XNLI-dev are in hüküm1nlp’s. On the 260 human-gold HakemBench questions the four strongest models lie within the ± 6 -point interval.

7.5 Forgetting

Stage 2 improved the NLP sets but, despite 25% replay, lowered the general decision sets by 1–5 points (Table 6): typed-decisions -2.0 , cisimcik -1.9 , MMLU-Pro-TR -3.1 and HakemBench human-gold -4.6 , while the HakemBench 7-track mean rose by 1.1. Among the hukum-bench tasks, TrClaim check-worthiness (-8.5) and TurkishMMLU (-3.9) dropped most. hüküm1nlp is therefore the better model for Turkish text-understanding decisions, and hüküm1 for general and human-gold judgement decisions.

8 Known contamination

hüküm1’s training data contains items from evaluation splits of public benchmarks: all 1,800 Belebele (tur_Latn) items, 285 MMLU few-shot dev questions, 57 XCOPA validation rows and ~ 440 GLUE MNLI dev/test premises. hüküm1nlp’s stage-2 data additionally contains 830 XNLI-tr rows built from the XNLI validation split and 265 Belebele passages in the replay part. Scores on Belebele, MMLU dev, XNLI validation and MNLI dev for these models are not clean; none of these splits is used in the results above (XCOPA and XNLI are reported on their test splits).

9 Limitations

- One seed per model; small sets carry ± 3 –5 points of noise.
- Knowledge is bounded by the 4B backbone; decision training does not add facts.
- Several labels are silver: absa-tr is LLM consensus, parts of TrGLUE are LLM-assisted, Turkish Decisions Max uses LLM-panel labels.

- Grammaticality judgements are weak (CoLA near chance, TurBLiMP 65–69); comparing two sentences side by side may suit the task less than sentence likelihood.
- Text only: the backbone’s vision tower is not used by the decision forward pass.
- The models answer every question they are asked; they neither abstain nor explain.

10 License

The weights derive from Gemma and are distributed under the Gemma Terms of Use and Prohibited Use Policy [3], which among other things prohibit automated decisions that affect individuals’ rights in areas such as finance, law, employment, health, housing, insurance and social welfare. Some training sources are non-commercial or share-alike; we recommend research and other non-commercial use. The training data is not distributed.

References

- [1] YTU COSMOS. Turkish-Gemma-4b-T1-Scout. <https://huggingface.co/ytu-ce-cosmos/Turkish-Gemma-4b-T1-Scout>, 2026.
- [2] Gemma Team. Gemma 3 technical report. arXiv:2503.19786, 2025.
- [3] Google. Gemma Terms of Use and Gemma Prohibited Use Policy. <https://ai.google.dev/gemma/terms>.
- [4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.
- [5] D. Han, M. Han, and the Unsloth team. Unsloth. <https://github.com/unslothai/unsloth>.
- [6] Cloudflare. Clef: joint-schema decision head, 2026.
- [7] TypeSafe. Jev: a System One decision model. <https://typesafe.ai>, 2026.
- [8] D. Sileo. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *LREC-COLING*, pages 15655–15684, 2024.
- [9] A. V. Biswas. BEV decision mix. <https://huggingface.co/datasets/avbiswas/bev-decision>, 2026.
- [10] U. S. Praveen. jev-bench. <https://huggingface.co/datasets/Praveenrajus/jev-bench>, 2026.
- [11] ahmetege. turkish_jev_noul. https://huggingface.co/datasets/ahmetege/turkish_jev_noul, 2026.
- [12] M. Nazeem. Hmm: a small open model for typed decisions. <https://github.com/n4ze3m/hmm>, 2026.
- [13] cisimcik. Turkish Decisions Max 19K. <https://huggingface.co/datasets/cisimcik/turkish-decisions-max-19k>, 2026.
- [14] LocalLLaMA. typed-decisions. <https://huggingface.co/datasets/LocalLLaMA/typed-decisions>, 2026.

- [15] YTU COSMOS. absa-tr. <https://huggingface.co/datasets/ytu-ce-cosmos/absa-tr>, 2026.
- [16] A. U. Öztürk, Y. Cemek, P. Karagoz. IronyTR: Irony detection in Turkish informal texts. *International Journal of Intelligent Information Technologies*, 2021.
- [17] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *EMNLP*, 2018.
- [18] D. Altınok. Introducing TrGLUE and SentiTurca: A comprehensive benchmark for Turkish general language understanding and sentiment analysis. arXiv:2512.22100, 2025.
- [19] J. FitzGerald et al. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *ACL*, 2023.
- [20] O. Karasoy, S. Ballı. Spam SMS detection for Turkish language with deep text analysis and deep learning methods. *Arabian Journal for Science and Engineering*, 2022.
- [21] F. C. Akyon et al. Automated question generation and question answering from Turkish texts. arXiv:2111.06476, 2022.
- [22] M. Artetxe, S. Ruder, D. Yogatama. On the cross-lingual transferability of monolingual representations. In *ACL*, 2020.
- [23] M. Novák et al. Findings of the third shared task on multilingual coreference resolution. In *CRAC*, 2024.
- [24] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, A. Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *EMNLP*, 2020.
- [25] T. A. Chang et al. Global PIQA: Evaluating physical commonsense reasoning across 100+ languages and cultures. arXiv:2510.24081, 2025.
- [26] A. Uzunoğlu, G. G. Şahin. Benchmarking procedural language understanding for low-resource languages: A case study on Turkish. In *IJCNLP-AACL*, 2023.
- [27] A. Gupta, V. Srikumar. X-FACT: A new benchmark dataset for multilingual fact checking. In *ACL*, 2021.
- [28] M. A. Bayram et al. Setting standards in Turkish NLP: TR-MMLU for large language model evaluation. arXiv:2501.00593, 2025.
- [29] A. Yüksel, A. Köksal, L. K. Şenel, A. Korhonen, H. Schütze. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of EMNLP*, 2024.
- [30] S. Singh et al. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. arXiv:2412.03304, 2024.
- [31] A. Romanou et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. arXiv:2411.19799, 2024.
- [32] M. Hardalov, T. Mihaylov, D. Zlatkova, Y. Dinkov, I. Koychev, P. Nakov. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *EMNLP*, 2020.
- [33] E. Başar, F. Padovani, J. Jumelet, A. Bisazza. TurBLiMP: A Turkish benchmark of linguistic minimal pairs. In *EMNLP*, 2025.
- [34] TokSuite authors. TokSuite. arXiv:2512.20757, 2025.