Turkish Language Capabilities of Large Language Models: A Comprehensive Performance Analysis

Furkan AKSOY
Faculty of Engineering and Natural Sciences
Maltepe University
Istanbul, Turkey
furkanaksoy178@gmail.com

***Abstract*—This paper presents a comprehensive evaluation of five state-of-the-art large language models (LLMs) on Turkish language tasks. We assess GPT-5 Thinking, Claude 4 Sonnet, Gemini 2.5 Pro, Grok 3, and DeepSeek R1 across six categories: spelling accuracy, grammar, cultural understanding, logical reasoning, standardized exam performance, and text comprehension. Using a dataset of 500 carefully designed test questions based on current Turkish Language Institution (TDK) standards, we evaluate each model's proficiency in handling the morphologically rich and agglutinative structure of Turkish. Our results show that GPT-5 Thinking (91.2% overall accuracy) and Gemini 2.5 Pro (91% overall accuracy) significantly outperform other models. While all models demonstrate strong performance in basic grammar rules (85-95% accuracy), they show considerable variation in cultural context understanding (70-100% range). The study reveals that Turkish's structural complexity, including vowel harmony and extensive suffix system, poses unique challenges for current LLMs. We provide practical recommendations for model selection in different applications and identify key areas for improvement in Turkish language processing.**

***Index Terms*—Large Language Models, Turkish Language Processing, Natural Language Understanding, Model Evaluation, Morphological Analysis, Multilingual AI**

I. Introduction

Large Language Models (LLMs) have revolutionized natural language processing, achieving remarkable performance across various linguistic tasks [1]. However, the evaluation of these models on morphologically rich languages like Turkish remains limited. Turkish presents unique challenges due to its agglutinative structure, complex vowel harmony system, and rich morphological features [2].

The motivation for this study stems from the increasing adoption of LLMs in Turkish-speaking applications across education, media, legal, and technology sectors. Despite this growing usage, there exists a lack of comprehensive evaluation frameworks specifically designed for Turkish language capabilities of modern LLMs.

This paper makes the following contributions:

- A comprehensive evaluation framework for Turkish language capabilities of LLMs
- Performance analysis of five state-of-the-art models across six distinct categories
- Identification of specific challenges posed by Turkish morphological complexity
- Practical recommendations for model selection in real-world applications
- Open-source dataset and methodology for reproducible research

II. Related Work

Early work on computational processing of Turkish focused on morphological analysis and disambiguation [2], [3]. Recent advances in neural language models have shown promise for morphologically rich languages [4].

Previous evaluations of multilingual language models have primarily focused on high-resource languages [5], [6]. Limited work exists on comprehensive Turkish language evaluation, with most studies focusing on specific tasks such as machine translation [7] or named entity recognition [8].

Our work differs from previous studies by providing a holistic evaluation framework that covers multiple aspects of Turkish language understanding, from basic orthographic rules to cultural context comprehension.

III. Methodology

A. Model Selection

We evaluate five contemporary LLMs representing different architectural approaches and training methodologies:

GPT-5 Thinking (OpenAI): Advanced reasoning-capable model with visible thinking processes, optimized for complex problem-solving tasks.

Claude 4 Sonnet (Anthropic): Constitutional AI-trained model emphasizing speed-accuracy balance and ethical considerations.

Gemini 2.5 Pro (Google): Multimodal model with extensive multilingual training, particularly strong in morphologically complex languages.

Grok 3 (xAI): Real-time internet-enabled model with dynamic information access capabilities.

DeepSeek R1: Open-source reasoning model with detailed thinking processes and etymological analysis capabilities.

B. Test Design

Our evaluation framework consists of six categories designed to assess different aspects of Turkish language competency:

1) Test 1: Spelling Accuracy: 200 words (100 correct, 100 with common errors) based on current TDK guidelines. Models respond with binary classification and provide corrections for misspelled words.

2) Test 2: Basic Grammar: 100 questions across seven subcategories:

- De/Da conjunction usage (15 questions)
- Ki conjunction identification (15 questions)
- Question particle harmony (15 questions)
- Vowel harmony rules (15 questions)
- Consonant assimilation (15 questions)
- Punctuation and formatting (15 questions)
- Morphological analysis (10 questions)

3) Test 3: Cultural Understanding: 50 questions covering Turkish proverbs and idioms, requiring both meaning explanation and contextual usage examples.

4) Test 4: Logical Reasoning: 40 questions split between verbal logic (20) and numerical logic (20), testing analytical thinking in Turkish context.

5) Test 5: Standardized Exam Performance: 40 questions from YKS TYT Turkish practice exams, including paragraph analysis and comprehensive grammar assessment.

6) Test 6: Text Comprehension: 10 complex texts featuring code-switching, archaic Turkish expressions, and mixed linguistic elements.

C. Evaluation Criteria

All evaluations follow current Turkish Language Institution (TDK) standards as of 2025. Scoring uses exact match for objective questions and expert validation for subjective responses. Response times are recorded to assess efficiency alongside accuracy.

IV. Results

A. Overall Performance

Table I presents the comprehensive performance across all test categories. GPT-5 Thinking and Gemini 2.5 Pro emerge as clear leaders with over 91% overall accuracy.

TABLE I
OVERALL PERFORMANCE RESULTS (% ACCURACY)

Model	T1	T2	T3	T4	T5	T6	Avg
GPT-5 Thinking	94.0	93.0	91.0	88.0	88.0	92.0	**91.2**
Gemini 2.5 Pro	91.5	92.0	98.0	80.0	85.0	88.0	**91.0**
DeepSeek R1	86.0	88.0	80.0	75.0	80.0	84.0	82.2
Claude 4 Sonnet	83.5	86.0	77.0	69.0	78.0	82.0	79.3
Grok 3	80.0	83.0	73.0	63.0	95.0	78.0	78.7

B. Detailed Category Analysis

1) Spelling Accuracy (Test 1): All models demonstrated competency in basic spelling rules, with GPT-5 Thinking achieving the highest accuracy (94%). DeepSeek R1 showed unique etymological analysis capabilities, providing detailed word origin explanations alongside corrections.

2) Grammar Assessment (Test 2): Performance was consistently high across all models (83-93%), indicating adequate representation of Turkish morphological rules in training data. Notable similarities between GPT-5 Thinking and DeepSeek R1 in punctuation usage suggest potential shared training sources.

TABLE II
GRAMMAR SUBCATEGORY RESULTS (% ACCURACY)

Subcategory	GPT-5	Gemini	DeepSeek	Claude	Grok
De/Da Usage	96	94	90	88	84
Ki Conjunction	94	92	88	86	82
Question Particle	98	96	92	90	88
Vowel Harmony	96	94	90	88	86
Consonant Rules	94	92	88	86	84
Punctuation	80	88	82	78	74
Morphology	92	90	86	84	82

3) Cultural Understanding (Test 3): Significant performance variation emerged in this category (73-98%), highlighting the impact of training data cultural representation. Gemini 2.5 Pro achieved perfect performance in proverbs, while Grok 3 showed the lowest cultural comprehension.

4) Logical Reasoning (Test 4): All models struggled with numerical logic in Turkish context, with performance dropping significantly from verbal to numerical tasks. This suggests challenges in cross-lingual transfer of mathematical reasoning.

5) Standardized Exam Performance (Test 5): Grok 3's exceptional performance (95%) was attributed to internet search capabilities rather than inherent language understanding, as the model accessed online answer keys rather than solving problems analytically.

6) Text Comprehension (Test 6): Models showed consistent performance in code-switching scenarios (78-92%), indicating robust multilingual understanding capabilities. GPT-5 Thinking excelled in complex text interpretation tasks.

C. Response Time Analysis

Table III shows significant variation in processing times, with Claude 4 Sonnet prioritizing speed while DeepSeek R1 emphasizes thorough analysis.

TABLE III
AVERAGE RESPONSE TIMES BY MODEL

Model	Avg Time (min)	Range (min)
Claude 4 Sonnet	2.5	1-4
Gemini 2.5 Pro	6.0	3-9
GPT-5 Thinking	10.5	8-13
Grok 3	8.0	5-12
DeepSeek R1	15.0	10-18

V. Discussion

A. Turkish Language Complexity Impact

Our analysis reveals that Turkish's agglutinative structure poses specific challenges for current LLMs. Words like "çalıştırabileceklerimizden" (from those we could make work)

contain seven morphemes, each contributing semantic and grammatical information. Models showed varying success in decomposing and understanding such complex structures.

Vowel harmony rules, particularly the interaction between major and minor harmony patterns, created difficulties for all models. The simultaneous application of vowel harmony and consonant assimilation in long word forms tested the limits of morphological processing capabilities.

B. Cultural Context Limitations

The significant performance variation in cultural understanding (70-100% range) highlights a critical limitation in current LLMs. Turkish cultural expressions, deeply rooted in Islamic tradition and Ottoman heritage, require contextual knowledge that appears underrepresented in training data.

This finding has important implications for applications requiring cultural sensitivity, such as educational content generation, literary translation, and social media content moderation.

C. Tool Integration Strategies

Models demonstrated different approaches to information access. Grok 3's systematic internet searching for unknown content contrasts with GPT-5 Thinking's reliance on internal knowledge. This variation suggests that optimal Turkish language processing may require hybrid approaches combining multiple models and external resources.

D. Training Data Quality Issues

Evidence of outdated training data emerged particularly in Claude 4 Sonnet's adherence to pre-2019 TDK rules. This highlights the importance of keeping training data current with evolving language standards, especially for officially regulated languages like Turkish.

VI. Practical Implications and Recommendations

A. Application-Specific Model Selection

Based on our findings, we provide specific recommendations:

Professional Writing: Multi-stage approach using Gemini 2.5 Pro for initial drafting followed by GPT-5 Thinking for logical consistency checking.

Journalism and Media: Grok 3 for real-time information access and fact-checking, supplemented by other models for linguistic accuracy.

Educational Applications: Gemini 2.5 Pro for basic language instruction, GPT-5 Thinking for advanced analytical tasks.

Cultural Content: Gemini 2.5 Pro shows superior cultural understanding, making it suitable for applications involving traditional Turkish expressions.

B. Models to Avoid for Specific Tasks

Our analysis identifies scenarios where certain models should be avoided:

- Grok 3: Not suitable for standalone spelling control (80% accuracy)
- Claude 4 Sonnet: Inadequate for cultural content (70-80% cultural accuracy)
- DeepSeek R1: Inappropriate for real-time applications (15-minute response times)

VII. Limitations and Future Work

This study has several limitations that should be addressed in future research:

Sample Size: While comprehensive, the 500-question dataset could be expanded to 1000+ questions for better statistical significance.

Model Coverage: Additional models, including domain-specific Turkish LLMs, should be evaluated.

Task Diversity: Future evaluations should include creative writing, technical translation, and domain-specific tasks.

Longitudinal Analysis: Tracking model performance changes over time as training data and architectures evolve.

Human Evaluation: Expert linguist evaluation alongside automated scoring would strengthen result validity.

VIII. Conclusion

This comprehensive evaluation of five state-of-the-art LLMs on Turkish language tasks reveals significant insights into current capabilities and limitations. GPT-5 Thinking (91.2%) and Gemini 2.5 Pro (91.0%) emerge as the most capable models overall, though no single model excels across all categories.

Key findings include:

- Strong performance in basic grammar (85-95%) across all models
- Significant variation in cultural understanding (70-100% range)
- Challenges in numerical reasoning within Turkish contexts
- Evidence of training data quality issues affecting some models
- Different tool integration strategies impacting performance

The study demonstrates that Turkish's morphological complexity and cultural richness require specialized attention in LLM development. While current models show promise, achieving near-perfect Turkish language processing will require enhanced training data quality, better cultural representation, and potentially hybrid approaches combining multiple models.

Our open-source dataset and evaluation framework provide a foundation for continued research in Turkish language AI, supporting the development of more culturally aware and linguistically accurate models for Turkish-speaking applications.

IX. Data Availability

All test data, detailed methodologies, and evaluation scripts are made available under open-source license at: https://github.com/FurkanAksoyy/tr-llm-turkce-performans-analizi

References

[1] T. Brown et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.

[2] K. Oflazer, "Two-level description of Turkish morphology," *Literary and linguistic computing*, vol. 9, no. 2, pp. 137-148, 1994.

[3] D. Hakkani-Tür, K. Oflazer, and G. Tür, "Statistical morphological disambiguation for agglutinative languages," *Computers and the Humanities*, vol. 36, no. 4, pp. 381-410, 2002.

[4] K. Kann et al., "Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 47-57, 2018.

[5] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440-8451, 2020.

[6] J. Hu et al., "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," *International Conference on Machine Learning*, pp. 4411-4421, 2020.

[7] J. Tiedemann and S. Thottingal, "OPUS-MT—Building open translation services for the world," *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479-480, 2020.

[8] G. G. Şahin and M. Steedman, "Data augmentation via dependency tree morphing for low-resource languages," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5004-5009, 2018.

[9] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171-4186, 2019.

[10] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842-866, 2020.