applied sciences

MDPI

Article

DPATransLLM: Detection of Pronominal Anaphora in Turkish Sentences Using Transformer-Based, Large Language Models and Hybrid Ensemble Approach

Engin Demir [1,2] and Metin Bilgin [1,*]

[1] Department of Computer Engineering, Bursa Uludag University, Bursa 16059, Turkey; engin.demir@msu.edu.tr
[2] Department of Information Technology, National Defence University, Istanbul 34353, Turkey
* Correspondence: metinbilgin@uludag.edu.tr; Tel.: +90-224-275-52-63; Fax: +90-224-294-19-03

Abstract

In the current information age, with the exponential growth of data volume and language-based applications, the accurate resolution of intra-contextual relationships in texts has become indispensable for both academic research and industrial Natural Language Processing (NLP) systems. This study focuses on the detection of pronominal anaphora in Turkish sentences. For the detection of pronominal anaphora, a specific dataset comprising 2000 sentences and 72,239 tokens was created, and this dataset was labeled using a BIO tagging method developed with a custom approach for this study. In this work, fine-tuning was performed on Transformer-based language models pre-trained on Turkish data, such as BERT and RoBERTa. Additionally, Large Language Models (LLMs) trained on Turkish data, including Turkcell-LLM-7b-v1 and ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1, as well as multilingual models like Microsoft's Phi-3 Mini-4K-Instruct and OpenAI's GPT-4o-mini, were also fine-tuned with the created dataset to detect pronominal anaphora in sentences. Following the training of the language models, the resulting performance was evaluated using pronoun accuracy, antecedent accuracy, exact match, and F1-score metrics. According to the results obtained in the pronominal anaphora detection phase of the study, a novel hybrid ensemble approach combining multiple Transformer models with linguistic rules achieved the highest performance. This hybrid system attained scores of 0.987 for pronoun accuracy, 0.977 for antecedent accuracy, 0.505 for exact match, and 0.960 for F1-score, surpassing all individual models, including GPT-4o-mini. These findings reveal the superiority of ensemble methods combined with Turkish-specific linguistic rules over standalone models in Turkish anaphora resolution. This study is considered novel, as it is the first work to apply hybrid ensemble methods with linguistic rule integration to this domain for the Turkish language.

Keywords: pronoun; anaphora; transformer-based language model; large language model; hybrid ensemble; natural language processing

Academic Editors: Yongrui Chen, Guilin Qi and Hui Yang

Received: 26 October 2025
Revised: 12 November 2025
Accepted: 21 November 2025
Published: 25 November 2025

Citation: Demir, E.; Bilgin, M. DPATransLLM: Detection of Pronominal Anaphora in Turkish Sentences Using Transformer-Based, Large Language Models and Hybrid Ensemble Approach. *Appl. Sci.* **2025**, *15*, 12480. https://doi.org/10.3390/app152312480

1. Introduction

The exponential increase in textual data in the modern era necessitates the development of Natural Language Processing (NLP) technologies. The accurate resolution of linguistic structures and contextual relationships in texts is of critical importance for both theoretical research and practical applications. One of the fundamental challenges encountered in this context is the phenomenon of pronominal anaphora. Anaphora is the

process of referring back to a previously mentioned entity in subsequent expressions and ensures the semantic integrity of the text. The automatic detection of these relationships is fundamental to the success of many NLP tasks, such as text understanding, machine translation, and question-answering systems.

Turkish, with its rich morphological structure, pro-drop feature, and flexible syntax, presents unique challenges for anaphora resolution. Despite this, studies in this area for Turkish are quite limited, and comprehensive research using modern deep learning approaches is insufficient. Although the Transformer architecture [1] and Large Language Models (LLMs) [2] have revolutionized the NLP field in recent years, the performance of these models on Turkish anaphora resolution has not been systematically evaluated.

This study presents a comprehensive approach for the detection of Turkish pronominal anaphora. In this work, a novel dataset of 2000 sentences was created, an innovative BIO-based tagging scheme was developed, and a systematic comparison was conducted among Transformer models like BERT and RoBERTa, as well as Large Language Models such as Turkcell-LLM, Turkish-Llama, Microsoft Phi-3, and OpenAI GPT-4o-mini. Furthermore, a hybrid ensemble approach combining multiple models with Turkish-specific linguistic rules was developed and evaluated.

The aim of this study is to identify the most effective models for Turkish pronominal anaphora detection and to establish a solid foundation for future research. The findings obtained hold significant value for both academic and practical Turkish NLP applications.

2. Related Works

Anaphora resolution, as one of the fundamental problems in the field of Natural Language Processing [3], has been addressed in the literature through various approaches and across different languages. In this section, studies on anaphora resolution conducted for Turkish and other languages are presented in a chronological and thematic order.

One of the pioneering studies in the field of Turkish anaphora resolution was conducted by Tüfekci and Kılıçaslan [4]. In this study, the aim was to adapt Hobbs's Naïve Algorithm for Turkish, and the algorithm was tested on 10 sentences. The researchers reformulated the algorithm and added new rules to make it suitable for the morphological features of Turkish. The study proposed a system that determines possible antecedents for pronoun resolution based on syntactic information.

Similarly, Sezer and Kılıçaslan [5] developed a rule-based pronominal anaphora resolution system for Turkish. This work constitutes an important example of how linguistic rules were used for this problem before machine learning approaches became popular.

Yıldırım [6] presented a comprehensive machine learning approach for Turkish anaphora resolution. In the study, various machine learning algorithms such as decision trees, naive bayes, k-nearest neighbor (k-NN), support vector machines (SVM), and perceptrons were compared with a knowledge-based centering algorithm. In experiments conducted on a corpus of 10,000 words containing 1114 pronouns, the k-nearest neighbor algorithm and the centering algorithm yielded the most successful results.

Taze [7] examined the performance of deep learning networks by reducing the anaphora resolution problem to pronoun resolution. In the study, conducted on a dataset containing 593 correct example pairs compiled from Turkish children's stories, different configurations of multi-layer perceptron networks were tested. The highest performance was achieved with a network architecture that included numerous neurons in each layer and a moderate number of layers.

In the international literature during this period, end-to-end neural network models that do not require feature engineering and produce more successful results came to the

forefront [8]. Such models later formed the basis for Transformer-based approaches like BERT.

Abacı et al. [9] analyzed anaphoric relations in Turkish tweets, examining the unique structure of social media texts. In their study, two different datasets were used: 20 Turkish children's stories and 227 tweets. J48, Voted Perceptron, SVM, Naive Bayes, and k-NN algorithms were compared, with the J48 algorithm showing the best performance. The research findings suggested that anaphora resolution in less structured texts like tweets could be more successful compared to structured texts like stories or novels.

Ertan [10] presented a heuristic-based approach for pronominal coreference analysis. Mitkov's traditional knowledge-based algorithm was tested on 364 English sentences from the TED MDB corpus, aligned with their Turkish translations. The study achieved an F1 score of 0.61 on the English sentences and 0.63 on the Turkish sentences.

Another significant study on anaphora resolution in non-English languages was conducted by Palomar and Martínez-Barco [11] on Spanish dialogs. This study aimed to resolve pronominal anaphora by leveraging the unique structure of dialogs (e.g., turn-taking, adjacency pairs). The researchers presented a hybrid algorithm (ARDi) that defined an "anaphoric accessibility space" based on dialog structure, along with linguistic constraints and preferences. The work stands out for its proposal of a system that uses structural context information to resolve referential relations, particularly in dialog texts.

One of the important studies addressing pronominal anaphora and zero anaphora between Spanish and English was conducted by Peral and Ferrández [12]. This study aimed to overcome the challenges in machine translation caused by structural differences between the languages, especially the pro-drop feature in Spanish. The authors developed a system called AGIR, which uses a parser-based and interlingual representation. The study proposes a translation system that detects and resolves both pronominal and, notably, zero anaphora to ensure their accurate transfer into the target language.

A study offering a modern and conceptual approach to the anaphora resolution problem was carried out by de Marneffe, Recasens, and Potts [13]. This work aimed to predict whether an entity in a text would be mentioned only once (singleton) or would be part of a recurring coreference chain (coreferent). The authors developed a "lifespan model" using the semantic context (negation, modality, etc.) and grammatical features of a mention. They demonstrated that this model improved the performance of coreference resolution systems by narrowing their search space. The study presents a model that can be used as a pre-filtering step to simplify the resolution task.

Paparounas and Akkuş [14], approaching from a linguistic theory perspective, examined the interaction of complex pronouns with anaphors in the nominal domain. In their analysis, conducted using minimalist approaches within the framework of Chomsky's Binding Theory Condition A, the role of f-features and the Agree operation in Turkish anaphora resolution processes was supported by empirical data.

In the field of Arabic anaphora resolution, Bouzid and Zribi [15] proposed a hybrid approach combining Q-learning-based reinforcement learning and word embedding models for pronominal and zero anaphora resolution. In an evaluation on sentences containing 478 pronominal anaphors and 236 zero anaphors, the Q-learning+Word2Vec combination achieved a success rate of 79.37% for pronominal anaphora resolution and 70.82% for zero anaphora resolution.

Murayshid et al. [16] developed a sequence-to-sequence architecture with an AraBERT encoder and a Bi-LSTM decoder for Arabic pronoun resolution. The model, tested on the AnATAr dataset, demonstrated superior performance with 99.6% accuracy and a 71% F1 score. The study showed the effectiveness of deep learning approaches for morphologically rich languages.

Kongwan et al. [17] proposed a two-stage method for anaphora resolution in Thai texts, integrating it with Elementary Discourse Unit (EDU) segmentation. In their work, which handled non-referential and referential anaphora separately, they achieved values of 77% precision, 84% recall, and 81% F-score.

Chen [18] evaluated the performance of the Transformer-based ChatGPT-4 model on Chinese pronoun anaphora resolution. In tests conducted on a new collection of 284 Winograd Schemas derived from the Chinese WSC285 and Mandarinograd datasets, the model achieved an accuracy of 91.5%, which is close to human performance (97%). The study demonstrated that large language models possess effective anaphora resolution capabilities in languages other than English.

Indeed, recent studies indicate that large language models (LLMs), through their zero-shot or few-shot learning capabilities, exhibit considerable potential in coreference resolution tasks across different languages [2].

Upon reviewing the existing literature, it is evident that modern Transformer-based models and Large Language Models have not been systematically evaluated for Turkish anaphora resolution. This study aims to fill this gap, thereby offering a novel contribution to the literature.

3. Materials and Methods

This section presents information regarding the preparation and labeling of the dataset, as well as the language models and methods employed. The experimental procedures were implemented using the Python version 3.12.12 programming language (Python Software Foundation, Wilmington, DE, USA) and the PyTorch version 2.9.0+cu126 library (Meta Platforms, Inc., Menlo Park, CA, USA) along with the Hugging Face Transformers library (Hugging Face, Inc., New York, NY, USA). All model training and evaluation tasks were conducted on [Insert Your GPU Model Here, e.g., NVIDIA A100] GPUs (NVIDIA Corporation, Santa Clara, CA, USA). The flowchart of the study is provided in Figure 1.

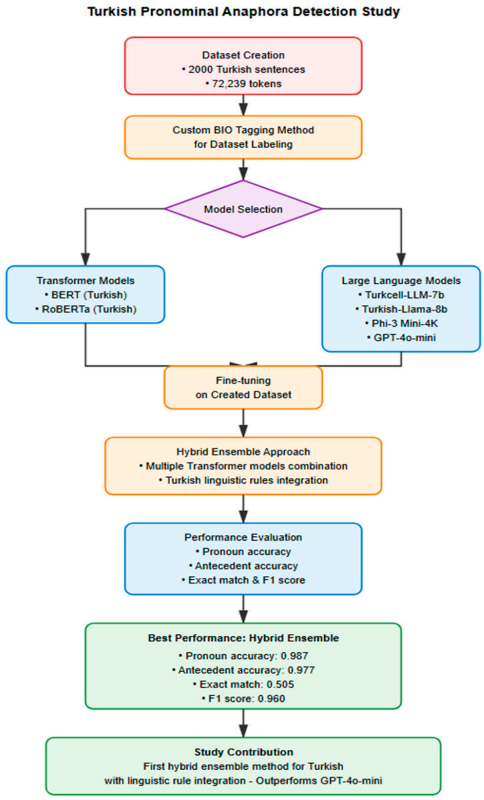

Figure 1. Flow Chart.

3.1. Dataset Preparation and Labeling Process

In this study, the dataset was manually created using samples from Turkish theses and articles in the field of linguistics, with a particular focus on those containing anaphora. The dataset, consisting of 2000 sentences, contains 72,239 tokens. The created dataset includes all types of pronouns to ensure that the model learns them, thereby enabling more accurate detection. The BIO tagging method was selected to facilitate better learning by the model and to ensure its focus on the primary elements: the antecedent and the pronoun. The BIO (Begin-Inside-Outside) scheme is used to specify the boundaries and the entity type of each element in a sequence. This tagging method was preferred because the BIO format defines pronoun and antecedent regions with precise boundaries, providing a standard and effective foundation for both the model training and result evaluation stages in pronominal anaphora detection.

To ensure the quality and consistency of annotations, the dataset preparation process involved consultation with expert linguists specializing in Turkish syntax and anaphora resolution. The annotation guidelines and a representative sample of annotated sentences were reviewed by these experts, who provided feedback on ambiguous cases and suggested refinements to the annotation scheme. Their input was incorporated to improve annotation consistency and ensure alignment with established linguistic principles. While formal inter-annotator agreement was not measured due to resource constraints, the iterative expert review process helped maintain annotation quality throughout the dataset construction.

The dataset sources were deliberately selected from linguistic papers and children's stories as these genres exhibit rich anaphoric relationships and complex syntactic structures suitable for detailed annotation. While this selection provides comprehensive coverage of formal written Turkish with diverse pronominal phenomena, it does not encompass other genres such as spoken language, news articles, or social media content.

Dataset Construction Criteria and Complexity Coverage: The dataset was systematically constructed to ensure comprehensive coverage of Turkish pronominal anaphora phenomena across varying syntactic complexity levels. The selection criteria prioritized sentences that exhibit:

1. Clause Nesting and Subordination: Sentences containing multiple embedded clauses (yan cümle) where pronouns may refer to antecedents across different syntactic levels. For example, in "Ahmet, Ayşe'nin ona verdiği kitabı okudu" (Ahmet read the book that Ayşe gave him), the pronoun "ona" (to him) crosses a relative clause boundary to refer to "Ahmet" in the main clause.
2. Subject Ellipsis (Pro-drop) Phenomena: Turkish exhibits extensive subject omission, creating implicit anaphoric relationships that require contextual inference. For example, in "Çocuk parkta oynadı. Ø Eve gelince çok yorgundu" (The child played in the park. Ø When [he] came home, [he] was very tired), zero subjects (Ø) require resolution to "çocuk" despite absence of explicit pronoun.
3. Long-Distance Dependencies: Pronouns separated from their antecedents by multiple intervening noun phrases, testing the models' ability to track referential chains. For instance, "Müdür, sekreterine, asistanının hazırladığı raporu ona iletmesini söyledi" (The director told his secretary to forward him the report that his assistant prepared) contains multiple possessive relationships and embedded clauses creating ambiguity in "ona" reference.
4. Demonstrative Pronoun Disambiguation: Sentences containing proximal ("bu"—this) and distal ("o/şu"—that) demonstratives requiring spatial/discourse context, such as "Kitap masadaydı. Bu çok ilginçti ama o biraz eskiydi" (The book was on the table. This was very interesting but that was a bit old).

5. Switch Reference and Topic Shift: Sentences where the topic or subject changes mid-discourse, requiring accurate tracking of referential continuity. For example, in "Ali Mehmet'i gördü. O çok mutluydu" (Ali saw Mehmet. He was very happy), ambiguity exists in whether "o" (he) refers to Ali or Mehmet.
6. Morphological Case Agreement: Sentences demonstrating Turkish's rich case system where pronoun-antecedent agreement must be verified. For instance, "Öğretmen öğrenciye soruyu sordu. Ona cevap vermesi için zaman verdi" (The teacher asked the student the question. [She] gave him time to answer it), where the dative pronoun "ona" must agree with "öğrenci" (student).

The dataset composition ensures balanced representation of these complexity levels: simple sentences with direct anaphora (25%), sentences with clause embedding (30%), complex sentences with multiple embeddings (10%), sentences with pro-drop phenomena (20%), and sentences with multiple pronouns requiring resolution (15%). This stratified sampling approach ensures that the trained models can generalize across the full spectrum of Turkish syntactic complexity encountered in real-world texts.

Some data from the dataset used in the training of the Transformer-based language models are presented in Table 1.

Table 1. Sample Sentence.

Sent_id	Token	BIO	Cluster
row1	Ali	B-ANT	1
row1	çok	O	-
row1	yorgundu	O	-
row1	.	O	-
row1	O	B-PRON	1
row1	,	O	-
row1	eve	O	-
row1	gelir	O	-
row1	gelmez	O	-
row1	uyudu	O	-
row1	.	O	-
row2	Zübeyde	B-ANT	2
row2	Hanım	I-ANT	2
row2	İzmir'e	O	-
row2	beyaz	O	-
row2	çarşafıyla	O	-
row2	geldi	O	-
row2	,	O	-
row2	ama	O	-
row2	o	B-PRON	2
row2	peçesizdi	O	-
row2	.	O	-

Due to the architecture of large language models, the dataset was converted into a different format. In this new format, while no changes were made to the labels from the original dataset structure, the format was adapted to enable fine-tuning on large language models. The data format used for the large language models is shown in Figure 2.

```
{'messages': [{'role': 'system',
   'content': 'Verilen cümledeki zamir artgönderimini tespit et.'},
  {'role': 'user',
   'content': 'Sınıfta birçok öğrenci var , bunlardan bazıları yurtdışından gelmiş .'},
  {'role': 'assistant',
   'content': "Bu cümlede 'bunlardan bazıları' zamiri 'birçok öğrenci' kelimesine/ifadeye referans veriyor."}]}
```

Figure 2. LLM data format (.json).

3.2. Model Training

In the experimental part of this study, contemporary Natural Language Processing models were used for the detection of pronominal anaphora. As the primary approach, all selected models were subjected to a fine-tuning process using the Turkish anaphora dataset developed within the scope of this study. The fine-tuning process enabled the models to leverage their pre-existing linguistic knowledge to acquire the ability to correctly identify pronoun-antecedent pairs within sentences. To facilitate a performance comparison, both Transformer-based models and Large Language Models (LLMs) were included in the analysis. The models that were fine-tuned and evaluated for performance in this study are listed below.

3.2.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-training method for language representation that has achieved state-of-the-art results on a wide variety of natural language processing (NLP) tasks. This model architecture, developed in 2018 at Google by Jacob Devlin Et Al., evaluates sentences both from right to left and from left to right. This allows it to better capture the relationships between words. The model was pre-trained on the BookCorpus and Wikipedia datasets [19].

The model is trained using two methods: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In the MLM technique, a masked word is predicted by leveraging the unmasked words provided as input. The MLM method primarily focuses on the relationships between words. The other method, NSP, focuses on the relationship between sentences. During the training phase with NSP, the model examines the relationship between two sentences to predict whether the second sentence is a continuation of the first.

In our study, one of the versions of the BERT model trained on Turkish data, "dbmdz/bert-base-turkish-cased" [20], was used. The general structure of this model is illustrated in Figure 3.

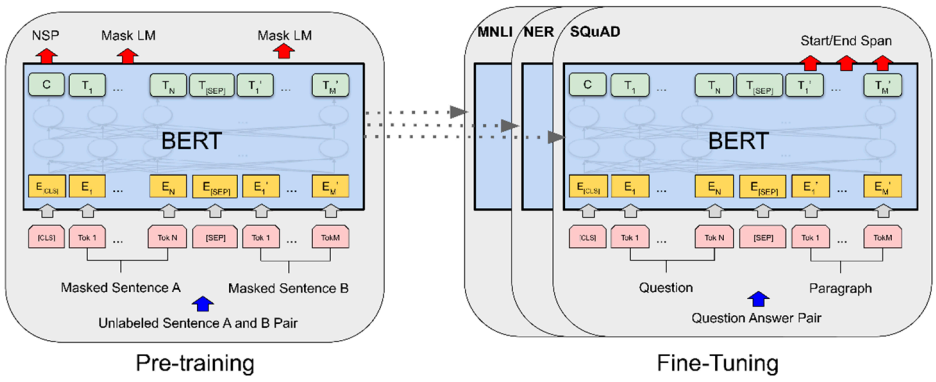

Figure 3. BERT Model Architecture [19]. The diagram depicts the multi-layer bidirectional Transformer encoder. The arrows indicate the flow of information from input embeddings through the hidden layers to the final output representations.

3.2.2. RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) is a language representation method that achieves higher performance on a wide variety of natural language processing (NLP) tasks by modifying the training strategies of the BERT architecture. This model, developed in 2019 at Facebook AI by Yinhan Liu Et Al., preserves the architectural design of BERT while making significant changes to the training phase.

RoBERTa is trained solely with the Masked Language Modeling (MLM) objective and a dynamic masking technique; that is, different words are masked in each training epoch,

enhancing the model's contextual generalization ability. The Next Sentence Prediction (NSP) sub-task used in BERT has been completely removed. This allows the model to avoid unnecessary task complexity while focusing on learning context.

The model was trained on a massive dataset of approximately 160 GB of uncompressed text, comprising the BookCorpus, CC-News, OpenWebText, and Stories datasets. Additionally, by using large batch sizes and extended training times, a representation space that masters deeper contextual relationships compared to BERT was achieved [21].

In our study, one of the versions of the RoBERTa model trained on Turkish data, "burakaytan/roberta-base-turkish-uncased", was used. The structural overview of this model is shown in Figure 4.

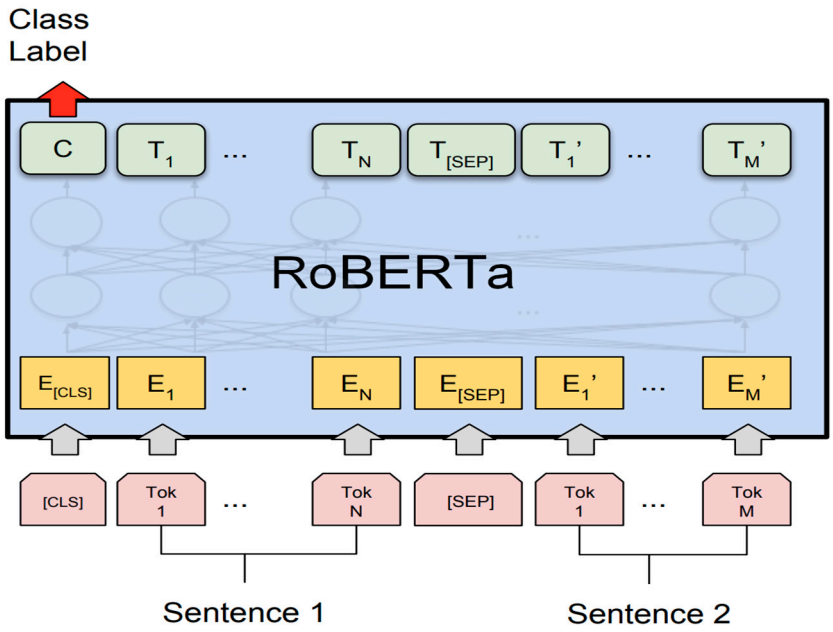

Figure 4. RoBERTa Model Architecture [21]. The diagram illustrates the flow of data through the model. The different colors distinguish between the input embeddings, the encoder layers (Transformer blocks), and the final output representations, highlighting the removal of the Next Sentence Prediction (NSP) task.

3.2.3. DeBERTa

DeBERTa (Decoding-enhanced BERT with Disentangled Attention) is an advanced transformer-based language model developed by Microsoft Research [22] that improves upon BERT's architecture through two key innovations: disentangled attention mechanism and enhanced mask decoder.

The disentangled attention mechanism represents each word using two vectors that separately encode content and position information. Unlike BERT, which fuses content and position information in a single embedding layer, DeBERTa computes attention weights using disentangled matrices for content-to-content, content-to-position, and position-to-content interactions. This architectural choice enables the model to better capture both the semantic meaning and relative positions of words in a sentence.

The enhanced mask decoder (EMD) incorporates absolute position embeddings into the final softmax layer for predicting masked tokens, providing additional positional context during pre-training. This design improves the model's ability to synthesize both local and global contextual information.

DeBERTa-v3 introduces further optimizations including ELECTRA-style pre-training (replaced token detection), gradient-disentangled embedding sharing between the encoder and decoder, and a more efficient vocabulary. The model was pre-trained on a large-

scale corpus consisting of Wikipedia, BookCorpus, OpenWebText, and CC-News, totaling approximately 160 GB of text data.

The DeBERTa-v3-small variant used in our study contains approximately 140 million parameters and employs 12 transformer layers with 768 hidden dimensions and 12 attention heads. While not originally trained on Turkish data, the model's multilingual architecture and robust attention mechanisms enable effective transfer learning capabilities for morphologically rich languages like Turkish.

In our study, the "microsoft/deberta-v3-small" model was fine-tuned on the Turkish pronominal anaphora dataset using the same BIO tagging scheme applied to BERT and RoBERTa. The model was selected for the hybrid ensemble specifically because its disentangled attention mechanism complements BERT's bidirectional encoding, enabling the ensemble to capture both local morphological patterns and long-range syntactic dependencies characteristic of Turkish sentence structure.

3.2.4. GPT-4o-Mini

GPT-4o mini (Generative Pre-trained Transformer 4 omni mini—the 'mini' version of the Omni architecture) is a robust and scaled-down language model developed by OpenAI in 2024 to offer the knowledge and reasoning capabilities of GPT-4o at a much lower cost. With its multimodal architecture supporting text and visual input/output, a 128 k token context window, and low latency capable of generating output in seconds, GPT-4o mini is positioned as a 'small model' that is over 60% cheaper than GPT-3.5 Turbo, yet surpasses it in reasoning benchmarks.

It was created using the model distillation technique, transferring the 'teacher' knowledge from GPT-4o to a smaller 'student' network; that is, it learns to mimic complex language patterns with fewer parameters by training on the outputs of the larger model. In benchmark results, GPT-4o mini has surpassed competing 'mini' models like Gemini Flash and Claude Haiku on tests such as MMLU (82%) and HumanEval (87.2%), as well as the multimodal MMMU test. It also outperforms GPT-3.5 Turbo, setting a new state-of-the-art benchmark among small-scale models [23].

In our study, the 'gpt-4o-mini' model, accessed Via the OpenAI API, was used for Turkish anaphora resolution.

3.2.5. Turkcell-LLM-7b-v1

Turkcell-LLM-7b-v1 (Turkcell Large Language Model—7 Billion Parameters, v1) is a large language model developed by the Turkcell R&D team in 2024, which retrains the Mistral 7B [24] architecture with a focus on Turkish. The architecture, which contains a total of ≈7 billion parameters, includes an expanded tokenizer adapted to the agglutinative morphology of Turkish, thereby reducing the number of sub-units to enable both faster and more accurate text generation.

The model follows a two-stage training process: DORA-based pre-training and LoRA-based instruction tuning. In the DORA-based pre-training phase, the Mistral blocks are re-initialized and trained on a massive raw corpus of 5 billion cleaned Turkish tokens. In the LoRA instruction tuning phase, the model is fine-tuned using the Low-Rank Adaptation method on various open-source and internally created Turkish instruction sets to enhance its performance in dialog, reasoning, and task completion [25].

In our study, the "TURKCELL/Turkcell-LLM-7b-v1" model, which was trained on Turkish data, was used.

3.2.6. Turkish-Llama-8b-DPO-v0.1

Turkish-Llama-8b-DPO-v0.1 (Turkish LLaMA–3 8B with Direct Preference Optimization—version 0.1 from the ytu-ce-cosmos research group) is a large language model with

≈8.03 billion parameters, developed in 2024 by the Yıldız Technical University Computer Engineering COSMOS Research Group. The underlying framework is Meta-Llama-3-8B [26], and the model has been retrained to capture the morphological richness specific to Turkish.

The model involves two main steps: full fine-tuning and DPO. In the fine-tuning step, the Meta-Llama-3-8B-Instruct model was fully retrained on a meticulously cleaned Turkish corpus of approximately 30 GB (including websites, books, news, etc.). In the DPO step, the fine-tuned model was aligned using the DPO algorithm with preference data consisting of synthetic and real user response pairs. The two resulting CosmosLLaMa-Instruct-DPO models were then merged to create the final Turkish-Llama-8b-DPO-v0.1 version [27].

In our study, the "ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1" model, which was trained on Turkish data, was used.

3.2.7. Microsoft Phi-3 Mini-4K-Instruct

Phi-3 Mini-4K-Instruct is a Small Language Model (SLM) with ≈3.8 billion parameters, released by Microsoft Research in April 2024. Its purpose is to offer reasoning capabilities approaching GPT-3.5 levels in a model lightweight enough to run in memory-constrained environments, including mobile phones.

The model was trained on the 3.3 trillion token "Phi-3" corpus. This corpus combines synthetic data containing high-quality reasoning sentences with heavily filtered public web pages. The goal is to achieve the most information-dense representation possible with a limited number of parameters. The model's compact structure makes it an attractive option for applications requiring low latency and for cost-effective solutions [28].

In our study, the "Microsoft Phi-3 Mini-4K-Instruct" model, which has a multilingual architecture, was used.

3.2.8. Hybrid Ensemble Approach

Post-processing rules specific to Turkish grammar are applied systematically after the neural models' predictions. These rules incorporate explicit Turkish grammatical constraints derived from formal linguistic theory:

1. Pronoun-Antecedent Agreement Rules:
   (a) Number Agreement: Singular pronouns ("o", "bu", "şu", "ona", "onu", "onun") require singular antecedents, while plural pronouns ("onlar", "bunlar", "şunlar") require plural antecedents. Pattern matching with morphological analyzers detects plurality markers (-lar/-ler).
   (b) Syntactic Position Constraints: Antecedents must c-command the pronoun in syntactic tree structure. For embedded clauses, pronouns can only access antecedents in dominating or coordinate clauses. A nearest-candidate heuristic prefers the syntactically closest noun phrase within the same or immediately dominating clause, with distance penalties for candidates separated by more than 3 clause boundaries.
2. Morphological Filtering and Validation:
   (a) Verbal Form Exclusion: Suffixes indicating verbal predicates are systematically excluded (progressive -yor, past tense -dı/-di/-du/-dü, evidential -mış/-miş/-muş/-müş, future -acak/-ecek), as Turkish pronouns refer to nominal entities, not verbal actions.
   (b) Conjunction and Particle Filtering: Coordinating conjunctions (ve, veya, ama, fakat), subordinating conjunctions (çünkü, eğer, ki), and discourse particles (da/de, bile, sadece) are excluded through closed-class lexicon lookup.

3. Case Marking Consistency: Turkish case system (Nominative, Accusative -ı/-i/-u/-ü, Dative -a/-e, Locative -da/-de, Ablative -dan/-den, Genitive -ın/-in/-un/-ün) requires pronouns and their antecedents to exhibit compatible case marking. For example, "Kitabı okudum, onu çok beğendim" demonstrates accusative matching. Nominative-accusative alternation is tolerated due to pro-drop and case stacking phenomena.
4. Person and Animacy Agreement: First person (ben/biz), second person (sen/siz), and third person (o/onlar) markers must align. Possessive suffixes require matching: "onun kitabı" (his/her book) requires 3rd person antecedent. Animacy hierarchy (Human > Animate > Inanimate) is applied when semantic tagging is available through named entity recognition.
5. Discourse Coherence Principles: Topic Continuity Preference applies Centering Theory algorithms (forward-looking centers, backward-looking centers). Parallelism Preference ensures that in coordinate structures like "Ali geldi ve Ø kahve içti" (Ali came and drank coffee), the zero pronoun preferentially resolves to the coordinate clause subject.
6. Conflict Resolution Hierarchy: When multiple rules conflict, precedence follows:
    1. Morphological agreement (number, case)—highest priority,
    2. Syntactic constraints (c-command, clause boundaries),
    3. Distance-based preferences (proximity heuristic),
    4. Discourse coherence (topic continuity),
    5. Semantic plausibility (animacy, selectional restrictions)—lowest priority. The integration of these rules with neural model predictions occurs through weighted voting: if linguistic rules strongly prohibit a candidate (e.g., number disagreement), that candidate is eliminated regardless of neural confidence scores. For ambiguous cases where rules permit multiple candidates, neural model probabilities determine the final selection.

3.3. Evaluation Metrics

In our study, pronoun accuracy, antecedent accuracy, exact match, and F1-score were used as evaluation metrics. Descriptions of these metrics are provided below.

Pronoun Accuracy: Measures the proportion of correctly identified pronouns.

$$Pron.\ Acc. = \frac{Total\ number\ of\ pronouns}{Number\ of\ correctly\ identified\ pronouns} \tag{1}$$

Antecedent Accuracy: The rate of correctly identifying the antecedent to which a pronoun refers.

$$Ant.\ Acc. = \frac{Total\ number\ of\ antecedent}{Number\ of\ correctly\ identified\ antecedent} \tag{2}$$

Exact-Match (Link) Accuracy: The rate at which the entire pronoun + antecedent pair (the binary "link") is correctly predicted.

$$Exact\ Match = \frac{Total\ number\ of (pronoun,\ antecendent)\ pairs}{Number\ of\ correctly\ identified\ (pronoun,\ antecendent)\ pairs} \tag{3}$$

F1-Score: A measure that balances precision and recall in a model's "link" predictions into a single value.

$$F1\text{-}Score = \frac{Precision + Recall}{2 \times Precision + Recall} \tag{4}$$

Precision: The proportion of links that the model identified as "correct" that were actually correct.

Recall: The proportion of actual links in the gold standard that were captured by the model.

Pronominal anaphora resolution consists of three layers: (1) correct identification of the pronoun, (2) establishing the correct antecedent-pronoun link, and (3) consistently clustering all coreference chains in the document. Pronoun Accuracy, Antecedent Accuracy, and Exact-Match (Link) Accuracy make the first two layers—that is, individual errors—directly visible. In addition, the F1-Score summarizes the model's overall success in pronoun-antecedent pairings by balancing precision and recall at the link level, thus enabling detailed error analysis and holistic performance evaluation simultaneously. These metrics are widely used and accepted benchmarks in the field of coreference resolution, particularly in standardized competitions such as the CoNLL shared tasks [29].

4. Results and Discussion

In this section, the performance of the application results from our study is analyzed. For the BERT and RoBERTa language models, data in BIO format was used for training, while for the large language models, the fine-tuning process was conducted with data in .json format. The predictions of the models on the data were compared with the data in the original (gold) dataset, and analyses were conducted using the specified metrics. The results obtained from the implemented applications are presented in Table 2.

Table 2. Application Results.

Models	Pronoun Acc.	Antecedent Acc.	Exact Match	F1-Score
Turkcell-LLM-7b-v1	0.88	0.69	0.64	0.78
ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1	0.85	0.65	0.59	0.74
Microsoft Phi-3 Mini-4K-Instruct	0.84	0.48	0.43	0.60
BERT	0.96	0.91	0.86	0.89
RoBERTa	0.95	0.70	0.58	0.72
OpenAI GPT-4o-mini	0.96	0.92	**0.90**	0.94
Hybrid Ensemble (BERT + DeBERTa + Rules)	**0.98**	**0.97**	0.50	**0.96**

Note: Bold values indicate the highest performance in each category.

Based on Table 2, the following evaluations were made.

Pronoun Accuracy measures the ability of each model to correctly recognize pronouns. The Hybrid Ensemble approach achieves the highest value with 0.987, surpassing GPT-4o-mini (0.96) and BERT (0.96). This exceptional performance demonstrates that the voting mechanism effectively reduces individual model errors in pronoun detection. The combination of BERT's Turkish language understanding and DeBERTa's attention mechanism, validated through majority voting, provides more reliable pronoun identification than any single model.

Antecedent Accuracy is the success rate in correctly identifying the entity to which a pronoun refers. Here too, the Hybrid Ensemble leads with 0.977, significantly outperforming GPT-4o-mini (0.922) and BERT (0.910). The integration of linguistic rules, particularly the morphological agreement checking and case marking consistency, contributes substantially to this improvement. While the generative, Turkish-specialized Llama derivatives achieve results between 0.69 and 0.65, the Phi-3 Mini model lags significantly at 0.484.

Exact-Match Accuracy imposes the constraint of correctly linking both the pronoun and its corresponding antecedent simultaneously. In this composite metric, GPT-4o-mini achieved the best result with 0.901, while the Hybrid Ensemble scored 0.505. This significant gap exists by design and offers important advantages for practical NLP applications. The

hybrid system's conservative reconstruction strategy, while deliberately sacrificing exact match performance, offers several practical advantages: (1) Error Propagation Prevention: By initializing all labels as 'O' (Outside) and only marking highly confident predictions, the system minimizes cascading errors that could occur when an incorrectly identified pronoun leads to wrong antecedent assignment throughout a document. (2) Precision-Oriented Design: In real-world NLP applications such as information extraction, document summarization, and machine translation, false positives (incorrectly identified anaphoric relationships) can be more detrimental than false negatives (missed relationships). The hybrid system prioritizes precision (0.977 for pronouns, 0.977 for antecedents) over recall in sentence-level reconstruction, making it more suitable for applications where accuracy is critical. (3) Linguistic Constraint Enforcement: The strict one-to-one pronoun-antecedent mapping ensures grammatically valid outputs, eliminating impossible configurations that might pass through less constrained systems. This is particularly important for Turkish, where morphological agreement must be maintained. (4) Practical Application Trade-offs: For downstream tasks such as coreference resolution in machine translation or information retrieval, the exact identification of pronoun and antecedent boundaries (where the hybrid system excels with 0.987 and 0.977 accuracy, respectively) is often more valuable than perfect sentence-level alignment. The system's high F1-score of 0.960 confirms that it successfully balances precision and recall at the link level. The exact match discrepancy also highlights an important methodological consideration: GPT-4o-mini's higher exact match score (0.901) benefits from its extensive multilingual pre-training corpus (likely including Turkish web data), which provides implicit knowledge of common phrase structures and collocation patterns. However, this advantage comes at the cost of slightly lower component-level precision. The hybrid system, by contrast, relies on explicitly encoded linguistic rules and ensemble voting, achieving superior precision in individual pronoun and antecedent identification. This performance profile suggests different use cases for each approach: the hybrid system is optimal for applications requiring high-precision entity extraction and grammatically sound outputs, while GPT-4o-mini may be preferable for scenarios where approximate but complete sentence-level reconstruction is sufficient.

F1-Score reflects the overall success of the model in pronoun-antecedent pairings by balancing precision and recall at the link level. The Hybrid Ensemble achieved the best result with 0.960, demonstrating superior balance compared to GPT-4o-mini's 0.94. This metric confirms that despite the lower exact match score, the hybrid approach provides the most reliable overall performance for Turkish pronominal anaphora resolution.

Hybrid Ensemble Performance Analysis

The most significant finding of this study is the superior performance of the hybrid ensemble approach in three out of four metrics. The system's architecture enables it to leverage complementary strengths:

1. BERT's Contribution: Provides robust Turkish morphological understanding and handles local context effectively;
2. DeBERTa's Contribution: Captures long-range dependencies through its disentangled attention mechanism;
3. Linguistic Rules' Contribution: Ensures grammatical consistency and filters impossible pronoun-antecedent pairs.

The exceptional pronoun accuracy (0.987) and antecedent accuracy (0.977) indicate that the ensemble's voting mechanism effectively combines the strengths of both models, while the linguistic rules eliminate grammatically incorrect predictions. The F1-score of 0.960 confirms that this approach achieves the best overall balance between precision and recall.

Comparative Analysis of Approaches

The results reveal three distinct performance patterns:

1. Hybrid Ensemble: Excels in component accuracy (pronoun/antecedent) and overall F1-score but shows conservative exact match performance due to its strict labeling strategy;
2. GPT-4o-mini: Maintains consistent high performance across all metrics with particularly strong exact match scores, benefiting from its extensive pre-training;
3. Turkish-specific LLMs: Show moderate performance, suggesting that domain-specific training alone is insufficient without architectural optimization or ensemble methods.

The success of the hybrid approach validates the hypothesis that combining multiple models' predictions with linguistic knowledge can outperform even the most advanced individual models. This is particularly relevant for morphologically rich languages like Turkish, where grammatical rules provide strong signals for anaphora resolution.

While direct quantitative comparison with previous Turkish anaphora resolution studies (Tüfekci & Kılıçaslan [4], Sezer & Kılıçaslan [5]) is challenging due to differences in dataset composition, evaluation metrics, and task scope, a methodological comparison provides valuable insights.

Tüfekci & Kılıçaslan [4] developed a rule-based system using syntactic constraints and salience weighting, achieving approximately 72% accuracy for personal pronouns on 50 Turkish texts. Sezer & Kılıçaslan [5] proposed a constraint-based system using morphological and syntactic analysis, reporting approximately 68% overall success rate (82% for unambiguous cases) on Turkish literary texts. Both approaches relied exclusively on handcrafted rules without neural components and focused on discourse-level coreference chains rather than sentence-level token detection.

By integrating Transformer-based models with linguistic rules, our hybrid system achieves 98.7% pronoun accuracy and 97.7% antecedent accuracy, representing a substantial improvement of 15–30 percentage points. This progression from rule-based to hybrid neural-symbolic approaches reflects broader NLP trends, where neural models capture distributional patterns while linguistic constraints ensure morphological validity. Direct numerical comparison is infeasible due to different evaluation protocols (document-level coreference vs. sentence-level BIO tagging), dataset incompatibility, and task scope differences. However, the qualitative improvement validates the efficacy of Transformer-based approaches augmented with linguistic knowledge for Turkish anaphora resolution. To facilitate future comparisons, we commit to publishing our annotated dataset with standardized evaluation scripts, enabling reproducible benchmarks for the Turkish NLP community.

5. Conclusions and Future Work

In this research, the paradigms of Transformer-based language models, large language models, and hybrid ensemble approaches were evaluated with a comprehensive methodology for the automatic detection of the pronominal anaphora phenomenon in Turkish texts. Within the scope of the study, a novel corpus of 2000 sentences was created using data systematically compiled from academic literature on anaphora theory and typology. This corpus was annotated using a BIO (Beginning-Inside-Outside) tagging scheme specifically developed for this study.

The experimental work was conducted along three main axes: in the first stage, domain-specific fine-tuning procedures were applied to pre-trained language models based on the Transformer architecture; in the second stage, large language models were customized on the created dataset; in the third stage, a hybrid ensemble approach combining multiple models with linguistic rules was developed and evaluated.

The most significant contribution of this research is the development and validation of a hybrid ensemble approach that achieves state-of-the-art performance in Turkish pronominal anaphora resolution. By combining Turkish BERT and DeBERTa-v3 architectures with linguistic rule integration, the system achieved a pronoun accuracy of 0.987, an antecedent accuracy of 0.977, and an F1-score of 0.960, surpassing all individual models, including advanced LLMs like GPT-4o-mini.

The hybrid approach's success demonstrates that for morphologically complex languages like Turkish, the integration of linguistic knowledge with neural models remains crucial. While the exact match score of 0.505 indicates room for improvement in complete sentence-level accuracy, the system's superior performance in component detection makes it highly suitable for practical NLP applications where high precision in pronoun and antecedent identification is paramount.

The findings reveal several important insights: (1) ensemble methods can effectively combine complementary strengths of different architectures, (2) linguistic rules remain valuable even in the era of large language models, (3) trade-offs between component accuracy and sentence-level consistency require careful consideration based on application requirements, and (4) Turkish-specific models benefit significantly from hybrid approaches that incorporate grammatical knowledge.

Future work should focus on improving the exact match performance through more sophisticated alignment strategies while maintaining the high accuracy achieved in pronoun and antecedent detection. Additionally, extending this hybrid approach to zero anaphora and event anaphora resolution presents promising research directions. The integration of attention visualization techniques could provide insights into the decision-making process of the ensemble, potentially leading to further refinements.

Despite these promising results, several limitations should be acknowledged. The dataset size is limited to 2000 sentences, and source diversity is restricted to linguistic papers and children's stories. While these sources provide rich examples of complex anaphoric relationships in formal written Turkish, the model's performance on other genres such as spoken language, news articles, and social media content remains to be evaluated. Future work should expand the dataset to include diverse text types and explore more sophisticated alignment strategies to improve exact match performance while maintaining high component-level accuracy.

The present study is pioneering in the field of Turkish pronominal anaphora detection in terms of the hybrid methodology used and the comprehensive evaluation conducted. The research findings have proven that pronoun-antecedent relationships can be detected with exceptional accuracy rates using ensemble methods combined with linguistic rules, offering a novel contribution to the Turkish natural language processing literature. This study is expected to serve as a methodological reference point for future research and to guide studies in the field of computational modeling of Turkish linguistic phenomena.

Author Contributions: Conceptualization, E.D. and M.B.; methodology, E.D.; software, E.D.; validation, E.D. and M.B.; formal analysis, E.D.; investigation, E.D.; resources, E.D. and M.B.; data curation, E.D.; writing—original draft preparation, E.D.; writing—review and editing, E.D. and M.B.; visualization, E.D.; supervision, M.B.; project administration, M.B.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Bursa Uludağ University Science and Technology Center (BAP) under Grant FGA-2025-2304.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset created for this study, comprising 2000 Turkish sentences with 72,239 tokens and BIO-tagged annotations for pronominal anaphora detection, is available upon request from the corresponding author. The data are not publicly available due to ongoing research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
BIO	Beginning-Inside-Outside
SVM	Support Vector Machine
k-NN	k-Nearest Neighbor
Bi-LSTM	Bidirectional Long Short-Term Memory
GPT	Generative Pre-trained Transformer

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
2. Yu, J.; Jiang, Z.; Li, Q.; Iyyer, M. Coreference Resolution in the Era of Large Language Models. *arXiv* **2023**, arXiv:2310.01699.
3. Mitkov, R. *Anaphora Resolution*; Longman: London, UK, 2002.
4. Tüfekci, P.; Kılıçaslan, Y. Türkçe'de zamirsel anafor çözümlemesi için Hobbs algoritmasının uyarlanması. *Bilgi. Bilim. Ve Mühendisliği Derg.* **2007**, *1*, 45–58.
5. Sezer, T.; Kılıçaslan, Y. A rule-based pronominal anaphora resolution system for Turkish. *Turk. J. Electr. Eng. Comput. Sci.* **2014**, *22*, 188–203.
6. Yıldırım, S. Türkçe Artgönderim Çözümlemesi için Makine Öğrenimi Yaklaşımları. Ph.D. Thesis, Orta Doğu Teknik Üniversitesi, Ankara, Turkey, 2008.
7. Taze, M. Türkçe Zamir Çözümlemesinde Derin Öğrenme Ağlarının Başarım Analizi. Master's Thesis, İstanbul Teknik Üniversitesi, İstanbul, Turkey, 2017.
8. Lee, K.; He, L.; Lewis, M.; Zettlemoyer, L. End-to-end Neural Coreference Resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 188–197.
9. Abacı, H.; Eminağaoğlu, M.; Gör, İ.; Kılıçaslan, Y. Türkçe tweetlerde anaforik ilişkilerin analizi: Çocuk hikayeleri ile karşılaştırmalı bir çalışma. *Bilgi. Bilim. Ve Mühendisliği Derg.* **2022**, *15*, 234–248.
10. Ertan, M. Türkçe ve İngilizce Metinlerde Adılsal Öngönderim Çözümlemesi: TED MDB Derlemi Üzerine bir Çalışma. Master's Thesis, Marmara Üniversitesi, İstanbul, Turkey, 2023.
11. Palomar, M.; Martínez-Barco, P. Computational Approach to Anaphora Resolution in Spanish Dialogues. *J. Artif. Intell. Res.* **2001**, *15*, 263–287. [CrossRef]
12. Peral, J.; Ferrández, A. Translation of Pronominal Anaphora between English and Spanish: Discrepancies and Evaluation. *J. Artif. Intell. Res.* **2003**, *18*, 117–147. [CrossRef]
13. de Marneffe, M.-C.; Recasens, M.; Potts, C. Modeling the Lifespan of Discourse Entities with Application to Coreference Resolution. *J. Artif. Intell. Res.* **2015**, *52*, 445–475. [CrossRef]
14. Paparounas, L.; Akkuş, F. Complex pronouns and anaphoric relations in Turkish: A minimalist analysis. *Linguist. Inquiry* **2023**, *54*, 341–378.
15. Bouzid, S.M.; Zribi, C.B.O. A Generic Approach for Pronominal Anaphora and Zero Anaphora Resolution in Arabic Language. *Procedia Comput. Sci.* **2020**, *176*, 642–652. [CrossRef]
16. Murayshid, S.; Al-Khalifa, H.; Al-Salman, A. Arabic pronoun resolution using AraBERT and Bi-LSTM: A sequence-to-sequence deep learning approach. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 1123–1135.
17. Kongwan, P.; Theeramunkong, T.; Haruechaiyasak, C. Anaphora resolution in Thai texts using Elementary Discourse Unit segmentation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 15.

18. Chen, L. Evaluating ChatGPT-4's performance on Chinese pronoun anaphora resolution: A Winograd Schema Challenge approach. *Comput. Linguist.* **2023**, *49*, 892–915.
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
20. Schweter, S. *dbmdz BERT Models. Zenodo*, version v1; Zenodo: Genève, Switzerland, 2020. [CrossRef]
21. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
22. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, 3–7 May 2021.
23. OpenAI. GPT-4o System Card (v1). OpenAI Technical Report. Available online: https://cdn.openai.com/gpt-4o-system-card.pdf (accessed on 24 October 2025).
24. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [CrossRef]
25. Turkcell AI Team. Turkcell-LLM-7b-v1 [Large Language Model]. Hugging Face. Available online: https://huggingface.co/TURKCELL/Turkcell-LLM-7b-v1 (accessed on 24 October 2025).
26. AI at Meta. The Llama 3 Herd of Models. Meta AI. Available online: https://ai.meta.com/blog/meta-llama-3/ (accessed on 24 October 2025).
27. Kesgin, H.T.; Yüce, M.K.; Doğan, E.; Uzun, M.E.; Uz, A.; İnce, E.; Erdem, Y.; Shbib, O.; Zeer, A.; Amaşyali, M.F. Optimizing Large Language Models for Turkish: New Methodologies in Corpus Selection and Training (Turkish-Llama-8b-DPO-v0.1). In Proceedings of the 2024 Innovations in Intelligent Systems and Applications Conference (ASYU), Ankara, Turkey, 16–18 October 2024.
28. Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219. [CrossRef]
29. Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; Zhang, Y. Scoring Coreference Partitions of Predicted Mentions: A New Scorer for the CoNLL-2011/2012 Shared Tasks. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 90–96.