Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Citation Intent Classification for Turkish Academic Literature with Prompt-Optimized LLMs and Language-Specific Encoders

Kemal Sami Karaca

kemalsami.karaca@tubitak.gov.tr

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu

Bahaeddin Eravci

TOBB University of Economics and Technology

Research Article

Keywords: Citation intent classification, Turkish NLP, Large Language Models, Prompt Optimization, Bibliometric analysis

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10740883/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Financial interests: The authors declare they have no financial interests. No specific grant or external funding was received for the work reported in this manuscript. Non-financial interests: Kemal Sami Karaca is employed by TÜBİTAK ULAKBİM, the institution that operates the TRDizin and DergiPark platforms discussed in this study as potential deployment environments, and that hosts both the annotation platform (atif.ulakbim.gov.tr) and the open data repository (Aperta) used in this work. Computational resources were provided by TÜBİTAK ULAKBİM (TRUBA) and TOBB University of Economics and Technology as part of the authors' institutional duties. Bahaeddin Eravci declares no competing interests. Neither author holds an editorial, advisory, or other role with this journal.

Citation Intent Classification for Turkish Academic Literature with Prompt-Optimized LLMs and Language-Specific Encoders

Kemal Sami Karaca[1*] and Bahaeddin Eravci[2]

[1*]TÜBİTAK ULAKBİM, Yüzüncüyıl, İşçi Blokları Mahallesi Muhsin Yazıcıoğlu Caddesi No:51/C, Ankara, 06530, Türkiye.
[2]Department of Artificial Intelligence Engineering, TOBB ETU, Söğütözü Cd. No:43, Ankara, 06510, Türkiye.

*Corresponding author(s). E-mail(s): kemalsami.karaca@tubitak.gov.tr;
Contributing authors: beravci@etu.edu.tr;

Abstract

Citation intent classification (CIC) can enrich bibliometric analysis by distinguishing how publications are cited, but existing resources remain concentrated in English, limiting the ability of non-English indexing platforms to move beyond raw citation counts. This paper introduces TurkCite, a CIC framework and expert-annotated dataset for Turkish academic literature. A translation-based transfer baseline shows that direct transfer from English resources is insufficient, motivating a language-specific corpus of 2,746 citation instances annotated under a five-class citation-intent schema. TurkCite evaluates two complementary modeling branches. The generative branch applies in-context learning with multiple large language models and prompt-refinement strategies, providing a training-free baseline for cold-start settings. Its best configuration reaches 86.9% accuracy, with a best Macro-F1 of 71.7% in the few-shot setting. The discriminative branch fine-tunes five Transformer encoder backbones under a common supervised protocol, combining structural and contextual variants designed for class-imbalanced Turkish CIC. Turkish-specific encoders form the strongest group, reaching a best mean accuracy of 90.11% and a best mean Macro-F1 of **72.58%**. Richer citation context improves overall accuracy, whereas hierarchical decomposition provides no reliable class-balanced advantage, leaving rare intent categories as the main remaining challenge. Beyond the Turkish case, TurkCite offers a methodological workflow for adapting CIC to non-English bibliometric settings, with complementary pathways for cold-start analysis and scalable citation indexing.

Keywords: Citation intent classification, Turkish NLP, Large Language Models, Prompt Optimization, Bibliometric analysis

1 Introduction

Citations are the fundamental mechanism connecting new research to established literature, serving as the primary indicator of knowledge transfer. Consequently, academic impact has traditionally been quantified through citation counts, which drive standard metrics such as the h-index and journal impact factors. However, relying solely on citation counts treats every reference as an equivalent unit of impact, disregarding whether it signifies a methodological dependency, a contrasting result, or a simple background mention. This lack of granularity obscures the communicative function of citations, limiting the ability to measure the true nature of scientific influence.

To address this limitation, Citation Intent Classification (CIC) has emerged as a key task in bibliometrics, using NLP and ML techniques to analyze the communicative function of references. By categorizing citations into functional classes (such as methodological use, contrast, or general context), CIC enables more nuanced bibliometric analysis and information retrieval. While important progress has been made for English, driven by relatively large-scale datasets like SciCite (Cohan, Ammar, van Zuylen, & Cady, 2019) and ACL-ARC (Jurgens, Kumar, Hoover, McFarland, & Jurafsky, 2018), no comparable framework, model, or annotated benchmark has been established for most non-English languages, despite isolated smaller-scale efforts in a few. Turkish-language scholarly publishing represents a large and growing academic ecosystem supported by dedicated national indexing infrastructure such as TRDizin (TÜBİTAK ULAKBİM, 2024b), yet no native CIC system exists to serve it. Turkish differs substantially from English in linguistic structure and discourse conventions, rendering cross-lingual transfer and translation-based approaches largely ineffective. Furthermore, the wide diversity of disciplines, topics, and writing styles across academic literature compounds this difficulty. A viable CIC system must therefore be designed as a generalizable framework, adaptable through domain-specific annotated data to accommodate heterogeneous discourse norms, rather than being confined to a single disciplinary scope.

This study addresses this gap by introducing a dedicated dual-branch framework for citation intent classification in non-English, low-resource academic settings, instantiated and validated for Turkish. The **TurkCite** framework comprises two complementary classification branches. The *generative branch* employs Large Language Models (LLMs) through In-Context Learning (ICL), exploiting their broad pre-trained knowledge for data-efficient classification without parameter updates. The *discriminative branch* fine-tunes Transformer-based encoders on domain-specific annotated data for deterministic, cost-efficient production deployment. This dual-branch design is a deliberate architectural choice, as the generative branch excels in low-resource scenarios with scarce labeled data, while the discriminative branch provides a reproducible, adaptable, high-throughput solution for large-scale bibliometric indexing. By developing and validating both branches on the first domain-specific Turkish CIC dataset,

we establish a methodology and a set of design decisions that address the challenges posed by the Turkish language, specifically its distinct linguistic structure, increased contextual requirements for disambiguation, and data scarcity, and that are intended to be transferable to other non-English languages facing similar resource constraints.

Beyond a classification benchmark, this study aims to establish a practical decision support system for academic stakeholders. When integrated into national databases such as TRDizin (TÜBİTAK ULAKBİM, 2024b), the proposed system functions as an advanced analytical tool, allowing researchers to filter literature based on specific communicative functions (e.g., identifying papers that *Differ* from a methodology). For editorial management systems like DergiPark (TÜBİTAK ULAKBİM, 2024a), this expert system provides a diagnostic layer to flag potentially superficial citations, thereby supporting academic integrity through automated evaluation. This capability is particularly pressing in an era where the proliferation of AI-generated academic text has introduced new threats to scholarly reliability, including hallucinated references that cite non-existent works. An intent-aware system can surface such anomalies by detecting citations whose communicative profile does not match any plausible function. For instance, an editorial board could deploy the system to audit whether submitted manuscripts contain substantive methodological citations or rely predominantly on generic background references. More broadly, by aggregating citation intents across an author's or journal's publications, the system enables a *citation impact profile* that distinguishes methodological influence from background mentions, offering a more nuanced alternative to raw citation counts and providing a foundation for automated anomaly detection in bibliometric patterns.

Driven by the need to establish a strong baseline for Turkish CIC and to evaluate the dual-branch design, this study addresses two primary research questions. First, we examine the effectiveness of LLMs in a data-scarce setting, specifically questioning whether ICL can serve as a viable alternative to traditional supervised fine-tuning. Within this scope, a critical sub-question is whether automated prompt optimization frameworks can address the consistency problems of manual prompt writing and achieve competitive performance for Turkish academic texts. Second, we aim to develop a trainable, cost-efficient classification system within the discriminative branch that can complement the generative branch in production deployment. To this end, we first construct TurkCite, an expert-annotated Turkish CIC dataset that reflects the naturally skewed distribution of citation intents observed in CIC corpora, with Background citations forming the majority and rarer intent classes remaining underrepresented. Working under this constraint, we evaluate five diverse Transformer architectures to identify the most effective foundation for such a system, examining whether models pre-trained on Turkish corpora outperform multilingual alternatives. We further evaluate a hierarchical classification strategy to address this imbalance directly, investigating whether isolating the dominant "Background" class in a first stage enables more accurate identification of rarer citation types.

Addressing both the linguistic challenges of non-English academic texts and the practical need for automated bibliometric evaluation, the primary contributions of this study are as follows.

- **Turkish Citation Intent Dataset (TurkCite):** We introduce a revised and expanded expert-annotated dataset for Turkish citation analysis, extending the preliminary resource released in Karaca and Eravcı (2025). The current release contains 2,746 citation contexts, rigorously categorized into five intent classes, and provides a scalable foundation comparable in annotation rigor to established English benchmarks such as ACL-ARC. Additionally, an open-access web annotation platform (TÜBİTAK ULAKBİM, 2026) is developed to enable sustainable and continuous corpus expansion, following the longitudinal growth model that has proven effective for English CIC resources.
- **Training-Free Generative CIC Pipeline with Prompt Optimization:** For the generative branch, we develop an LLM-based citation intent classification pipeline using zero-shot and few-shot In-Context Learning. DSPy/MIPROv2 is integrated to automate prompt and demonstration selection, reducing dependence on ad hoc manual prompt engineering. Without parameter updates, the branch reaches 86.9% accuracy with an expert-crafted prompt and 85.8% accuracy under warm-start DSPy optimization.
- **Language-Specific Encoder System for Turkish CIC:** For the discriminative branch, we develop a supervised Transformer-based classification system by fine-tuning five encoder architectures covering Turkish-specific, multilingual, and general-purpose pre-trained models. Beyond model comparison, the system systematically evaluates key design choices for class-imbalanced Turkish CIC, including hierarchical decomposition, contextual enrichment with section and surrounding-sentence information, and metric-aware checkpoint selection across multiple random seeds. Turkish-pretrained encoders emerge as the most effective foundation, with Turkish ELECTRA reaching a peak mean accuracy of $\sim$90% and a peak mean Macro-F1 of $72.58 \pm 0.20\%$. The analysis further shows that contextual enrichment significantly improves accuracy, whereas hierarchical decomposition and Macro-F1-based checkpoint selection do not provide reliable gains in class-balanced performance. The resulting system provides a reproducible and production-oriented baseline for high-throughput citation intent analysis in Turkish and comparable resource-constrained academic settings.

The remainder of this paper is organized as follows. Section 2 surveys the evolution of citation intent classification from rule-based systems to Transformer architectures, reviews automated prompt optimization and situates the current state of Turkish academic NLP within the broader context of non-English bibliometric analysis. Section 3 details the construction, annotation, and linguistic characteristics of the TurkCite dataset, including a cross-lingual feasibility analysis that empirically motivates the need for native classification systems. Section 4 formalizes the proposed dual-branch methodology, covering both the generative and discriminative branches and training procedures. Section 5 presents the experimental results for both branches, ablation studies, and computational cost analyses. Section 6 discusses the key findings, practical deployment implications, and methodological trade-offs. Finally, Section 7 concludes the paper and outlines concrete directions for extending the framework to other non-English academic ecosystems.

2 Related work

The methodological approaches to citation analysis have undergone substantial evolution over the past decades. Early conceptual studies, such as the work by Moravcsik and Murugesan (1975), demonstrated that citations are not mere quantitative units. Rather, each reference serves a specific functional purpose, such as adopting a methodology or contrasting a claim. Alongside these conceptual developments, early computational research by Edmundson (1969) introduced "cue words" as linguistic indicators to automatically identify significant sentences in academic texts. To computationally extract these intents, initial systems relied heavily on rigid, rule-based frameworks. For instance, Garzone and Mercer (2000) employed grammar-based parsing, while Teufel (1999) introduced "Argumentative Zoning" to map citations to the rhetorical structure of a paper.

To overcome the tedious task of manual feature engineering and the lack of cross-domain adaptability in rule-based systems, statistical machine learning approaches such as Support Vector Machines (SVM) and Random Forests were subsequently adopted. While these models improved generalization, they still struggled to capture deep semantic dependencies within complex academic discourse. A major paradigm shift occurred with the emergence of deep learning and Transformer architectures like BERT (Devlin, Chang, Lee, & Toutanova, 2019). To mitigate the limitations of general-domain language models on specialized scientific jargon, Beltagy, Lo, and Cohan (2019) introduced SciBERT, a BERT variant pre-trained on a large corpus of scientific publications, establishing a new standard for computational bibliometrics. Building on this, Cohan et al. (2019) demonstrated that incorporating structural metadata (such as section headers) could significantly enhance intent disambiguation.

Despite these technological advancements, a persistent challenge in the evolution of CIC is the inherent class imbalance prevalent in scholarly datasets. Background citations typically dominate academic corpora, often constituting over 70% of all instances, while semantically rich but rare intents such as methodological use or disagreement remain underrepresented. This skewed distribution biases models toward majority classes, leading to poor recall for rare but significant intents. While various strategies have been explored to address this imbalance, including cost-sensitive loss functions such as Focal Loss (Lin, Goyal, Girshick, He, & Dollár, 2017), class-weighted training, and hierarchical decomposition of the label space, achieving high discriminative power across all rhetorical functions remains an open challenge for automated classification systems. A recent example is CiteFusion (Paolini et al., 2025), which decomposes the multi-class CIC task into one-vs-all binary subtasks and pairs a domain-specific encoder (SciBERT, Beltagy et al., 2019) with a general-purpose one (XLNet, Z. Yang et al., 2019) for each intent, aggregating their outputs through a feed-forward meta-classifier. This design reports Macro-F1 scores of 89.60% on SciCite and 76.24% on ACL-ARC, indicating that decomposing the label space can be effective when each class is modeled independently rather than through a sequential cascade.

As modeling capabilities have expanded beyond individual citation sentences, recent research has explored the structural context provided by citation networks

and knowledge graphs, integrating graph embedding techniques to capture inter-document dependencies that purely text-based models overlook (Kipf & Welling, 2017). GraphCite (Berrebbi, Huynh, & Balalau, 2022), for example, combines high-level citation patterns with textual clues to improve intent prediction, while Tsai, Yen, Huang, and Chen (2023) enrich classification by mapping rhetorical intents between citing and cited papers. While these graph-based approaches offer complementary structural signals, text-level classification remains the foundational capability required for any downstream network analysis. In parallel, decision support systems for academic literature analysis have gained traction within the expert systems community. Van Dinter, Catal, and Tekinerdogan (2021) proposed a DSS employing multi-channel CNN architectures for automated citation classification during document screening. However, such decision support systems have been developed almost exclusively for English-language corpora, leaving a significant gap for non-English academic ecosystems that maintain their own national indexing infrastructure and discourse conventions.

To address this data and infrastructure scarcity in non-English domains, an alternative to supervised fine-tuning has emerged through prompt-based learning, driven by the recognition that while fine-tuning yields high discriminative accuracy, it remains bottlenecked by the requirement for extensive annotated datasets. A fundamental shift in methodology occurred with the introduction of Large Language Models (LLMs) such as GPT-3 (Brown et al., 2020). These large-scale architectures introduced In-Context Learning (ICL), enabling zero-shot or few-shot classification without parameter updates and offering a flexible alternative for low-resource languages where annotated data is scarce. Whether ICL can substitute for supervised fine-tuning in citation intent classification, however, remains unsettled: Istrate et al. (2024) found that neither zero-shot nor few-shot prompting with GPT-3.5 and GPT-4 approached fine-tuned performance on software citation intent, concluding that the task requires additional training rather than prompting alone. Koloveas, Chatzopoulos, Vergoulis, and Tryfonopoulos (2026) report the same pattern for open-weight models, where in-context learning remained below specialized encoder baselines until supervised fine-tuning was applied. Their parameter analysis further indicates that the prompt configuration, in particular the number of demonstrations and the query template, was significantly associated with performance, whereas decoding temperature was not. A further major challenge with generative LLMs is their sensitivity to prompt phrasing, and automated optimization frameworks have emerged to address this instability. Systems such as Automatic Prompt Engineer (APE) (Zhou et al., 2023) and Optimization by PROmpting (OPRO) (C. Yang, He, Li, Cao, & Tang, 2023) treat prompts as learnable parameters, while PromptBreeder (Fernando, Banarse, Michalewski, Osindero, & Rocktäschel, 2024) employs evolutionary algorithms to refine task-specific instructions. Among these frameworks, DSPy (Khattab et al., 2024) introduces a significant advancement by transitioning from heuristic "string-based" engineering to a declarative, programmatic optimization pipeline. By decoupling the inferential program flow from task-specific instructions, DSPy enables the systematic optimization of LLM interactions.

While prompt-based approaches reduce the dependency on labeled data, fine-tuning pre-trained encoders on task-specific corpora remains the dominant paradigm when deterministic, reproducible outputs are required. The effectiveness of this approach depends heavily on the availability of high-quality language-specific models. In this context, Turkish NLP has undergone a significant transformation over the past decade, moving beyond foundational morphological analysis (Oflazer & Saraçlar, 2018) to complex downstream tasks. Substantial progress has been achieved in domains such as Named Entity Recognition (NER) (Yeniterzi, 2011), Sentiment Analysis (Dehkharghani, Saygin, Yanikoglu, & Oflazer, 2016), and text categorization (Amasyali & Diri, 2006). To address the complex morphological dependencies of the Turkish language, recent advances have shifted focus toward native pre-trained models. The release of monolingual Transformer models, such as BERTurk and a Turkish ELECTRA variant trained on the same corpora (Schweter, 2020), has provided the necessary semantic foundation to process Turkish text effectively. The latter adapts the replaced-token-detection objective of ELECTRA (Clark, Luong, Le, & Manning, 2020) to Turkish. Additionally, multilingual models like XLM-RoBERTa (Conneau et al., 2020) have demonstrated strong cross-lingual transfer capabilities. However, despite these advances in general-purpose NLP, specialized academic tasks such as Citation Intent Classification remain largely unexplored.

Consequently, despite these broad NLP advancements, the automated analysis of citations in Turkish is still at an early stage. Previous exploratory research in this area mostly relied on manual labor. For instance, Atılgan, Atakan, and Bulut (2008) manually categorized citations to understand their functions. Early computational attempts, such as those by Taşkın and Al (2018), employed basic machine learning algorithms that lacked deep contextual understanding. Our preliminary work (Karaca & Eravcı, 2025) introduced an initial Turkish CIC resource and examined LLM-based citation intent classification, but was confined to a generative-only feasibility setting. More broadly, CIC research outside English remains limited and the few existing efforts have followed markedly different methodological routes. Budi and Yaniasih (2023) constructed an end-to-end pipeline for Indonesian journal articles drawn from the national indexing platform SINTA, extracting citation sentences with GROBID and manually annotating 8,566 instances across five disciplines under a five-class citation-function schema. Their multi-output convolutional network classifies citation sentiment, role, and function jointly, reaching a Macro-F1 of 0.88 for citation function. This work demonstrates that language-specific CIC resources can be developed on top of national indexing infrastructure, although it relies on word-embedding and CNN architectures rather than pre-trained language models. Following a different route Deng, Tong, Zhang, and Li (2025) developed an LLM-based citation intent classification scheme for Chinese patent literature, explicitly adopting LLMs because they significantly reduce dependence on large-scale annotated data sets and are particularly suitable for applications in sparse-data contexts. Taken together, these efforts indicate that non-English CIC has so far been approached either through supervised shallow architectures or through prompting alone. To our knowledge, no existing framework combines language-specific Transformer encoders with a training-free generative pathway within a single non-English academic ecosystem, and none addresses

the constraints that agglutinative morphology imposes on subword representation for this task.

3 TURKCITE dataset

3.1 Data sources and selection

To construct a representative corpus for Turkish citation analysis, we sourced research articles from DergiPark (TÜBİTAK ULAKBİM, 2024a) and TRDizin (TÜBİTAK ULAKBİM, 2024b), the two primary indexing platforms for academic publishing in Türkiye. The scope was restricted to the domain of Computer Science for the period 2020–2025. Computer Science literature typically exhibits explicit citation patterns by referencing specific methods, algorithms, and datasets, minimizing the ambiguity often found in humanities or social sciences. Furthermore, this choice aligns the corpus with other benchmarks such as SciCite and ACL-ARC, facilitating cross-lingual comparisons. While the current instantiation targets Computer Science, the annotation platform and data collection framework are designed for extension to other disciplines as domain-specific labeled data becomes available. We exclusively selected original research articles, excluding review papers to prevent the overrepresentation of "Background" citations and to ensure a balanced inclusion of methodological references.

3.2 Citation extraction pipeline

The extraction of citation contexts from PDF documents was automated using the CEX (Citation Extractor) module, developed under the GRASPOS project (GRASPOS, 2024). Using the GROBID (GeneRation Of BIbliographic Data) framework (Lopez, 2009), this pipeline parses unstructured PDF content into structured TEI format. We developed a custom wrapper to process the output into JSON, ensuring compatibility with our subsequent machine learning pipelines. The comprehensive workflow of this citation extraction and string normalization process is illustrated in Figure 1. To handle Turkish-specific tokenization and character encoding issues, we used pre-trained GROBID models optimized for scientific literature. The pipeline was applied to 723 candidate documents and produced 20,026 raw citation strings. Documents that were not original research articles and documents from which no citation context could be reliably recovered were discarded, together with malformed extractions. Contexts that were identical after normalization, including records the parser had emitted separately although they differed only in punctuation or whitespace, were then de-duplicated. This yielded 18,837 unique citation contexts from 712 distinct articles, which formed the pool from which the annotated subset was drawn. Each entry contains the citation sentence, the referenced work, and the structural section of the paper (e.g., Introduction, Method) in which it appears. This pipeline is designed to be reproducible for other non-English academic ecosystems with comparable national indexing infrastructure.

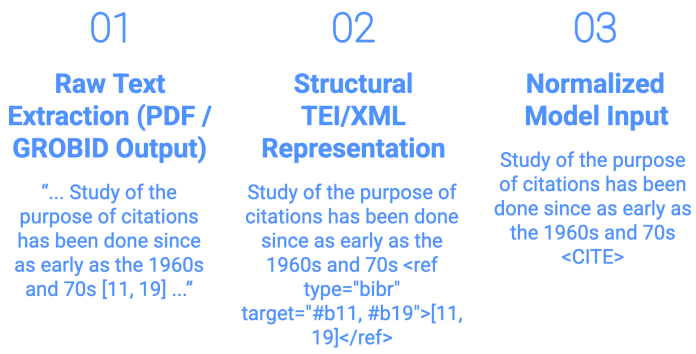


Fig. 1: Visual representation of the citation extraction and normalization pipeline. The process transitions from a raw PDF sentence to a structured TEI/XML format, concluding with a normalized string where the target citation is labeled as `<CITE>` and the secondary reference is masked as `[REF]`.

3.3 Annotation schema

Defining an appropriate taxonomy for citation intents is critical for downstream model performance. While unsupervised clustering analysis (using the Elbow method (Jin & Han, 2011) and Silhouette scores (Rousseeuw, 1987)) on our raw data suggested a latent structure beyond a simple binary classification, we sought a schema grounded in established bibliometric practice. Consequently, we adopted the 5-class Web of Science (WoS) (Clarivate, 2022) schema.

- **Background:** Citations providing context or historical overview.
- **Basis:** Citations indicating the use of a method, dataset, or algorithm.
- **Support:** Citations whose findings reinforce the current study.
- **Differ:** Citations contrasting with the current study's results.
- **Discuss:** Citations that are critically analyzed or elaborated upon.

This granular schema was selected over simpler, ternary taxonomies (e.g., SciCite's 3-class system) to capture the semantic depth of academic discourse, particularly to isolate the key distinction between using a method (Basis) and agreeing with a prior result (Support).

3.4 Annotation process and quality assurance

To ensure high-quality labels and reduce the cognitive burden on annotators, we developed a custom web-based annotation interface (illustrated in Figure 2) that consolidates the target citation, surrounding context, section metadata, labeling options, and original PDF view within a single workspace. The platform is publicly accessible at `atif.ulakbim.gov.tr` (TÜBİTAK ULAKBİM, 2026), supports multi-annotator workflows with configurable review stages, and is designed for longitudinal corpus expansion as new disciplines and annotators are onboarded.

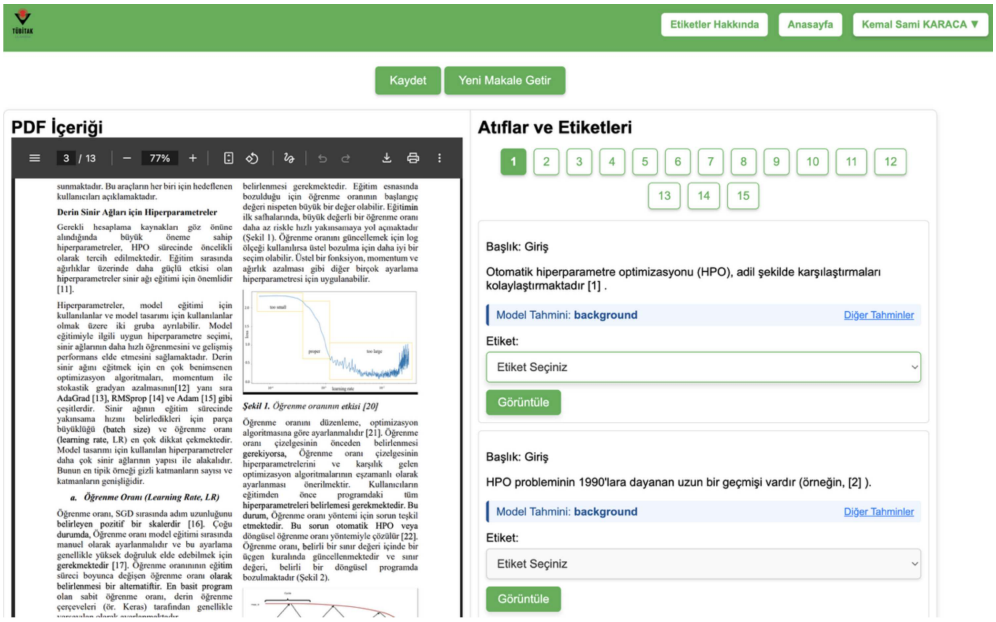


Fig. 2: The web-based annotation interface developed for the TurkCite dataset, displaying the original PDF context, section metadata, and active labeling options for the five-class schema.

Given the interpretive nature of citation intent classification, annotation was conducted through an expert-centered workflow. Domain experts assigned labels using the five-class schema described in Section 3.3, consulting the expanded document context when needed to resolve ambiguous cases.

To validate annotation reliability, we calculated Cohen's Kappa (κ) (Cohen, 1960) on a blinded subset of the data. The inter-annotator agreement between the two domain experts yielded a Cohen's Kappa score of $\kappa = 0.837$, which is categorized as "*Almost Perfect*" agreement according to the Landis and Koch (Landis & Koch, 1977) scale. This value indicates a high level of consistency while also reflecting the inherent interpretive difficulty of citation intent classification. Agreement was lower for the "Discuss" category than for more deterministic classes such as "Background," suggesting that the main source of disagreement lies in distinguishing substantive intellectual engagement from general contextual mention.

3.5 TurkCite corpus characteristics

The final **TurkCite** dataset consists of 2,746 expert-annotated citation sentences. Consistent with global trends observed in datasets like ACL-ARC, our corpus exhibits a significant class imbalance. The Background class dominates, constituting 78% of the data, while classes such as Differ and Support are comparatively underrepresented. Table 1 reports the class distribution. This imbalance presents a primary challenge for

classification models, motivating the use of specialized loss functions and hierarchical architectures in CIC systems.

An earlier and smaller version of this corpus was released alongside our preliminary study (Karaca & Eravcı, 2025), which examined LLM-based classification in a generative-only feasibility setting. Relative to that release, duplicate and near-duplicate contexts introduced by the extraction pipeline were removed, the retained contexts were normalized as described in Section 3.6, and the removed instances were replaced with newly annotated non-Background contexts. These revisions slightly reduce the severity of the class imbalance without altering its overall shape, yielding 2,746 instances compared with 2,650 in the preliminary release.

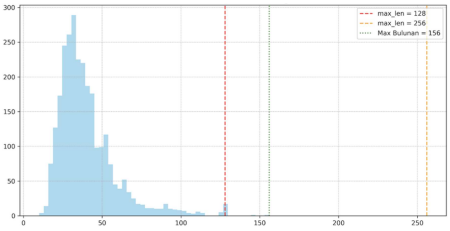
Table 1: Class distribution of the dataset.

Intent Class	**Count**
Background	2139
Basis	313
Discuss	134
Support	106
Differ	54
Total	**2746**

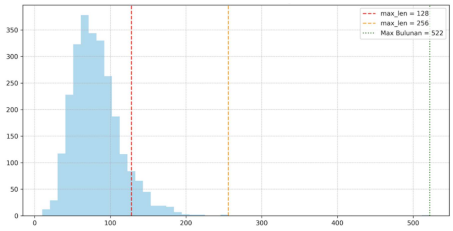
The current version of the dataset (Version 1) is released as an open-source resource, freely available to researchers and academic stakeholders at `aperta.ulakbim.gov.tr` (Karaca & Eravcı, 2026). To situate TurkCite within the broader CIC landscape, Table 2 compares its key characteristics with established English benchmarks.

Beyond the class distribution, we conducted an empirical analysis of the lexical properties of the corpus. An Out-of-Vocabulary (OOV) analysis revealed that 28.65% of unique tokens in the Turkish academic text are OOV when using traditional static word-level embeddings (e.g., GloVe). This high rate is a direct consequence of the highly agglutinative and productive nature of Turkish morphology, especially within specialized scientific domains. Such a significant lexical gap demonstrates the inadequacy of word-level representations for Turkish and provides a strong justification for the adoption of subword-based Transformer architectures, such as BERTurk or generative LLMs, which handle these morphological complexities more effectively.

Furthermore, we examined the distribution of sequence lengths to determine optimal input constraints for our models. As illustrated in Figure 3, single-citation sentences are generally concise, with nearly all instances fitting within a 128-token window. However, the introduction of richer context (incorporating the paper title and surrounding sentences) shifts the distribution, with some samples extending up to 522 tokens. While the maximum length reaches 522, the distribution analysis confirms that a 256-token limit (and 512 for specific context-heavy configurations) captures the vast



(a) Single-sentence token length distribution (majority within 128 tokens).



(b) Rich-context token length distribution (extending up to 522 tokens).

Fig. 3: Token length distributions for different input representations in TurkCite.

majority of the contextual information. This empirical evidence provides the methodological justification for our sequence length parameters, balancing computational efficiency with the preservation of critical contextual cues.

Table 2: Comparison of TurkCite with established English CIC benchmarks.

Property	**ACL-ARC**	**SciCite**	**TurkCite**
Language	English	English	Turkish
Samples	∼3,000	∼11,000	2,746
Classes	6	3	5
Domain	Comp. Ling.	Multi-domain	Comp. Science
Annotation	Expert	Expert+Crowd	Expert

3.6 Preprocessing and normalization

Prior to model training, the extracted citation contexts were cleaned and normalized through a standardized preprocessing pipeline. Structural artifacts introduced during PDF parsing, such as truncated sentences and encoding inconsistencies, were corrected where possible. To focus the models on semantic context rather than surface-level bibliographic formats, which smaller encoder models can otherwise exploit as a spurious cue, the target citation marker was replaced with a special `<CITE>` token. Other citation markers appearing within the same context were masked with `[REF]` tokens, allowing the models to distinguish the target citation from secondary references.

3.7 Cross-lingual transfer baseline

Before developing native Turkish classifiers, we assessed whether an English-centric transfer pipeline could provide a viable shortcut for Turkish citation intent classification. Turkish citation contexts were translated into English using the Gemini Flash (`[gemini-2.5-flash]`) (Gemini Team, Google, 2024) and then classified with

SciBERT (Beltagy et al., 2019), a widely used encoder for scientific text. This translation-based baseline reached a peak accuracy of 0.73 but only approximately 0.50 Macro-F1, indicating limited class-balanced performance.

The performance pattern suggests that translation alone does not preserve all of the discourse cues needed for citation intent classification. Two factors are likely involved: semantic and stylistic shifts introduced when complex Turkish academic prose is rendered into English, and a distributional mismatch between native English scientific writing and machine-translated Turkish academic text. These findings motivate the development of native Turkish CIC resources and classifiers rather than relying solely on English models through translation. More broadly, the analysis indicates that translation-based CIC may be insufficient for languages whose linguistic structure and academic discourse conventions differ substantially from English.

4 Methodology

4.1 TurkCite framework overview

The TurkCite framework is organized around two complementary classification branches: a generative branch based on Large Language Models (LLMs) and a discriminative branch based on fine-tuned Transformer encoders. As illustrated in Figure 4, the workflow begins with raw PDF ingestion and citation-context extraction, followed by preprocessing, annotation, and normalization into the TurkCite dataset. The resulting citation contexts are then processed through the two branches, which are developed and evaluated as complementary modeling strategies for Turkish citation intent classification.

The generative branch uses In-Context Learning to classify citation intents without parameter updates. Within this branch, DSPy/MIPROv2 is used to systematize prompt and demonstration selection, reducing dependence on ad hoc prompt design. The discriminative branch fine-tunes five pre-trained encoder architectures under different structural and contextual configurations to identify effective modeling choices for class-imbalanced Turkish academic text.

This dual-branch framework reflects two practical usage scenarios. The generative branch is suitable for low-resource or cold-start settings where labeled data are limited and rapid adaptation is needed. The discriminative branch, by contrast, supports deterministic, low-cost, and high-throughput classification for large-scale bibliometric indexing. In this sense, TurkCite is not a single fixed model architecture, but a methodological framework that compares complementary classification strategies for Turkish academic literature analysis.

4.2 Generative branch: prompt-optimized in-context learning

The generative branch provides a training-free classification pathway for adapting citation intent classification to languages and domains where annotated data are limited. In this study, we instantiate this pathway for Turkish academic literature using pre-trained instruction-following multilingual LLMs. We formulate the task as In-Context Learning (ICL), where classification is guided by natural language task instructions

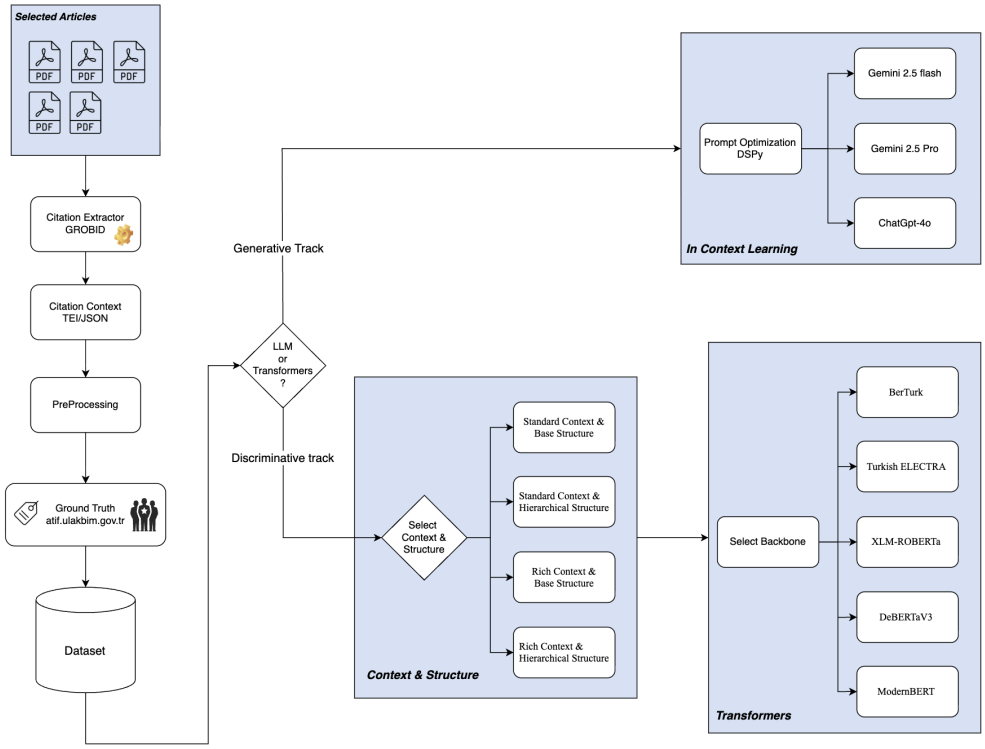


Fig. 4: End-to-end workflow of the **TurkCite** framework. The pipeline traces the progression from selected PDF articles through citation extraction, citation-context construction, preprocessing, annotation, and dataset formation, followed by two complementary classification branches: a generative LLM branch and a discriminative Transformer encoder branch.

and, in the few-shot setting, a small number of labeled exemplars supplied directly in the prompt. Unlike fine-tuned encoders, ICL does not require task-specific parameter updates. We therefore evaluate both zero-shot prompts, which contain only task instructions, and few-shot prompts, which combine task instructions with k labeled examples.

We represent a prompt as $\mathcal{P} = (\mathcal{I}, \mathcal{D}_k)$, where $\mathcal{I}$ denotes the instruction component, including the task description and citation-intent definitions, and $\mathcal{D}_k = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ denotes a set of k labeled exemplar pairs drawn from the training data. In the zero-shot setting, $\mathcal{D}_0 = \emptyset$, and classification relies only on $\mathcal{I}$. Given a citation context x, the LLM φ_{LLM} predicts the citation intent label as:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_{\varphi_{\text{LLM}}}(y \mid \mathcal{P}, x) \tag{1}$$

where $\mathcal{Y} = \{\text{Background, Basis, Support, Differ, Discuss}\}$ is the set of citation intent labels defined in Section 3. This formulation provides a conceptual representation of

label selection. In implementation, model outputs are constrained to the predefined label set and parsed accordingly.

A practical challenge in ICL is that performance can vary with prompt wording, class definitions, and the choice of labeled exemplars. This sensitivity is especially relevant for non-English academic text, fine-grained intent labels, and class-imbalanced datasets. To make prompt selection more reproducible and to improve the overall configuration of the generative branch, we integrate DSPy (Khattab et al., 2024), a declarative framework for optimizing LLM pipelines. Specifically, we use MIPROv2 to search over candidate instructions and few-shot exemplar sets. In this setup, instructions and exemplars are treated as optimizable prompt components, while the underlying LLM parameters remain fixed. The optimizer evaluates candidate prompt configurations on an internal validation split and selects the configuration that maximizes exact matches. The configuration used in our DSPy pipeline is summarized in Table 3.

Table 3: DSPy Configuration Settings

Configuration Parameter	**Value**
Module Strategy	ChainOfThought
Train / Validation Split	80% / 20%
Instruction Candidates	18
Few-Shot Candidates	9
Max Bootstrapped Demos (k)	6
Optimization Trials	27

A key design choice is the initialization of the prompt search. We adopt a warm-start strategy using a domain-expert seed prompt $\mathcal{P}_{\text{seed}}$ that encodes the WoS class definitions and task-specific citation cues. This allows the search to begin from a task-informed prompt rather than an empty instruction space, while preserving the training-free nature of the generative branch. The full prompt optimization procedure is provided in Appendix A.1.

4.3 Discriminative branch: fine-tuned encoders

The discriminative branch provides a supervised classification pathway based on fine-tuned encoder models. Its purpose is not only to identify the strongest pre-trained backbone for Turkish citation intent classification, but also to evaluate which modeling choices are effective under class imbalance and semantic ambiguity. We selected five encoder architectures, covering Turkish-specific, multilingual, and general-purpose pre-trained models, to compare the effect of language-specific pre-training and encoder design under a common classification protocol. These comprise BERTurk and a Turkish ELECTRA variant (Clark et al., 2020; Schweter, 2020), the multilingual XLM-RoBERTa (Conneau et al., 2020), and two English-pretrained general-purpose encoders, DeBERTaV3 (He, Gao, & Chen, 2023) and ModernBERT (Warner et al., 2024).

Given a citation context X consisting of n tokens, a pre-trained encoder $\varphi_{\text{enc}}^{\text{pre}}$ produces contextualized hidden states $\mathbf{H} = \varphi_{\text{enc}}^{\text{pre}}(X) = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, where $\mathbf{h}_i \in \mathbb{R}^d$. To ensure comparability across the five architectures, we apply the same classification head to each encoder. Instead of relying only on the `[CLS]` token, we compute a mean-pooled context vector, $\mathbf{h}_{\text{pool}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_i$, to obtain a representation of the full citation context.

To mitigate overfitting, dropout is applied with probability $p = 0.1$, yielding $\mathbf{h}_{\text{drop}} = \text{Dropout}(\mathbf{h}_{\text{pool}})$. A linear projection then maps this representation to class logits, $\mathbf{z} = \mathbf{W}\mathbf{h}_{\text{drop}} + \mathbf{b}$, where $\mathbf{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{Y}|}$. The probability distribution over the five citation intent classes is obtained as $\hat{\mathbf{p}} = \text{softmax}(\mathbf{z})$, and the predicted label is selected by maximum probability. The complete encoder classifier can therefore be written as

$$\varphi_{\text{enc}}(X) = \arg\max_{y \in \mathcal{Y}} \left[\text{softmax}\left(\mathbf{W}\, \text{Dropout}\left(\frac{1}{n} \sum_{i=1}^{n} \varphi_{\text{enc}}^{\text{pre}}(X)_i \right) + \mathbf{b} \right) \right]_y . \tag{2}$$

This standardized classifier isolates the contribution of the encoder backbone while keeping the downstream decision layer constant. On top of this base formulation, we introduce two modifications motivated by the properties of citation intent data. The first is a structural modification that addresses the dominance of the *Background* class by decomposing the five-class problem into a two-stage hierarchical pipeline. The second is a contextual modification that addresses semantic ambiguity by enriching the input with section headers and surrounding sentences. This choice follows evidence that structural cues carry semantic signal beyond their organizational role, as Paolini et al. (2025) report mean accuracy gains from prepending section titles to citation contexts. These modifications are independent and can be applied separately or jointly, producing four classifier configurations.

Input representations: We define two levels of input context. The standard input S contains the target citation sentence. The rich-context input S_{rich} augments this sentence with section-level and neighboring-sentence information:

$$S_{\text{rich}} = [\text{CLS}]\, S_{\text{title}}\, [\text{SEP}]\, S_{p-1} \oplus S_p(\langle \text{CITE} \rangle) \oplus S_{p+1}\, [\text{SEP}]. \tag{3}$$

Here, S_{title} denotes the section header, S_p denotes the target citation sentence, and S_{p-1} and S_{p+1} denote the preceding and following sentences.

Model variants: Based on the classification objective, three encoder configurations are trained:

- $\varphi_{\text{enc}}^{(1)}(X)$: a single-stage classifier over all $|\mathcal{Y}| = 5$ citation intent classes.
- $\varphi_{\text{enc}}^{(2)}(X)$: a binary classifier for *Background* versus *Non-Background*.
- $\varphi_{\text{enc}}^{(3)}(X)$: a four-class classifier for *Basis*, *Support*, *Differ*, and *Discuss*.

The hierarchical configurations combine $\varphi_{\text{enc}}^{(2)}$ and $\varphi_{\text{enc}}^{(3)}$ in cascade. The binary classifier first determines whether the citation belongs to the dominant *Background* class. If not, the citation is passed to the four-class classifier:

$$\hat{y} = \begin{cases} c_{\text{bg}}, & \text{if } \varphi^{(2)}_{\text{enc}}(X) = c_{\text{bg}} \\ \varphi^{(3)}_{\text{enc}}(X), & \text{if } \varphi^{(2)}_{\text{enc}}(X) \neq c_{\text{bg}}. \end{cases} \tag{4}$$

Classifier configurations: The combinations of the structural and contextual modifications produce four configurations within the discriminative branch, summarized in Table 4.

Table 4: Discriminative branch classifier configurations. Each configuration is defined by the combination of a structural choice, single-stage or hierarchical, and a contextual choice, standard or rich input.

Classifier Configuration	Classifier Structure	Context	Definition
Base-Standard	Single-stage	S	$\hat{y} = \varphi^{(1)}_{\text{enc}}(S)$
Hierarchical-Standard	Hierarchical	S	$\hat{y}$ via Eq. 4, $X = S$
Base-Rich	Single-stage	S_{rich}	$\hat{y} = \varphi^{(1)}_{\text{enc}}(S_{\text{rich}})$
Hierarchical-Rich	Hierarchical	S_{rich}	$\hat{y}$ via Eq. 4, $X = S_{\text{rich}}$

4.4 Encoder training and model selection

For the discriminative branch, all encoder models were trained under a common optimization protocol to ensure comparability across architectures and classifier configurations. Because the TurkCite label distribution is strongly imbalanced, with *Background* citations constituting approximately 78% of the dataset, standard Cross-Entropy loss can favor the majority class during training. We therefore used Focal Loss (Lin et al., 2017) as the training objective. This loss introduces a confidence-based modulating factor that reduces the contribution of well-classified examples and assigns greater relative weight to harder cases, which is particularly relevant for minority intent classes such as *Differ* and *Basis*.

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{5}$$

where p_t is the model-estimated probability of the ground-truth class, α_t is the class-specific weighting factor, and γ is the focusing parameter. The values of α_t and γ were selected through hyperparameter optimization rather than fixed manually.

In addition to the training loss, we treat the validation metric used for model selection as a separate design choice. All encoder models are optimized with focal loss, since Macro-F1 is not differentiable and therefore cannot be used directly as a gradient-based objective. However, hyperparameter search, early stopping, and checkpoint selection can be guided by either validation Accuracy or validation Macro-F1. Each classifier configuration in Table 4 is therefore tuned under both selection targets. This allows us to assess whether optimizing for overall correctness or class-balanced validation performance leads to better generalization on the minority citation intents.

This separation between optimization loss and selection metric is important in class-imbalanced CIC, where the model with the highest validation accuracy may not be the model with the best minority-class behavior.

All encoder models were optimized with AdamW, a linear learning-rate scheduler, and early stopping with a patience of five epochs. Hyperparameters, including learning rate, batch size, dropout rate, and focal loss parameters, were tuned with 50 trials per configuration. The best checkpoint from each trial was selected according to the corresponding validation target and evaluated only on the held-out test set.

5 Experiments and results

5.1 Experimental setup

TurkCite was divided into fixed training, validation, and test partitions using a stratified hold-out strategy. The final split assigns 68% of the instances to training, 12% to validation, and 20% to testing. Stratification preserves the citation-intent distribution across partitions, ensuring that minority classes such as *Differ* and *Discuss* remain represented during model development and final evaluation. The validation set was used for prompt selection, hyperparameter tuning, early stopping, and checkpoint selection, while the held-out test set was reserved exclusively for final reporting.

The evaluation was conducted across the two branches of the TurkCite framework. The generative branch was implemented through official API endpoints for instruction-following LLMs and evaluated under zero-shot, few-shot, and prompt-optimized configurations. The discriminative branch was implemented using the HuggingFace Transformers library and fine-tuned locally on NVIDIA A100 and Tesla V100 GPUs. For the encoder-based models, hyperparameter optimization was performed with Optuna using 50 trials per configuration, following the model-selection targets defined in Section 4.4. Final results are reported on the held-out test set, with accuracy and Macro-F1 used as the primary evaluation metrics to capture both overall correctness and class-balanced performance.

5.2 Generative branch results

This section evaluates the generative branch as a training-free pathway for citation intent classification. We analyze prompt configurations of the form $\mathcal{P} = (\mathcal{I}, \mathcal{D}_k)$, where $\mathcal{I}$ denotes the instruction component and $\mathcal{D}_k$ denotes the set of k labeled exemplars. The evaluation focuses on three questions: how much instruction design affects performance, whether automated prompt optimization improves over manually specified prompts, and how sensitive the models are to few-shot exemplars. The prompt templates used in this progression are provided in Appendix A.2.

The prompt refinement results in Table 5 show that instruction design is the dominant factor in generative CIC performance. With the basic instruction prompt, which contains only label names and a generic classification directive, Gemini 2.5 Flash reached 0.616 accuracy and GPT-4o reached 0.587 accuracy. Replacing this generic instruction with a detailed prompt containing citation-intent descriptions increased accuracy to 0.833 for Gemini 2.5 Flash and 0.825 for GPT-4o. This large gain across

both models indicates that Turkish citation intent classification cannot be reliably induced from label names alone. Instead, the models require explicit definitions of the communicative function associated with each citation-intent class.

Table 5: Prompt refinement and optimization results for the generative branch. Stages 1–3 show the effect of progressively richer instruction design under zero-shot settings. Stages 4–5 report DSPy/MIPROv2 optimization results for Gemini 2.5 Flash. Reference rows provide additional model-capacity and cross-model comparisons.

Stage	Model	Configuration	Acc
1	Gemini 2.5 Flash	$(\mathcal{I}_{\text{basic}}, \mathcal{D}_0)$	0.616
	GPT-4o	$(\mathcal{I}_{\text{basic}}, \mathcal{D}_0)$	0.587
2	Gemini 2.5 Flash	$(\mathcal{I}_{\text{detailed}}, \mathcal{D}_0)$	0.833
	GPT-4o	$(\mathcal{I}_{\text{detailed}}, \mathcal{D}_0)$	0.825
3	Gemini 2.5 Flash	$(\mathcal{I}_{\text{expert}}, \mathcal{D}_0)$	0.869
	GPT-4o	$(\mathcal{I}_{\text{expert}}, \mathcal{D}_0)$	0.821
4	Gemini 2.5 Flash (DSPy)	$\mathcal{P}^*_{\text{cold}}$	0.740
5	Gemini 2.5 Flash (DSPy)	$\mathcal{P}^*_{\text{warm}}$	0.858
Ref.	Gemini 2.5 Pro	$(\mathcal{I}_{\text{expert}}, \mathcal{D}_0)$	0.861
Ref.	GPT-4o	$(\mathcal{I}_{\text{expert}}, \mathcal{D}_1)$[a]	0.827

[a]For GPT-4o, the strongest configuration was obtained with a single demonstration rather than in the zero-shot setting. Full k-shot results are reported in Table 6.

The expert-crafted instruction further clarifies the role of domain-aware task specification. Under the zero-shot expert prompt, Gemini 2.5 Flash achieved the strongest generative accuracy, reaching 0.869. GPT-4o, by contrast, did not improve beyond its detailed-prompt result under the corresponding zero-shot expert setting, reaching 0.821. Its strongest configuration overall was instead obtained with a single demonstration (0.827), reported as a reference row in Table 5 and examined in the k-shot analysis below. This suggests that additional expert framing is not uniformly beneficial across model families, but that strong instruction design remains necessary for both. The main benefit of the expert prompt is therefore not merely longer prompting, but a more precise encoding of citation-intent semantics and Turkish academic discourse cues.

DSPy/MIPROv2 was then evaluated to determine whether automated prompt search could reduce the need for manual prompt engineering. The cold-start configuration reached 0.740 accuracy, improving over the basic prompt but falling short of both the detailed and expert-crafted instructions. This shows that prompt optimization can identify partially useful configurations, but it does not infer the full task semantics required for fine-grained citation intent classification. The warm-start configuration performed better, reaching 0.858 accuracy when initialized from a domain-informed seed prompt. However, it still did not surpass the best expert-crafted zero-shot result. We therefore interpret DSPy as a conditional refinement tool: it can improve and stabilize a reasonable initial prompt, but its effectiveness depends strongly on the semantic quality of the seed. This makes it useful for reducing prompt iteration effort

when expert input is limited or costly, but not for replacing expert specification of the label schema.

The Gemini 2.5 Pro result provides an additional model-capacity reference point. Under the expert-crafted zero-shot configuration and without automated optimization, Gemini 2.5 Pro reached 0.861 accuracy. This closely matches the DSPy-optimized Gemini 2.5 Flash configuration but does not surpass the best Flash zero-shot result. The result suggests that increasing model capacity alone does not eliminate the remaining limitations of the generative branch, particularly for class-balanced recognition. It also supports the use of Gemini 2.5 Flash as the main optimization target, since it provides comparable or better performance in the evaluated configurations with a more cost-efficient profile.

The k-shot sensitivity analysis in Table 6 shows that few-shot prompting does not yield a clear universal winner. With the expert-crafted instruction fixed, Gemini 2.5 Flash achieved its highest accuracy in the zero-shot setting, while its highest Macro-F1 was obtained with one labeled exemplar. GPT-4o, in contrast, reached its best accuracy and Macro-F1 in the one-shot setting. Larger values of k did not provide consistent gains for either model. This pattern suggests that few-shot examples can help the models attend to minority-class distinctions, but their effect depends strongly on which examples are selected. In class-imbalanced CIC, a small set of exemplars may fail to represent the full range of rhetorical patterns, may overemphasize majority-class cues, or may anchor the model to surface similarities that do not generalize. A comparable pattern has been reported for English software citation intent classification, where few-shot prompting did not reliably improve over the zero-shot setting (Istrate et al., 2024). Koloveas et al. (2026), by contrast, found few-shot configurations to dominate their best-performing settings, indicating that the effect of k depends on the label schema and model family rather than being monotonic. Therefore, the results should be interpreted not as evidence that few-shot prompting is ineffective, but as evidence that naive increases in k are insufficient without deliberate exemplar selection.

Table 6: k-shot prompt sensitivity results for the generative branch. All configurations used the expert-crafted instruction $\mathcal{I}_{\text{expert}}$ with varying numbers of demonstrations $\mathcal{D}_k$.

	$\mathcal{D}_0$		$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_5$	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GPT-4o	0.821	0.574	0.827	0.651	0.801	0.627	0.807	0.582
Gemini 2.5 Flash	0.869	0.675	0.867	0.717	0.858	0.655	0.841	0.687

Overall, the generative branch establishes a strong training-free baseline for Turkish citation intent classification when guided by explicit, domain-aware instructions. Its main limitation is the persistent gap between accuracy and Macro-F1, which indicates that minority citation intents remain difficult to identify reliably under

prompting alone. This motivates the discriminative encoder branch, where supervised fine-tuning, loss-level reweighting, and controlled contextual and structural configurations are used to address class imbalance more directly.

5.3 Discriminative branch results

The discriminative branch was evaluated as the supervised, high-throughput pathway of the TurkCite framework. Table 7 reports the performance of five encoder backbones across the four classifier configurations defined in Section 4.3. Each configuration was evaluated over three independent random seeds, and results are reported as mean ± sample standard deviation. Pairwise model comparisons were assessed using paired bootstrap tests with 10,000 resamples and Holm–Bonferroni correction.

Table 7: Performance of five encoder architectures across four configurations. Each cell reports mean ± SD (%) over N=3 random seeds. Base = $\varphi_{\text{enc}}^{(1)}$; Hierarchical = $\varphi_{\text{enc}}^{(2)} \to \varphi_{\text{enc}}^{(3)}$. Best mean per column in **bold**.

Context	Structure	Model	Optimized for Accuracy		Optimized for Macro-F1	
			Accuracy	Macro-F1	Accuracy	Macro-F1
Standard (S)	Base	BERTurk	87.45 ± 0.55	69.60 ± 1.28	88.05 ± 0.42	**72.11** ± 1.54
		Turkish ELECTRA	88.11 ± 0.38	**72.58** ± 0.20	87.44 ± 1.11	71.71 ± 3.25
		XLM-RoBERTa	85.81 ± 0.32	67.02 ± 0.63	85.02 ± 1.34	65.06 ± 1.83
		DeBERTaV3	82.36 ± 1.27	60.92 ± 3.41	83.87 ± 0.76	60.65 ± 1.80
		ModernBERT	79.99 ± 1.11	51.62 ± 3.09	78.54 ± 1.27	48.23 ± 1.43
	Hierarchical	BERTurk	88.24 ± 1.03	68.47 ± 2.37	87.45 ± 0.31	67.59 ± 3.11
		Turkish ELECTRA	87.75 ± 0.84	68.30 ± 1.13	87.51 ± 0.27	66.85 ± 3.85
		XLM-RoBERTa	86.36 ± 0.73	64.08 ± 3.87	85.26 ± 1.46	65.14 ± 3.20
		DeBERTaV3	83.81 ± 0.19	58.29 ± 1.38	81.57 ± 0.27	55.77 ± 3.08
		ModernBERT	79.51 ± 1.41	47.66 ± 2.09	77.93 ± 2.00	45.76 ± 0.72
Rich (S_{rich})	Hierarchical	BERTurk	88.78 ± 0.69	65.41 ± 2.42	88.42 ± 0.64	64.51 ± 1.54
		Turkish ELECTRA	89.20 ± 0.28	65.38 ± 0.87	88.78 ± 0.21	64.26 ± 1.15
		XLM-RoBERTa	87.99 ± 0.48	64.76 ± 1.28	88.96 ± 1.37	65.36 ± 3.63
		DeBERTaV3	85.02 ± 1.52	57.40 ± 5.12	85.57 ± 0.90	59.75 ± 1.27
		ModernBERT	84.05 ± 1.03	54.27 ± 2.49	81.08 ± 0.95	54.33 ± 2.42
	Base	BERTurk	89.93 ± 0.46	68.82 ± 2.99	**89.39** ± 1.05	67.87 ± 1.16
		Turkish ELECTRA	**90.11** ± 0.73	68.48 ± 0.67	88.60 ± 1.15	68.45 ± 0.92
		XLM-RoBERTa	89.33 ± 0.76	67.96 ± 2.24	87.38 ± 3.11	67.16 ± 2.10
		DeBERTaV3	86.17 ± 1.93	61.96 ± 3.68	86.96 ± 1.00	66.71 ± 2.49
		ModernBERT	81.93 ± 1.64	55.29 ± 4.47	81.45 ± 1.49	54.49 ± 5.33

We first consider the Base-Standard configuration, $\varphi_{\text{enc}}^{(1)}(S)$, which serves as the unmodified single-stage reference setting. In this configuration, Turkish ELECTRA and BERTurk formed the strongest overall group among the evaluated encoders. Under accuracy-based checkpoint selection, Turkish ELECTRA achieved 88.11 ± 0.38% accuracy and 72.58 ± 0.20% Macro-F1, while BERTurk achieved 87.45 ± 0.55% accuracy and 69.60 ± 1.28% Macro-F1. The accuracy difference between these two models in the Base-Standard setting was statistically significant ($p_{\text{boot}} < 0.001$). Outside this baseline setting, the observed differences between Turkish ELECTRA and BERTurk were smaller and were not statistically reliable after correction for multiple

comparisons. The result is therefore best interpreted as identifying a strong Turkish-specific encoder group, rather than a fixed ordering between the two backbones across all configurations. In practical terms, Turkish-specific pre-training appears to provide the more stable advantage, while the relative choice between Turkish ELECTRA and BERTurk depends on the context representation and checkpoint-selection criterion.

The remaining encoders followed a generally lower performance pattern. XLM-RoBERTa provided the strongest multilingual baseline, while DeBERTaV3 and ModernBERT obtained lower scores in most settings. This result is consistent with the linguistic and domain-specific requirements of Turkish academic text, where language-specific pre-training appears to be beneficial. At the same time, the differences across configurations indicate that backbone choice interacts with the available contextual information, particularly for models without Turkish-specific pre-training.

Table 8 provides a class-level view of the Base-Standard setting using BERTurk from one representative seed. The model achieved strong performance on the dominant Background class, with an F1-score of 0.93. Performance was lower for the minority and semantically more demanding classes. In particular, Discuss and Differ obtained recall values of 0.43 and 0.50, respectively. This class-level pattern explains the persistent gap between accuracy and Macro-F1 in Table 7. The models can classify the majority class reliably, but distinguishing rarer rhetorical functions remains substantially harder.

Table 8: Per-class performance metrics for BERTurk (Base-Standard).

Intent Class	Precision	Recall	F1-score
Background	0.90	0.96	0.93
Basis	0.84	0.73	0.78
Support	0.69	0.69	0.69
Differ	0.83	0.50	0.62
Discuss	0.81	0.43	0.57
Macro Average	**0.81**	**0.66**	≈ 0.71

Representative single run, optimized for accuracy. Mean Macro-F1 across N=3 seeds: $69.60 \pm 1.28\%$; $N_{\text{test}} = 550$.

We next evaluate the structural modification introduced by the hierarchical classifier, $\varphi_{\text{enc}}^{(2)} \rightarrow \varphi_{\text{enc}}^{(3)}$. In the Standard-context setting, hierarchical decomposition did not improve performance over the single-stage baseline. Across the five encoders under accuracy-based checkpoint selection, the mean accuracy change was negligible (-0.05 percentage points, $p = 0.8746$), while Macro-F1 decreased by approximately 3.02 percentage points ($p < 0.001$, Cohen's $d = -0.882$). For example, Turkish ELECTRA decreased from $72.58 \pm 0.20\%$ Macro-F1 in the Base-Standard configuration to $68.30 \pm 1.13\%$ in the Hierarchical-Standard configuration. BERTurk showed a smaller but similar decrease, from $69.60 \pm 1.28\%$ to $68.47 \pm 2.37\%$. These results suggest

that separating Background from Non-Background citations does not, by itself, provide enough additional discriminative signal for minority-intent classification. Error propagation from the first stage may also contribute to this pattern.

The contextual modification produced a different effect. Moving from the standard input S to the rich input S_{rich} yielded a statistically significant mean accuracy gain of 2.31 percentage points across the tested configurations ($p < 0.001$, Cohen's $d = 1.335$). The strongest mean accuracy was obtained by Turkish ELECTRA in the Rich-Base configuration, reaching 90.11±0.73%, followed closely by BERTurk with 89.93±0.46%. This indicates that section-level and neighboring-sentence information can improve overall classification reliability.

However, the same contextual enrichment did not produce a comparable improvement in Macro-F1. Across the tested configurations, the mean Macro-F1 change associated with rich context was only +0.12 percentage points and was not statistically significant ($p = 0.8824$). In the Rich-Base configuration, Turkish ELECTRA achieved $68.48 \pm 0.67\%$ Macro-F1 and BERTurk achieved $68.82 \pm 2.99\%$, both below the best Macro-F1 observed in the Base-Standard setting. This asymmetry suggests that additional context mainly improves overall accuracy, likely by helping models identify common citation patterns more reliably, but it does not fully resolve the ambiguity among minority citation intents.

We also examined whether Macro-F1-based checkpoint selection could improve class-balanced performance. Across the 20 model–configuration combinations in Table 7, selecting checkpoints by validation Macro-F1 did not yield a consistent improvement over selecting by validation accuracy. The average Macro-F1 change was -0.32 percentage points, with positive gains in only 6 of the 20 cells. The largest isolated improvement occurred for BERTurk in the Base-Standard configuration, where Macro-F1 increased from $69.60 \pm 1.28\%$ to $72.11 \pm 1.54\%$. Nevertheless, the highest mean Macro-F1 in the full experimental design remained Turkish ELECTRA under the Base-Standard configuration with accuracy-based checkpoint selection, at $72.58 \pm 0.20\%$. This indicates that changing the checkpoint-selection target alone is not sufficient to overcome the minority-class limitations of the current dataset.

To better understand whether this limitation is related to data availability, Figure 5 presents learning curves for BERTurk in the Base-Standard configuration. Accuracy plateaued relatively early, whereas Macro-F1 continued to increase as the training set grew. This divergence suggests that the model learns the dominant Background class with comparatively little data, while minority-intent discrimination remains under-saturated at the current corpus size. The result indicates that further improvements in class-balanced performance are likely to require targeted expansion of underrepresented classes, rather than relying only on additional architecture-level tuning.

Because the hierarchical models depend on the first-stage Background versus Non-Background decision, we further examined whether their performance was sensitive to the Background confidence threshold. This diagnostic analysis tests whether a less conservative Background gate could route more instances to the second-stage intent classifier and thereby improve class-balanced performance. We therefore re-evaluated a single representative run of the Hierarchical-Standard configuration under Background

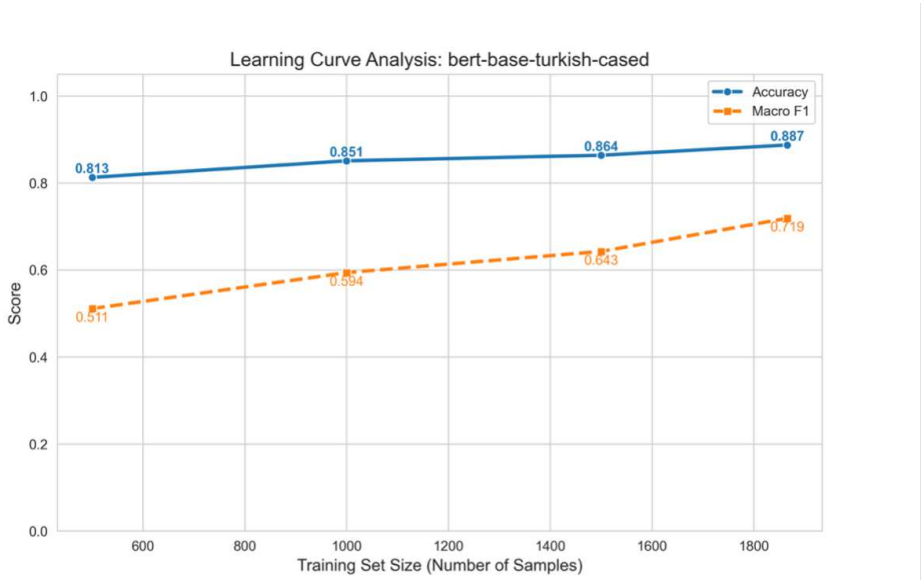


Fig. 5: Learning curves for BERTurk (Base-Standard configuration) showing Accuracy and Macro-F1 as a function of training set size. Accuracy plateaus early while Macro-F1 continues to improve, indicating unsaturated minority class learning.

thresholds of 0.70 and 0.85. Raising the threshold changed the accuracy–Macro-F1 trade-off, but did not produce a consistent improvement in Macro-F1. For Turkish ELECTRA, accuracy decreased from 0.854 to 0.716 and Macro-F1 from 0.692 to 0.578 across the two thresholds, while XLM-RoBERTa remained comparatively stable. The broader pattern therefore does not support threshold adjustment as a reliable remedy for the limitations of the hierarchical configuration.

Overall, the discriminative branch provides the strongest controlled pathway for large-scale Turkish citation intent classification. Turkish-specific encoders, particularly Turkish ELECTRA and BERTurk, achieved the best overall results, with mean accuracies approaching 90% in enriched-context configurations. Contextual enrichment was the most effective intervention for improving accuracy. Hierarchical decomposition did not provide a reliable benefit and was associated with lower class-balanced performance in several settings. Taken together, these findings indicate that the supervised encoder branch is suitable for scalable bibliometric indexing, while further improvements in minority citation intents will likely require targeted expansion of the annotated corpus.

6 Discussion

The dual-branch framework evaluated in this study indicates that Turkish citation intent classification should not be treated as a single-model selection problem. The generative and discriminative branches occupy different positions in the trade-off

between labeled-data requirements, inference cost, output stability, and class-balanced performance. This complementarity is important for bibliometric infrastructures such as TRDizin, where citation intent classification may be needed both for exploratory analysis in low-resource settings and for continuous indexing once annotated data are available.

The generative branch provides a strong training-free baseline and is particularly useful for cold-start or auditing workflows. Its performance shows that large language models can recover meaningful citation-intent signals from Turkish academic text without parameter updates. However, the branch remains sensitive to prompt design, exemplar selection, provider-side model changes, and API cost. The discriminative branch, by contrast, is the more controlled pathway for scalable indexing. Turkish-specific encoders achieved the strongest reported mean results, with rich-context inputs improving overall accuracy and locally deployable models offering lower marginal inference cost after training. Table 9 summarizes these practical trade-offs.

Table 9: Summary of practical trade-offs between the generative and discriminative branches.

Property	**Generative Branch**	**Discriminative Branch**
Best reported Accuracy	0.869	0.901
Best reported Macro-F1	0.717	0.726
Labeled Training Data	Not required for zero-shot use	Required for fine-tuning
Inference Setting	API-dependent	Local deployment possible
Marginal Inference Cost	Higher	Lower after training
Output Stability	Sensitive to prompts and provider updates	Stable under fixed checkpoint
Adaptation Mechanism	Prompt revision or optimization	Fine-tuning or retraining

The most important limitation across both branches is the persistent gap between accuracy and Macro-F1. Contextual enrichment improved accuracy, but it did not produce a comparable gain in class-balanced performance. Similarly, hierarchical decomposition did not reliably improve minority-intent recognition, suggesting that the main bottleneck is not the classifier configuration alone. The learning-curve analysis further indicates that Macro-F1 remains under-saturated at the current corpus size. For bibliometric use, this distinction is important: Background citations dominate the corpus, but minority classes such as Support, Differ, and Discuss are central to richer citation impact profiles. Accuracy alone can therefore obscure weaknesses in analytically important citation-intent categories.

These findings also have implications for non-English citation intent classification. Direct comparison with English benchmarks such as SciCite is difficult because TurkCite differs in language, dataset size, and label granularity. The results should therefore be viewed as an encouraging Turkish benchmark rather than as a direct

numerical counterpart to English datasets. A more informative comparison is with non-English CIC work conducted under broadly similar conditions. Budi and Yaniasih (2023) report a Macro-F1 of 0.88 for citation function on Indonesian journal articles, above the best Macro-F1 obtained here. Two differences are relevant for interpreting this gap. Their labeled set is roughly three times larger, and their citation-function distribution is considerably more balanced, with the largest class covering approximately 44% of instances compared with 78% in TurkCite. This distribution was itself the result of a deliberate schema revision intended to produce more balanced categories. The comparison therefore suggests that class-balanced performance in CIC depends not only on corpus size and model capacity but also on the granularity of the label schema. The five-class WoS taxonomy adopted here preserves distinctions that are analytically valuable for bibliometric use, particularly between Basis and Support, but it also concentrates a large proportion of instances in a single class. This trade-off is consistent with our observation that the minority-class bottleneck persists across architectures, context representations, and checkpoint-selection targets. At the same time, the weak translation-transfer baseline and the strong performance of Turkish-specific encoders support the need for annotated resources and encoder models developed for the target language. TurkCite therefore provides both a Turkish CIC resource and a controlled evaluation framework for studying citation function in scholarly ecosystems that remain underrepresented in English-centered bibliometric research.

7 Conclusion and future work

This study introduced TurkCite, a citation intent classification framework and expert-annotated dataset for Turkish academic literature. The corpus contains 2,746 citation instances annotated under a five-class schema adapted from established citation-intent taxonomies. The expert-centered annotation workflow achieved high inter-annotator agreement ($\kappa = 0.837$), supporting the reliability of the dataset while also reflecting the interpretive nature of fine-grained citation intent classification. By focusing on Turkish academic texts, TurkCite addresses a gap in citation intent classification research, which has largely been shaped by English-language resources and benchmarks.

The experiments indicate that translation-based transfer from English is insufficient for Turkish citation intent classification, motivating the need for language-specific resources and modeling workflows. The generative branch provided a strong training-free baseline, making it useful for cold-start settings where annotated data are unavailable or limited. Prompt optimization was most effective when anchored in a meaningful task-specific seed prompt, suggesting that automated prompt refinement can support prompt stability but does not replace task-informed prompt design. The discriminative branch offered the more controlled pathway for scalable indexing once annotated data were available. Turkish-specific encoders achieved the strongest overall results, with the best mean accuracy reaching 90.11% and the best mean Macro-F1 reaching 72.58%. At the same time, the experiments showed that class-balanced

performance remains constrained by underrepresented citation intents. Richer context improved overall accuracy, but hierarchical decomposition and Macro-F1-based checkpoint selection did not resolve the minority-class bottleneck.

Beyond the Turkish case, TurkCite provides an example workflow for developing citation intent analysis in non-English bibliometric settings. The framework illustrates several practical pathways that can be selected according to data availability and institutional needs: a translation-based baseline for testing cross-lingual transfer, a generative branch for low-resource or exploratory use, and a supervised discriminative branch for high-throughput indexing when annotated data are available. This layered design may be useful for national indexing platforms and scholarly infrastructures that seek to move beyond raw citation counts toward more informative citation impact profiles, but lack mature citation-intent resources in the target language.

Future work should prioritize targeted expansion of underrepresented classes such as Differ and Support, since the learning-curve analysis suggests that minority-intent recognition remains under-saturated at the current corpus size. Since the citation extraction pipeline already yields a substantially larger pool of unannotated Turkish citation contexts than the annotated subset, semi-supervised approaches that exploit unlabeled data, such as the GAN-BERT-based framework proposed by Chatrinan, Noraset, and Tuarob (2025), offer a complementary route to improving minority-intent recognition without a proportional increase in annotation cost. Validation across additional disciplines beyond Computer Science is also needed, as citation practices may vary across fields. Incorporating citation-network structure, document-level metadata, and broader discourse signals may provide complementary evidence for difficult intent distinctions. More broadly, the TurkCite framework can inform similar efforts in other non-English scholarly ecosystems, provided that annotation design, model selection, and evaluation protocols are adapted to the linguistic and bibliometric conditions of the target context.

A Appendix

This appendix provides additional details for the generative branch. Appendix A.1 summarizes the DSPy/MIPROv2 prompt optimization procedure. Appendix A.2 presents the three prompt formulations used to evaluate prompt evolution: the basic instruction prompt ($\mathcal{I}_{\text{basic}}$), the detailed instruction prompt ($\mathcal{I}_{\text{detailed}}$), and the expert-crafted prompt ($\mathcal{I}_{\text{expert}}$).

A.1 Prompt Optimization Procedure

Algorithm 1 Prompt optimization procedure for the generative branch

Input: Training set D_{train}, internal validation set D_{val}, seed prompt $\mathcal{P}_{\text{seed}}$, number of trials $N = 27$

Output: Optimized prompt $\mathcal{P}^*$

Initialize prompt candidate space $\mathbb{P} \leftarrow \{\mathcal{P}_{\text{seed}}\}$
Generate 18 candidate instruction variants
Construct 9 few-shot exemplar sets from D_{train}, with at most $k = 6$ exemplars per set
Populate $\mathbb{P}$ with candidate instruction and exemplar-set combinations
Initialize best score $S_{\max} \leftarrow 0$
Initialize best prompt $\mathcal{P}^* \leftarrow \mathcal{P}_{\text{seed}}$
for $t = 1$ **to** N **do**
    Select candidate prompt $\mathcal{P}_t \in \mathbb{P}$
    Evaluate $S_t \leftarrow \text{ExactMatch}(\varphi_{\text{LLM}}(\mathcal{P}_t, D_{\text{val}}))$
    if $S_t > S_{\max}$ **then**
        $S_{\max} \leftarrow S_t$
        $\mathcal{P}^* \leftarrow \mathcal{P}_t$
    end if
end for
return $\mathcal{P}^*$

A.2 Prompt Templates

Initial Prompt for Citation Classification

```
You are a helpful AI assistant. Below, some academic citations are listed
↪  along
with their titles and contents.
Classify each citation into one of the following categories: background,
↪  basis,
discuss, support, differ, or other. Please return only a valid JSON array.
Do not add any other explanation.
Format example: [ {"id": "123", "intent": "background"} ... ]
```

Fig. 6: Stage 1: Basic instruction prompt ($\mathcal{I}_{\text{basic}}$) containing only label names and a classification directive, without category definitions or persona framing.

Detailed Prompt for Citation Classification

```
You are a helpful AI assistant. Below, some academic citations are listed
↪  along with their titles and contents.
Classify each citation into one of the following six categories:
- background: General knowledge in the literature or a general reference to
previous studies
- basis: The dataset, method, or theoretical framework that is the
↪  foundation of
this study
- discuss: A source that is discussed in detail within the study
- support: If the outputs of this study support the results of the cited
↪  study,
usually seen in the conclusion section
- differ: If the outputs of this study contradict the results of the cited
↪  study,
```

```
usually seen in the conclusion section
- other: Situations that fall outside of the above
Please return only a valid JSON array. Do not add any other explanation.
Format example: [{"id": "123", "intent": "background"}, ...]
```

Fig. 7: Stage 2: Detailed instruction prompt ($\mathcal{I}_{\text{detailed}}$) incorporating per-category descriptions in Turkish, providing structured guidance beyond bare label names.

Expert-crafted Prompt for Citation Classification

```
Role: You are an expert academic editor specializing in computer science and
↪  artificial
intelligence.
Task: Your task is to meticulously analyze and classify academic citations
↪  from  ...
Goal: Your goal is to classify each citation into one of the following five
↪  ...
Categories:
1. background:
    - Description: The cited work is referred to for general context,
    ↪  historical ...
    - Typical Turkish Sections: 'Giriş' (Introduction), 'Literatür Taraması'
    ↪  ...
    - Key Idea: Provides broader context or acknowledges foundational
    ↪  knowledge that ...
2. basis:
    - Description: The cited work is actively and substantively discussed,
    ↪  analyzed, ...
    - Typical Turkish Sections: 'Literatür Taraması' (Literature Review),
    ↪  'Tartışma' ...
    - Key Idea: The cited work is a subject of significant intellectual
    ↪  engagement ...
3. discuss: ...
4. support: ...
5. differ: ...
Important Considerations for Classification:
    - Language: The citation text and section titles in the input JSON will
    ↪  be  ...
    - Context is Key: While the 'section' (Turkish section title) where the
    ↪  citation ...
    - Input Format:
    Example of Input Data Structure:
    [{  "id": "unique_id_1",
        "citation": "Alanyazında bu konuda farklı yaklaşımlar mevcuttur
        ↪  (Karaca, 2022).",
        "section": "Giriş" }]
Expected Output Format:
    The output must be a JSON array of objects. Each object should contain
    ↪  the id of ...
    Example of Output Data Structure:
    [{ "id": "unique_id_1", "intent": "intent_1" }]
Please return your response strictly as a valid JSON array. Do not include
↪  any additional
commentary, explanation, text, or formatting outside of the JSON array
↪  itself.
```

Fig. 8: Stage 3: Expert-crafted prompt ($\mathcal{I}_{\text{expert}}$) incorporating Web of Science category definitions and academic-persona framing. Used as the fixed instruction across all k-shot configurations.

Acknowledgements. The citation extraction module (CEX) used in this study was developed within the GRASPOS project. The numerical calculations reported in this paper were partially performed at TÜBİTAK ULAKBİM, High Performance and Grid Computing Center (TRUBA resources). The authors thank the domain experts who contributed to the annotation of the TurkCite dataset.

Statements and Declarations

Funding. No specific grant or external funding was received for the work reported in this manuscript. The research was carried out as part of the authors' institutional duties, and computational resources were provided by TÜBİTAK ULAKBİM and TOBB University of Economics and Technology.

Competing Interests. Financial interests: The authors declare they have no financial interests. No specific grant or external funding was received for the work reported in this manuscript. Non-financial interests: Kemal Sami Karaca is employed by TÜBİTAK ULAKBİM, the institution that operates the TRDizin and DergiPark platforms discussed in this study as potential deployment environments, and that hosts both the annotation platform (https://atif.ulakbim.gov.tr) and the open data repository (Aperta) used in this work. Computational resources were provided by TÜBİTAK ULAKBİM (TRUBA) and TOBB University of Economics and Technology as part of the authors' institutional duties. Bahaeddin Eravci declares no competing interests. Neither author holds an editorial, advisory, or other role with this journal.

Ethics approval. Not applicable. This study did not involve human participants or animals. It is based on the analysis of publicly available scholarly articles indexed on DergiPark and TRDizin.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The dataset generated in this study, "Türkçe akademik makaleler için atıf niyeti veri seti [Citation Intent Dataset for Turkish Academic Articles]" (TurkCite, Version 1), comprising 2,746 expert-annotated Turkish citation contexts, is publicly available without restriction in the Aperta open access archive at https://doi.org/10.48623/aperta.286647. The source articles from which citation contexts were extracted are publicly accessible through the DergiPark (https://dergipark.org.tr) and TRDizin (https://trdizin.gov.tr) national indexing platforms. The web-based annotation platform developed for this study is publicly accessible at https://atif.ulakbim.gov.tr.

Code availability. The source code for the citation extraction pipeline, the prompt optimization procedure, and the encoder fine-tuning experiments is openly available at https://github.com/TOBB-University-Master/CitationIntentClassifier.

Author contributions. K.S.K. and B.E. conceived the study and designed the methodology. K.S.K. developed the software, curated the dataset, performed the formal analysis and investigation, prepared all figures and tables, and wrote the main manuscript text. B.E. supervised the work. K.S.K. and B.E. reviewed and edited the manuscript. All authors reviewed and approved the final version.

Prior publication. A preliminary version of this work appeared in Karaca and Eravcı (2025). The present manuscript is a substantially extended version; material carried over from the conference paper has been rewritten, and the relationship between the two is described in Sections 2 and 3.5.

References

Amasyali, M.F., & Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. *International conference on applications of natural language to information systems* (pp. 221–226).

Atılgan, D., Atakan, C., Bulut, B. (2008). Türkçe kütüphanecilik dergilerinin atıf analizi. *Türk Kütüphaneciliği*, *22*(4), 392–413,

Beltagy, I., Lo, K., Cohan, A. (2019, November). SciBERT: A pretrained language model for scientific text. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3615–3620). Hong Kong, China: Association for Computational Linguistics.

Berrebbi, D., Huynh, N., Balalau, O. (2022). GraphCite: Citation intent classification in scientific publications via graph embeddings. *Companion proceedings of the web conference 2022* (p. 779–783). New York, NY, USA: Association for Computing Machinery.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th international conference on neural information processing systems.*

Budi, I., & Yaniasih, Y. (2023). Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*, *128*, 735-759, https://doi.org/10.1007/s11192-022-04567-4

Chatrinan, K., Noraset, T., Tuarob, S. (2025). GAN-CITE: leveraging semi-supervised generative adversarial networks for citation function classification with limited data. *Scientometrics*, *130*(2), 679–703, https://doi.org/10.1007/s11192-025-05233-1

Clarivate (2022). *Web of science enriched cited references: Citation context classifications.* https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2022/05/202205-WoS-citation-context-final.pdf. (Access Date: 07.01.2026)

Clark, K., Luong, M., Le, Q.V., Manning, C.D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, april 26-30, 2020.*

Cohan, A., Ammar, W., van Zuylen, M., Cady, F. (2019, June). Structural scaffolds for citation intent classification in scientific publications. *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3586–3596). Minneapolis, Minnesota: Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46,

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Association for Computational Linguistics.

Dehkharghani, R., Saygin, Y., Yanikoglu, B., Oflazer, K. (2016). SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*, *50*(3), 667–685,

Deng, J., Tong, L., Zhang, J., Li, M. (2025). Contemporary technological value of ancient chinese medical books: an empirical patent citation analysis using large language models. *The Electronic Library*, *44*, 51-72, https://doi.org/10.1108/EL-06-2025-0221 Retrieved from https://www.sciencedirect.com/science/article/pii/S0264047325000281

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). 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: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Edmundson, H.P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264–285,

Fernando, C., Banarse, D.S., Michalewski, H., Osindero, S., Rocktäschel, T. (2024, 21–27 Jul). Promptbreeder: Self-referential self-improvement via prompt evolution. *Proceedings of the 41st international conference on machine learning* (Vol. 235, pp. 13481–13544). PMLR.

Garzone, M., & Mercer, R.E. (2000). Towards an automated citation classifier. *Conference of the canadian society for computational studies of intelligence* (pp. 337–346).

Gemini Team, Google (2024). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, , Retrieved from https://arxiv.org/abs/2312.11805

GRASPOS (2024). *GRASPOS.* https://graspos.eu/. (Access Date: 20.05.2025)

He, P., Gao, J., Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, may 1-5, 2023.*

Istrate, A.-M., Fisher, J., Yang, X., Moraw, K., Li, K., Li, D., Klein, M. (2024). Scientific software citation intent classification using large language models. *Natural scientific language processing and research knowledge graphs* (p. 80-99). Springer Nature Switzerland.

Jin, X., & Han, J. (2011). K-means clustering. C. Sammut & G.I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 563–564). Boston, MA: Springer US.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, *6*, 391–406,

Karaca, K.S., & Eravcı, B. (2025). A large-scale dataset and citation intent classification in turkish with llms. *2025 10th international conference on computer science and engineering (ubmk)* (p. 509-514).

Karaca, K.S., & Eravcı, B. (2026). *Türkçe akademik makaleler için atıf niyeti veri seti [citation intent dataset for turkish academic articles].* Ankara, Turkey: TÜBİTAK ULAKBİM. Retrieved from https://doi.org/10.48623/aperta.286647 (Data set, Aperta Open Access Archive)

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., A, S.V., ... Potts, C. (2024). Dspy: Compiling declarative language model calls into state-of-the-art pipelines. *The twelfth international conference on learning representations.*

Kipf, T.N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International conference on learning representations (iclr).*

Koloveas, P., Chatzopoulos, S., Vergoulis, T., Tryfonopoulos, C. (2026). Can llms predict citation intent? an experimental analysis of in-context learning and fine-tuning on open llms. *Linking theory and practice of digital libraries* (p. 207-224). Springer Nature Switzerland.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174,

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Research and advanced technology for digital libraries* (pp. 473–474). Berlin, Heidelberg: Springer Berlin Heidelberg.

Moravcsik, M.J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science*, *5*(1), 86–92,

Oflazer, K., & Saraçlar, M. (2018). *Turkish natural language processing.* Springer.

Paolini, L., Vahdati, S., Di Iorio, A., Wardenga, R., Heibi, I., Peroni, S. (2025). CiteFusion: an ensemble framework for citation intent classification harnessing dual-model binary couples and SHAP analyses. *Scientometrics*, *130*, 5911–5981, https://doi.org/10.1007/s11192-025-05418-8

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65,

Schweter, S. (2020). *BERTurk - BERT models for Turkish.* Zenodo.

Taşkın, Z., & Al, U. (2018, Jan 01). A content-based citation analysis study based on text categorization. *Scientometrics*, *114*(1), 335-357, https://doi.org/10.1007/s11192-017-2560-2

Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text* (Unpublished doctoral dissertation). University of Edinburgh Edinburgh, Scotland.

Tsai, H.-J., Yen, A.-Z., Huang, H.-H., Chen, H.-H. (2023). Citation intent classification and its supporting evidence extraction for citation graph construction. *Proceedings of the 32nd acm international conference on information and knowledge management* (p. 2472–2481).

TÜBİTAK ULAKBİM (2024a). *DergiPark Akademik.* https://dergipark.org.tr/. (Access Date: 20.05.2025)

TÜBİTAK ULAKBİM (2024b). *TR Dizin.* https://trdizin.gov.tr/. (Access Date: 20.05.2025)

TÜBİTAK ULAKBİM (2026). *TR Dizin citation index system.* https://atif.ulakbim.gov.tr. (Access Date: March 5, 2026)

van Dinter, R., Catal, C., Tekinerdogan, B. (2021). A decision support system for automating document retrieval and citation screening. *Expert Systems with Applications*, *182*, 115261, https://doi.org/10.1016/j.eswa.2021.115261 Retrieved from https://www.sciencedirect.com/science/article/pii/S095741742100693X

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., ... Poli, I. (2024). *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.* Retrieved from https://arxiv.org/abs/2412.13663

Yang, C., He, X., Li, C., Cao, L., Tang, J. (2023). Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, , Retrieved from https://arxiv.org/abs/2309.03409

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*, ,

Yeniterzi, R. (2011, June). Exploiting morphology in Turkish named entity recognition system. *Proceedings of the ACL 2011 student session* (pp. 105–110). Portland, OR, USA: Association for Computational Linguistics.

Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J. (2023). Large language models are human-level prompt engineers. *The eleventh international conference on learning representations.*