Supervision versus Demonstration-Based In-Context Learning for Multiword Expression Classification

Sercan Karakaş
University of Chicago
skarakas@uchicago.edu

Yusuf Şimşek
Fırat University
ysimsek@firat.edu.tr

Abstract

Turkish idiomatic light verb constructions (LVCs) are challenging for multiword expression processing because they often share the same surface form as fully literal verb–object combinations while functioning as a single, partially idiomatic predicate. We frame Turkish LVC detection as a binary classification task (literal meaning vs. idiomatic meaning) and evaluate on a manually created controlled set (N=147) with matched negatives: out-of-domain random sentences and in-domain literal controls (NLVC), alongside LVC positives. We compare a supervised Turkish encoder baseline (BERTurk with a classifier head) to three instruction-tuned LLMs from different families under zero-shot, one-shot, and few-shot prompting, and analyze how demonstrations shift error profiles. In zero-shot, LLMs perform well on negatives but show very low LVC recall. One-shot prompting sharply improves LVC detection but can induce strong, model-specific biases (over- vs. under-predicting LVC). A richer few-shot prompt improves calibration and yields robust overall performance for GPT-OSS-20B and Qwen 2.5-14B. Overall, the results highlight substantial prompt sensitivity in Turkish metalinguistic classification: the supervised baseline remains competitive, while prompted LLMs can match or exceed it on LVCs with carefully constructed demonstrations. We release code, prompts, and evaluation materials to support reproducibility.[1].

1 Introduction

Multiword expressions (MWEs) are, by definition, sequences of two or more words that behave as a single linguistic unit, often displaying semantic idiosyncrasy, non-compositional meaning, and/or distinctive syntactic behavior (Sag et al., 2002; Odijk, 2013; Mititelu et al., 2025). Because MWEs are pervasive in everyday language use, they pose persistent challenges for Natural Language Processing (NLP) systems and constitute a well-known hurdle for non-native speakers learning a language (Sag et al., 2002; Ramisch, 2015; Barman et al., 2024).

Psycholinguistic work indicates that children and adults in both L1 and L2 track distributional statistics of multiword sequences, and that frequent multiword units yield measurable processing advantages (Arnon and Clark, 2011; Siyanova-Chanturia et al., 2011; Castroviejo et al., 2024). If frequent MWEs are entrenched in comprehension and production, as acquisition research suggests (Arnon and Clark, 2011; He and Wittenberg, 2020; Schulz, 2024; Chiang, 2025), then NLP systems intended to model or support language use should handle them reliably across contexts. More broadly, NLP evaluation faces the challenge that many surface forms can realize the same content; automatic metrics are therefore difficult to interpret, and evaluation goals are often under-specified, making human-centered criteria essential (Zhou et al., 2022). In our setting, this motivates treating human-linguistic distinctions (e.g., whether a verb–nominal sequence forms an idiomatic/light-verb predicate unit vs. a literal verb–argument configuration) as a core evaluation target: for downstream Turkish NLP, models should reproduce these basic lexical-semantic generalizations stably across prompting regimes.

To address this, we frame Turkish LVC detection as a binary classification task and evaluate a supervised Turkish encoder baseline (BERTurk) alongside three instruction-tuned LLMs across varying prompting regimes on a manually created controlled set (N=147). Overall results highlight substantial prompt sensitivity : zero-shot LLMs show low LVC recall, whereas carefully constructed few-shot demonstrations improve calibration, allowing prompted LLMs to match or exceed the highly competitive supervised baseline. The remainder of the paper is organized as follows. Section 2 reviews prior work on MWE extraction in general and on

[1] Code and data repository

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 800–815

Turkish verbal MWEs in particular. Section 3 introduces the present study and motivates the literal–idiomatic contrast used to diagnose LVC detection. Section 4 describes the dataset construction, annotation procedure, and evaluation design. Section 5 presents the supervised BERTurk baselines and the instruction-tuned LLMs evaluated in the study. Section 6 reports the three experiments, comparing zero-shot, one-shot, and few-shot prompting regimes. Section 7 discusses the broader implications of the results, especially the role of prompt sensitivity and task-specific supervision.

2 Related Work

2.1 MWE extraction in general

A standard distinction in MWE processing is between discovery, which mines candidate expressions from raw text, and identification, which detects and types MWEs in context, typically via supervised tagging or structured prediction. Because many MWEs display substantial syntactic and morphological variability, they cannot be modeled simply as "words with spaces"; effective systems therefore combine lexical and morphosyntactic cues, sometimes in interaction with downstream tasks such as parsing or machine translation (Constant et al., 2017). Alongside work on specific MWE classes such as verb–particle constructions (Baldwin and Villavicencio, 2002), shared guidelines and multilingual benchmarks have advanced the field, especially through the PARSEME shared task on verbal MWEs, which explicitly targets variability and discontinuity across languages (Savary et al., 2017). More recently, MWE identification has largely shifted to neural contextual tagging models, from BiLSTM-CRF systems to transformer encoders, often with subword modeling to reduce sparsity and improve generalization across inflectional variants (Berk et al., 2018; Premasiri and Ranasinghe, 2022; Miletić and Schulte im Walde, 2024). However, strong token-level performance does not necessarily imply robust semantic modeling. Recent syntheses suggest that transformer-based systems remain inconsistent across settings and may over-rely on surface patterns or memorized lexical cues, especially for idiomatic or semi-lexicalized expressions. Meanwhile, newer PARSEME releases have expanded coverage and improved annotation consistency, enabling more controlled evaluation of discontinuous and morphologically complex VMWEs across languages (Savary et al., 2023).

2.2 Turkish

For Turkish, rich inflection and derivation greatly increase surface variation, making accurate lemmatization and morphosyntactic annotation critical for MWE typing and generalization (Oflazer, 1993; Oflazer et al., 2004; Karakaş and Şimşek, 2026); accordingly, early work tightly coupled morphological analysis with MWE processing (Oflazer et al., 2004). Turkish also has a highly productive inventory of verb–nominal predicates, especially light verb constructions (LVCs), where the nominal contributes core event semantics while the verb is semantically bleached and mainly provides functional material (e.g., event licensing, inflection) (Butt, 2010; Uçar, 2010; Özbek, 2010). Because many light verbs also appear in fully literal transitive uses, the boundary between lexicalized VMWEs and ordinary verb–object combinations is often blurred. Recent work makes use of syntactically annotated resources, including dependency-parsing studies (Eryiğit et al., 2011) and treebank-level MWE annotation supporting corpus analysis and supervised identification (Eryiğit et al., 2015). Within PARSEME, Turkish verbal MWEs are explicitly typed (including LVC subclasses), enabling shared evaluation and cross-lingual comparability (Savary et al., 2017). Still, representation is difficult: UD offers `compound:lvc`, but deciding whether a nominal is part of a lexicalized predicate (vs. an ordinary object) often requires semantic judgments beyond syntax-only cues (ud-, k). Consistently, Turkish corpus work shows that correcting lemma/POS information yields measurable gains in VMWE identification, critically showing how strongly LVC detection depends on robust morphosyntactic preprocessing (Öztürk et al., 2022) and possibly linguistically-motivated tokenization for large language models (Bayram et al., 2025, 2026).

Light verb predicates are also theoretically and experimentally informative because they can decouple surface argument structure from event interpretation. Superficially, many LVCs mirror ditransitives that encode transfer (e.g., *Ece Murat'a keman-ı ver-di* 'Ece gave Murat the violin'), where syntactic arguments map transparently onto giver, theme, and recipient roles. In LVCs such as *Ece Murat'a öpücük ver-di* ('Ece gave Murat a kiss'), however, the same frame admits competing analyses: comprehenders may treat the construction as a

two-participant predicate akin to *Ece Murat'ı öp-tü* ('Ece kissed Murat'), or as a three-role structure in which the nominal behaves like an additional theme (Özge et al., 2022). This mismatch between surface form and predicate meaning is precisely what makes Turkish LVCs a targeted test case for MWE-aware models.

3 Present Study

The central aim of this paper is to investigate Turkish verbal multiword expressions with an explicit focus on the *literal vs. idiomatic* contrast in verb–nominal predicates. MWEs are a well-known challenge for NLP because they frequently violate surface-based expectations about compositionality, lexical selection, and distribution, and their behavior varies from semi-compositional to strongly idiomatic (Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017; Savary et al., 2017). As we discussed in the previous section, for Turkish specifically, rich morphology and productive verb–nominal predicate formation make MWE identification particularly salient (Oflazer, 1993; Oflazer et al., 2004).

In addition to standard verbal MWEs, we explicitly target *light verb constructions* (LVCs) (and closely related idiomatic verb–nominal predicates) in order to reach human standards since human-centered criteria are especially important when defining what counts as success for NLP models (Zhou et al., 2022).

In these configurations in general and in Turkish, a nominal element contributes the core predicational content while the verb is partially semantically bleached, yielding a single predicate-like unit (Grimshaw and Mester, 1988; Sağ, 2015). Crucially, many Turkish light verbs also occur in fully literal transitive uses, creating hard minimal contrasts for both humans and NLP models.

3.1 Literal–Idiomatic (LVC) Contrast via Lexical Controls

Our design isolates idiomatic/LVC predication from ordinary verb–argument composition by constructing matched items where the *same verb* appears in (i) an idiomatic/LVC predicate and (ii) a literal use. This prevents trivial verb-only heuristics and forces models to attend to whether the verb–nominal sequence functions as a unitary predicate.

(1a) Ali Ayşe-ye ilham ver-di.
Ali Ayşe-DAT inspiration give-PST-3SG
'Ali inspired Ayşe.' (LVC; [1])

(1b) Ali Ayşe-ye kalem ver-di.
Ali Ayşe-DAT pen give-PST-3SG
'Ali gave Ayşe a pen.' (literal; [0])

Thus, (1a) realizes an LVC in which the nominal *ilham* supplies the core predicate meaning and *ver-* is light, whereas (1b) is a literal transfer event where *ver-* retains its basic 'give' meaning and *kalem* is an ordinary theme.

4 Methods

For this study, we evaluate (i) a Turkish encoder baseline with a task-specific classifier head (BERTurk) trained on Turkish treebanks data, and (ii) three modern large language models from different model families under three prompting regimes: zero-shot (instruction only), one-shot (instruction + one positive (LVC) and one negative (NLVC) example per target verb template), and few-shot (instruction + a compact set of labeled examples per template). These settings operationalize in-context learning and prompting-based adaptation (Brown et al., 2020; Liu et al., 2023).

To train the BERTurk classifier heads, we compiled supervision from Turkish Universal Dependencies (UD) treebanks distributed in CoNLL-U format: UD Turkish-Atis, UD Turkish-BOUN, UD Turkish-FrameNet, UD Turkish-GB, UD Turkish-IMST, UD Turkish-Kenet, UD Turkish-PUD, UD Turkish-Penn, and UD Turkish-Tourism (ud-, a,b,c,d,e,f,h,g,i). Each sentence includes POS tags, morphological features, and dependency relations.

To identify candidate LVC realizations from UD annotations, we exploited the dependency relations `compound:lvc` and `compound` (ud-, k,j). For treebanks that explicitly annotate LVCs with `compound:lvc`, we used those arcs directly as LVC candidates. For treebanks that do not contain `compound:lvc`, we followed an alternative procedure: (i) extract all `compound` dependencies, (ii) retain only those that form noun–verb dependencies (a nominal dependent linked to a verbal head), and (iii) manually review these candidates to remove non-LVC cases and to finalize a linguistically coherent set of target LVC patterns. Sentences containing validated LVC patterns were labeled as positive ([1]), and sentences without such patterns were labeled as negative ([0]).

Across the pooled UD data, we started with 82,884 sentences. From these, we automatically extracted 10,056 candidate LVC sentences using the

procedures described above. Importantly, manual verification was applied only to these automatically harvested candidates, not to the full dataset. Two annotators independently reviewed the candidates, and 565 items were identified as mislabeled (266 from the `compound:lvc`-based extraction and 299 from the compound-based extraction) and removed, leaving 82,319 sentences. After filtering, we obtained 9,491 LVC instances in total: 1,544 derived from `compound:lvc` arcs and 7,947 discovered via the noun–verb compound heuristic.

Across all prompting regimes, we cast the task as binary in-context classification: does a sentence contain an idiomatic/light verb construction (LVC) (`[1]`) or a literal verb–argument configuration (`[0]`)? To evaluate all models under controlled lexical conditions, we built a bespoke diagnostic dataset of 147 manually authored sentences, split into three balanced conditions (49 each): (i) LVC, (ii) NLVC (literal controls sharing the same target verbs as the LVC items), and (iii) RANDOM (unrelated negative controls). Items were validated by two annotators for naturalness, plausibility, and label correctness. Disagreements were resolved through discussion, and only agreed-upon items were retained in the final dataset. Evaluation items are held out from the one-/few-shot in-context exemplars, and the dataset was created specifically for this study. Annotators guarantee label correctness for the evaluation set, but the BERTurk baseline is trained on UD-derived proxy labels, where negatives may include unannotated LVCs because UD is not necessarily exhaustive.

Although $N = 147$ is modest, the dataset is intentionally constructed as a controlled diagnostic set with matched lexical controls. Prior work on behavioral test suites, contrast sets, and dynamic/challenge-style evaluations suggests that small, carefully-designed testbeds can be highly informative for revealing systematic decision-boundary failures that standard i.i.d. test accuracy can mask (Ribeiro et al., 2020; Gardner et al., 2020; Kiela et al., 2021; Yang et al., 2022; Zhao et al., 2024; He et al., 2025; Mayne et al., 2025; Karakaş and Şimşek, 2026). Accordingly, we interpret results primarily as evidence about regime-dependent behavior and calibration under controlled contrasts, rather than as an estimate of broad in-the-wild Turkish LVC detection accuracy. Figure 1 presents an overview of the experimental pipeline used in this study.

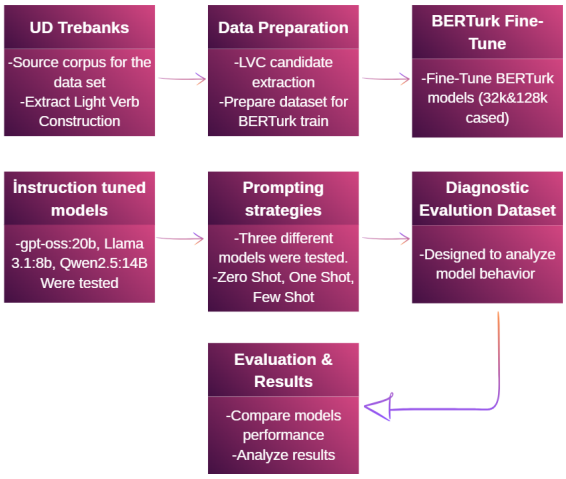

Figure 1: Flowchart of the experimental process

5 Models

We fine-tune BERTurk 32K cased and BERTurk 128K cased (Schweter, 2020) by adding a task-specific binary classification head over the final-layer `[CLS]` representation. We split the data 80/20 into train/test with stratified sampling and set hidden and attention dropout to 0.2 to reduce overfitting. Models are trained with learning rate 2×10^{-5}, batch size 32, weight decay 0.01, for up to 10 epochs, using early stopping on validation loss (patience 3) and selecting the checkpoint with the lowest validation loss. We report Accuracy, F1, and Loss. BERTurk 32K achieves 93.94% Accuracy, 0.7558 F1, and 0.1500 Loss; BERTurk 128K achieves 94.06% Accuracy, 0.7508 F1, and 0.1506 Loss, indicating comparable performance across vocabulary sizes.

All LLM experiments were run locally via Ollama. We evaluated instruction-tuned models using the Ollama tags `llama3.1:8b`, based on the Llama 3.1 model family (Grattafiori et al., 2024); `gpt-oss:20b`, based on OpenAI's gpt-oss model family (OpenAI, 2025); and `qwen2.5:14b`, based on the Qwen2.5 model family (Yang et al., 2025). The corresponding local Ollama configurations were: `llama3.1:8b` (architecture `llama`, 8.0B parameters, quantization `Q4_K_M`, context length 131,072), `gpt-oss:20b` (architecture `gptoss`, 20.9B parameters, quantization `MXFP4`, context length 131,072), and `qwen2.5:14b` (architecture `qwen2`, 14.8B parameters, quantization `Q4_K_M`, context length 32,768). The task is formulated as binary classification with labels `[1]` (contains an LVC) and `[0]` (no LVC). To improve output consistency, we set `temperature`=0.1 and left other decoding parameters at their Ollama defaults;

each sentence was processed individually, with no batching. Although generation length was not explicitly capped, prompts required outputting only a single label (`[0]` or `[1]`). For scoring, outputs were parsed automatically and counted as correct if and only if a valid label was produced; outputs lacking `[0]`/`[1]` were treated as invalid and manually inspected ($N = 2$).

6 Experiments

6.1 Experiment 1

Experiment 1 evaluates zero-shot prompting for instruction-tuned LLMs and, for comparison, supervised BERTurk classifier baselines. The task is binary LVC detection under three conditions with $n = 49$ items each: Random (negative; unrelated), NLVC (negative; literal in-domain), and LVC (positive; idiomatic/light-verb predicates under our labeling policy). Success rate is computed as correct/49 for each condition; Overall pools all 147 items.

Experiment 1 reveals a strong zero-shot asymmetry: LLMs perform well on negative controls (Random/NLVC) but largely miss LVC positives, consistent with a conservative bias toward [0]. By contrast, supervised BERTurk remains strong on both negatives and LVCs, mitigating the false-negative collapse (Table 1).

Model	Random	NLVC	LVC	Overall
Llama3.1-8B	0.980	0.959	0.000	0.646
GPT-OSS-20B	0.939	1.000	0.061	0.667
Qwen2.5-14B	0.918	0.857	0.122	0.633
BERTurk-32k (clf)	0.980	0.816	0.673	0.823
BERTurk-128k (clf)	0.980	0.816	0.796	0.864

Table 1: Experiment 1 success rates (0–1). LLM rows are zero-shot prompting; BERTurk rows are supervised classifier baselines (clf). Each condition has $n = 49$ items; Overall pools 147 items.

We tested Model×Correctness with chi-square tests per split (reporting Cramer's V). Among zero-shot LLMs, differences are not reliable on Random ($\chi^2(2) = 2.36$, $p = 0.307$, $V = 0.13$) but emerge on NLVC ($\chi^2(2) = 7.99$, $p = 0.018$, $V = 0.23$) and LVC ($\chi^2(2) = 9.89$, $p = 0.007$, $V = 0.26$). Including BERTurk yields a very large model effect on LVC ($\chi^2(4) = 123.83$, $p = 8.13{\times}10^{-26}$, $V = 0.71$). Notably, pooling conditions can hide this failure mode: Overall is not significant across LLMs ($\chi^2(2) = 0.38$, $p = 0.828$) but is across all models ($\chi^2(4) = 34.81$, $p = 5.07{\times}10^{-7}$).

In Experiment 1, the zero-shot LLMs are extremely conservative: they perform very well on the negative conditions (Random/NLVC) but collapse on the positive LVC condition (especially Llama3.1-8B at 0.000), which suggests a strong bias toward predicting the negative label when no demonstrations are provided. This may additionally reflect limited exposure to Turkish-specific linguistic patterns, as Llama 3.1 is not explicitly trained or optimized for Turkish. This creates a misleadingly "decent" Overall accuracy despite near-total failure on the phenomenon of interest (LVCs). By contrast, the supervised BERTurk classifier baselines maintain high LVC success (0.673–0.796) while staying strong on negatives, indicating that task-specific Turkish supervision captures the light-verb/idiom signal that zero-shot prompting does not. Statistically, model differences are driven primarily by the LVC condition (large effect when including BERTurk), while Overall differences among the three LLMs alone are not significant, which again shows why condition-wise reporting matters.

Although BERTurk has far fewer parameters than the instruction-tuned LLMs we test, it is evaluated here with a task-specific classifier head trained with labeled supervision derived from Turkish treebank resources. This kind of supervision may make certain morphosyntactic regularities more directly exploitable for a metalinguistic decision boundary (LVC vs. literal), even when overall model capacity is smaller. More generally, prior work on pretrained encoders suggests that additional supervised training—either as intermediate-task fine-tuning or task-specific fine-tuning—can sometimes improve downstream generalization and make particular linguistic distinctions more accessible to a simple classifier (Phang et al., 2018; Pruksachatkun et al., 2020). Related analyses also indicate that contextual encoders can encode multi-level linguistic information that downstream heads can leverage (Tenney et al., 2019). Since treebank annotation frameworks (e.g., Universal Dependencies) explicitly target predicate–argument structure and may include dedicated analyses for complex predicates (including light-verb-like patterns in some languages) (Nivre et al., 2016, 2020), supervised exposure to such representations might partially account for why a smaller encoder+head baseline can appear comparatively strong on this metalinguistic classification task. However, because the supervision signal is not identical to our labeling policy,

and because training data and evaluation items can differ in domain and lexical coverage, we treat this explanation as suggestive rather than causal, which requires further research.

6.2 Experiment 2: One-per-Class Prompting

In Experiment 2, we use one labeled example for LVC and NLVC per target verb template in-context demonstration along with instruction, and then require the model to output only a binary label for each test sentence in our evaluation corpus.

The one-shot prompt yields sharply different error profiles across model families (Table 2). Llama 3.1 8B attains high accuracy on LVC items but performs poorly on both negative splits, consistent with over-predicting the positive label. Qwen 2.5 shows the opposite bias: near-ceiling performance on negatives but substantially lower accuracy on LVC positives. GPT-OSS-20B is comparatively more balanced across conditions.

Model	LVC	NLVC	Random	Overall
GPT-OSS-20B	0.837	0.735	0.898	0.823
Llama 3.1 8B	0.878	0.286	0.469	0.544
Qwen 2.5 14B	0.490	1.000	0.959	0.816

Table 2: Experiment 2 success rates. Each split has $n = 49$ items; Overall pools $N = 147$ items.

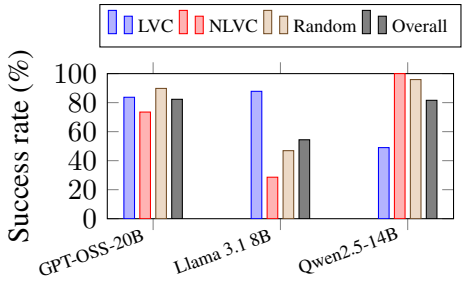

Figure 2: Experiment 2 (one-shot) success rates by condition and overall (percent).

To test whether models differ reliably within each split, we apply chi-square tests of independence (Model $\times$ Correctness) and use a Holm correction across splits. All splits show significant model differences after correction (LVC: $\chi^2(2) = 22.825$, $p_{Holm} = 1.106 \times 10^{-5}$; NLVC: $\chi^2(2) = 58.095$, $p_{Holm} = 9.704 \times 10^{-13}$; Random: $\chi^2(2) = 40.091$, $p_{Holm} = 5.909 \times 10^{-9}$; Overall: $\chi^2(2) = 37.574$, $p_{Holm} = 1.386 \times 10^{-8}$).

Post-hoc two-proportion tests (Holm-corrected within split) indicate that on LVC items, GPT-OSS-20B and Llama 3.1 8B outperform Qwen 2.5, whereas on negative splits, Qwen 2.5 and GPT-OSS-20B outperform Llama 3.1 8B; overall, GPT-OSS-20B and Qwen 2.5 are statistically indistinguishable and both outperform Llama 3.1 8B.

In general, the one-shot regime changes models' decision thresholds rather than uniformly improving "LVC understanding," producing clear bias trade-offs across families. Llama 3.1 8B appears to over-predict the positive label, which boosts LVC hit rate but collapses performance on both negative splits, while Qwen 2.5 shows the opposite pattern—near-ceiling negatives but substantially weaker LVC detection, suggesting a conservative, negative-skewed classifier. GPT-OSS-20B is the most balanced under this prompt, maintaining strong negative performance while still achieving relatively high LVC accuracy. Overall, the results imply that a single demonstration can induce strong, model-specific calibration shifts, so one-shot prompting is not reliably "better" without checking per-condition error profiles.

6.3 Experiment 3: Few-shot prompting

Experiment 3 evaluates the same three instruction-tuned LLMs under a few-shot prompt that provides multiple in-context demonstrations (one positive and one negative for several verbs) and constrains outputs to a binary label (`[1]` for LVC, `[0]` otherwise).

Across models, the few-shot prompt yields broadly high overall accuracy for GPT-OSS-20B and Qwen 2.5 (84–86%), while Llama 3.1 8B remains lower overall due to substantially weaker performance on the negative conditions despite strong LVC accuracy. Model differences are robust on Random and NLVC (and overall) under chi-square tests with Holm correction, but the omnibus cross-model difference on LVC items is weaker after correction, suggesting more similar positive-class performance under this prompt. The error profiles remain asymmetric: Qwen 2.5 is relatively conservative (very strong negatives but more missed LVCs), whereas Llama 3.1 8B is comparatively liberal (high LVC hit-rate but many false positives); GPT-OSS-20B is the most balanced of the three.

Model	Random	NLVC	LVC	Overall
GPT-OSS-20B	91.8	75.5	85.7	84.4
Llama 3.1 8B	51.0	61.2	87.8	66.7
Qwen 2.5 14B	87.8	98.0	71.4	85.7

Table 3: Experiment 3 success rates (%). Each condition has $n = 49$ items; Overall pools $N = 147$.

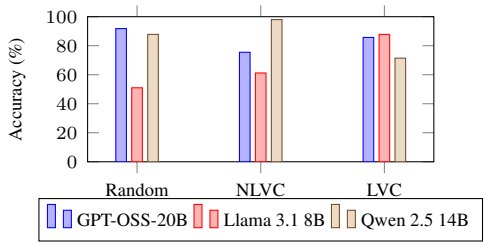

Figure 3: Experiment 3 accuracy by condition (few-shot prompting).

We tested cross-model differences per split using chi-square independence tests (Model×Correctness) with Holm correction over Random/NLVC/LVC/Overall. Effects are significant for Random ($\chi^2(2) = 27.85$, $p_{Holm} = 4.0{\times}10^{-6}$, $V = 0.44$), NLVC ($\chi^2(2) = 19.73$, $p_{Holm} = 1.4{\times}10^{-4}$, $V = 0.37$), and Overall ($\chi^2(2) = 19.95$, $p_{Holm} = 1.4{\times}10^{-4}$, $V = 0.21$), but not for LVC after correction ($\chi^2(2) = 5.17$, $p_{Holm} = 0.075$, $V = 0.19$). Post-hoc two-proportion tests (Holm-corrected within split) show that on Random, GPT-OSS-20B and Qwen 2.5 outperform Llama 3.1 8B, while GPT-OSS-20B and Qwen 2.5 do not differ; on NLVC, Qwen 2.5 outperforms both GPT-OSS-20B and Llama 3.1 8B, with no reliable GPT-OSS-20B vs. Llama difference; and on LVC, no pairwise differences survive correction. Finally, condition effects within models are significant for Llama 3.1 8B ($\chi^2(2) = 15.86$, $p_{Holm} = 0.0011$) and Qwen 2.5 ($\chi^2(2) = 14.33$, $p_{Holm} = 0.0015$), but not for GPT-OSS-20B ($\chi^2(2) = 5.05$, $p = 0.080$), consistent with GPT-OSS-20B's more even profile.

In general, experiment 3 (few-shot) largely eliminates the zero-shot "always-negative" failure mode, yielding high overall accuracy for GPT-OSS-20B and Qwen 2.5 (84–86 per cent), while Llama 3.1 8B remains substantially lower overall because it still over-predicts the positive label on the negative splits. The remaining differences are driven mainly by negative conditions: Qwen 2.5 is near-ceiling on NLVC/Random but misses more LVCs, whereas Llama shows the opposite bias (high LVC hit-rate, many false positives). GPT-OSS-20B is the most balanced profile across splits, and statistically, cross-model differences are robust for Random/NLVC (and overall) but noticeably weaker for LVC, suggesting that under a sufficiently rich prompt, models converge more on the positive class than on calibrating negatives.

7 General Discussion

Table 4 summarizes how prompt regime modulates Turkish LVC detection. Across LLMs, zero-shot performance is dominated by missed positives (high FN), while adding demonstrations mainly shifts the decision boundary and can flip the dominant error type (e.g., Llama becomes overly positive in one-shot). The few-shot prompt generally improves calibration by exposing broader positive/negative variability, reducing extreme one-shot failure modes while preserving high LVC hit-rates. For context, supervised BERTurk (classifier head) is more balanced in Experiment 1, though it is not directly comparable to prompted LLMs due to task-specific training.

Consistent with this, zero-shot cross-model differences are limited: after Holm correction, only NLVC shows a reliable Model×Correctness association ($\chi^2(2) = 6.45$, $p_{Holm} = 0.040$, $V = 0.26$), while Random and Overall do not ($p_{Holm} \geq 0.40$) and the LVC omnibus is not reliable ($\chi^2(2) = 6.39$, $p_{Holm} = 0.123$). In contrast, one-shot yields robust cross-model differences on every split (all $p_{Holm} \leq 1.11{\times}10^{-5}$). Under the few-shot prompt, cross-model differences remain reliable for Random ($\chi^2(2) = 27.85$, $p_{Holm} = 3.58{\times}10^{-6}$, $V = 0.44$), NLVC ($\chi^2(2) = 19.73$, $p_{Holm} = 1.40{\times}10^{-4}$, $V = 0.37$), and Overall ($\chi^2(2) = 19.95$, $p_{Holm} = 1.40{\times}10^{-4}$, $V = 0.21$), but not for LVC after correction ($\chi^2(2) = 5.17$, $p_{Holm} = 0.075$), consistent with partial convergence on positives under the prompt.

Within each model, prompt regime has a strong effect on LVC decisions. Regime×Correctness tests across the three prompting regimes show large and reliable regime effects on LVC for GPT-OSS-20B ($\chi^2(2) = 68.96$, $p_{Holm} = 4.25 \times 10^{-15}$, $V = 0.60$) and Qwen 2.5 ($\chi^2(2) = 35.48$, $p_{Holm} = 7.90 \times 10^{-8}$, $V = 0.42$). For Llama 3.1 8B, LVC accuracy jumps from zero-shot to prompted regimes (near-ceiling recall thereafter), but this co-occurs with a large rise in false positives on negatives (Table 4). Taken together, these regime-dependent FP/FN trade-offs are compatible with the view that in-context demonstrations primarily re-weight a model's decision boundary and calibration, that is, they change how aggressively the model predicts the positive class, rather than uniformly improving performance across splits (Zhao et al., 2021; Min et al., 2022; Liu et al., 2023; Akyürek et al., 2024). In this sense, few-shot

Model	Regime	Random	NLVC	LVC	Overall	FP	FN	Prec	Rec
GPT-OSS-20B									
	Zero-shot	93.9	100.0	6.1	66.7	3	46	50.0	6.1
	One-shot	89.8	73.5	83.7	82.3	18	8	69.5	83.7
	Few-shot (master)	91.8	75.5	85.7	84.4	16	7	72.4	85.7
Qwen 2.5 14B									
	Zero-shot	91.8	85.7	12.2	63.3	11	43	35.3	12.2
	One-shot	95.9	100.0	49.0	81.6	2	25	92.3	49.0
	Few-shot (master)	87.8	98.0	71.4	85.7	7	14	83.3	71.4
Llama 3.1 8B									
	Zero-shot	98.0	95.9	0.0	64.6	3	49	0.0	0.0
	One-shot	46.9	28.6	87.8	54.4	61	6	41.3	87.8
	Few-shot (master)	51.0	61.2	87.8	66.7	43	6	50.0	87.8
Supervised baselines									
BERTurk-32k (clf)	Supervised	98.0	81.6	67.3	82.3	10	16	76.7	67.3
BERTurk-128k (clf)	Supervised	98.0	81.6	79.6	86.4	10	10	79.6	79.6

Table 4: Prompt regime comparison across Experiments 1–3. Entries are success rates (%) for each condition (each has $n = 49$) and Overall ($N = 147$). FP/FN are pooled error counts where FP counts mistakes on negatives (Random+NLVC) and FN counts mistakes on LVC. Precision/Recall treat LVC as the positive class and pool Random+NLVC as negatives.

prompting appears to elicit metalinguistic judgments about LVC vs. literal usage, while the large regime effects also highlight substantial prompt sensitivity and contextual/label bias documented in prior work on in-context learning (Zhao et al., 2021; Min et al., 2022).

At first sight, the stronger performance of the fine-tuned BERT model over several larger recent LLMs may seem surprising. We do not interpret this result as evidence that BERT is generally more capable than state-of-the-art LLMs. However, we think that it reflects the difference between in-domain supervised sequence labeling and general-purpose prompting. Idiomatic light verb identification is a fixed, annotation-scheme-sensitive token classification problem: the model must decide whether a verb participates in a construction such as a noun plus *et-* 'do', *ol-* 'be/become', or *yap-* 'do/make' under the conventions of the annotated dataset. A fine-tuned encoder is directly optimized for this label space and can exploit recurring morphosyntactic, lexical, and positional cues. Prompted LLMs, by contrast, must infer the task definition from an instruction and map open-ended linguistic knowledge onto a discrete annotation decision without task-specific parameter updates. This exposes them to additional sources of error, including output-format instability, boundary mismatches, inconsistent treatment of borderline idiomaticity, and sensitivity to prompt wording. This interpretation is consistent with broader findings that task-specific supervision or parameter-efficient adaptation can match or outperform zero-/few-shot prompting and in-context learning, especially in narrow classification settings (Mosbach et al., 2023; Liu et al., 2022; Edwards and Camacho-Collados, 2024; Bucher and Martini, 2024; Pecher et al., 2024). Recent work also shows that prompted and in-context learning approaches remain sensitive to demonstration selection, number, order, and even the position of examples within the prompt (Schoch and Ji, 2025; Gao et al., 2025; Cobbina and Zhou, 2025). Thus, the relevant contrast in our experiments is not "BERT versus LLMs" in general, but supervised in-domain token-level classification versus prompted general-purpose inference. For a fine-grained morphosyntactic and idiomaticity-sensitive annotation task, direct supervision over the target label space can be more valuable than broad instruction-following ability.

BERTurk uses task-specific supervised training (a classifier head with labeled supervision), whereas the LLMs are evaluated via prompting without gradient updates; as a result, direct comparison may not be reliable. Still, the results suggest two descriptive trends: (i) relative to GPT-OSS-20B in zero-shot, BERTurk-128k achieves much higher LVC accuracy (0.796 vs. 0.061; two-proportion $z = 7.35$, $p = 2.0 \times 10^{-13}$) and higher Overall accuracy (0.864 vs. 0.667; $z = 3.99$, $p = 6.6 \times 10^{-5}$); (ii) once GPT-OSS-20B is provided demonstrations (one-/few-shot), its Overall performance becomes closer to BERTurk-128k (e.g., 0.844 vs. 0.864 in Experiment 3; $p = 0.62$), and it may even slightly exceed BERTurk on LVC positives (0.857 vs. 0.796; $p = 0.42$), though these differences are not statistically reliable at our sample size. One possible interpretation is that explicit supervised exposure to Turkish morphosyntax and annotation conventions may help stabilize metalinguistic judgments, while prompted LLMs can approximate this behavior when the prompt supplies

sufficient labeled structure; prior work suggests supervised objectives can sharpen linguistically relevant representations, but the mapping from such supervision to metalinguistic competence is not guaranteed and likely task-dependent (Tenney et al., 2019; Hewitt and Manning, 2019; Rogers et al., 2020).

8 Conclusion

This paper examined Turkish LVC detection through a literal–idiomatic contrast that blocks trivial verb-only heuristics and isolates whether models treat verb–nominal predicates as unitary multiword meanings. Across three prompting regimes, we find that instruction-tuned LLMs are conservative in zero-shot (high negative accuracy, near-collapse on LVC recall), but can be rapidly re-calibrated with minimal demonstrations, albeit with distinct family-specific biases in one-shot. A few-shot prompt generally stabilizes behavior and yields strong overall performance, with GPT-OSS-20B remaining the most balanced across conditions and Qwen 2.5 excelling on negatives. At the same time, the supervised BERTurk classifier provides a strong Turkish baseline that is competitive overall, suggesting that task-specific supervision may still offer advantages for metalinguistic decisions in morphologically rich languages. Thus, the findings appear to motivate treating LVC detection as a prompt-sensitive capability and show the value of controlled, condition-wise evaluation beyond aggregate accuracy.

Limitations

Our evaluation focuses on a targeted set of Turkish verb–nominal predicates (Random/NLVC/LVC; $N = 147$), so the conclusions may not fully generalize to other Turkish MWE families (e.g., fixed idioms, postpositional MWEs) or to broader domains beyond the curated test set. In addition, prompted LLM performance is sensitive to prompt design and demonstration choice, so reported one-/few-shot results should be interpreted as evidence about *prompt-regime effects* rather than model-intrinsic competence. From a mechanistic-interpretability perspective, prompts can be viewed as causal interventions on internal computation, and linking prompt-induced behavior shifts to stable, abstracted mechanisms remains an open challenge (Andreas, 2022; Bereska and Gavves, 2024; Geiger et al., 2025; Holtzman and Tan, 2025). Finally, BERTurk uses supervised training with a classifier head, whereas LLMs are evaluated via in-context prompting; we therefore treat the cross-family comparison as suggestive rather than strictly like-for-like.

References

a. Ud Turkish atis. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

b. Ud Turkish boun. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

c. Ud Turkish framenet. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

d. Ud Turkish gb. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

e. Ud Turkish imst. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

f. Ud Turkish kenet. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

g. Ud Turkish penn. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

h. Ud Turkish pud. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

i. Ud Turkish tourism. Universal Dependencies treebank hub page. Accessed: 2025-12-16.

j. Universal dependencies: compound. Universal Dependencies relations documentation. Accessed: 2025-12-16.

k. Universal dependencies: compound:lvc. Universal Dependencies relations documentation. Accessed: 2025-12-16.

Ekin Akyürek, Boxin Wang, Yoon Kim, and Jacob Andreas. 2024. In-context language learning: Architectures and algorithms. *Preprint*, arXiv:2401.12973.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779.

Inbal Arnon and Eve V. Clark. 2011. Why brush your teeth is better than teeth: Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, 7:107–129.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, 2 edition. Chapman and Hall/CRC.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*, pages 99–105.

A. Barman, D. Saha, and A. R. Pal. 2024. An approach for maintaining structural uniformity of multiword expressions. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. IEEE.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, and Savaş Yıldırım. 2025. Tokenization standards and evaluation in natural language processing: A comparative analysis of large language models on Turkish. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, Istanbul, Turkey. IEEE. Conference held June 25–28, 2025. Preprint available as arXiv:2508.13058.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, Savaş Yıldırım, and Demircan Çelik. 2026. Tokens with meaning: A hybrid tokenization approach for Turkish. *Preprint*, arXiv:2508.14292.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for AI safety: A review. *arXiv preprint arXiv:2404.14082*.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018. Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 248–253, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned "Small" LLMs still significantly outperform zero-shot generative AI models in text classification. *Preprint*, arXiv:2406.08660.

Miriam Butt. 2010. The light verb jungle: still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex Predicates: Cross-linguistic Perspectives on Event Structure*. Cambridge University Press.

Elena Castroviejo, Marta Ponciano, José V. Hernández-Conde, and Agustín Vicente. 2024. Development of nonliteral interpretations in typically developing spanish-speaking children: Light verb constructions and figurative expressions. *Studia Linguistica*, 78(1):8–34.

Lucy (Yu-Chuan) Chiang. 2025. *Acquisition of the Syntax Semantics of Light Verbs in Mandarin Bilinguals*. Phd dissertation, University of Michigan. Dissertation defended April 24, 2025.

Kwesi Adu Cobbina and Tianyi Zhou. 2025. Where to show demos in your prompt: A positional bias of in-context learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29560–29593, Suzhou, China. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Gülşen Eryiğit, Kübra Adali, Dilara Torunoğlu-Selamet, Umut Sulubacak, and Tuğba Pamay. 2015. Annotation and extraction of multiword expressions in Turkish treebanks. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 70–76, Denver, Colorado. Association for Computational Linguistics.

Gülşen Eryiğit, Tugay İlbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 45–55, Dublin, Ireland. Association for Computational Linguistics.

Xiang Gao, Ankita Sinha, and Kamalika Das. 2025. Learning to search effective example sequences for in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6152–6161, Albuquerque, New Mexico. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 6 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jane Grimshaw and Armin Mester. 1988. Light verbs and theta-marking. *Linguistic Inquiry*, 19(2):205–232.

Angela X. He and Eva Wittenberg. 2020. The acquisition of event nominals and light verb constructions. *Language and Linguistics Compass*, 14(2):e12363.

Linyang He, Qiaolin Wang, Xilin Jiang, and Nima Mesgarani. 2025. Layer-wise minimal pair probing reveals contextual grammatical-conceptual hierarchy in speech representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35338–35353, Suzhou, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ari Holtzman and Chenhao Tan. 2025. Prompting as scientific inquiry. *arXiv preprint arXiv:2507.00163*.

Sercan Karakaş and Yusuf Şimşek. 2026. From lemmas to dependencies: What signals drive light verbs classification? In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 220–227, Rabat, Morocco. Association for Computational Linguistics.

Sercan Karakaş and Yusuf Şimşek. 2026. Benchmarking source-sensitive reasoning in Turkish: Humans and LLMs under evidential trust manipulation. *Preprint*, arXiv:2604.24665.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*.

Harry Mayne, Ryan Othniel Kearns, Yushi Yang, Andrew M. Bean, Eoin D. Delaney, Chris Russell, and Adam Mahdi. 2025. LLMs don't know their own decision boundaries: The unreliability of self-generated counterfactual explanations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24161–24186, Suzhou, China. Association for Computational Linguistics.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Verginica Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic, and Ivelina Stoyanova. 2025. Survey on lexical resources focused on multiword expressions for the purposes of NLP. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 41–57, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*.

Jan Odijk. 2013. Identification and lexical representation of multiword expressions. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, pages 233–245. Springer, Berlin, Heidelberg.

Kemal Oflazer. 1993. Two-level description of Turkish morphology. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Kemal Oflazer, Özlem Çetinoğlu, and Bilge Say. 2004. Integrating morphology with multi-word expression processing in Turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Aydın Özbek. 2010. On çek- as a light verb: A contrastive view from japanese. *Journal of Language and Linguistic Studies*, 6(1):1–13.

Duygu Özge, Gülten Ünal, and İsa Kerem Bayırlı. 2022. Assigning meaning to light verbs in Turkish. *Dilbilim Araştırmaları Dergisi*, 33(1):1–27.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. Fine-tuning, prompting, in-context learning and instruction-tuning: How many labelled samples do we need? *Preprint*, arXiv:2402.12819.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplemental training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Damith Premasiri and Tharindu Ranasinghe. 2022. BERT(s) to detect multiword expressions. *arXiv preprint arXiv:2208.07832*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? *arXiv preprint arXiv:2005.00628*.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing (CICLing)*.

Yağmur Sağ. 2015. Complex predicate formation via adjunction to a head category: Evidence from light verb constructions in Turkish. Manuscript, Rutgers University.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Stephanie Schoch and Yangfeng Ji. 2025. Monte Carlo sampling for analyzing in-context examples. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 63–78, Albuquerque, New Mexico. Association for Computational Linguistics.

Johannes M. Schulz. 2024. *Multi-word-constructions and linguistic development in early foreign language classrooms: the role of input variability*. Doctoral dissertation, University of Oxford.

Stefan Schweter. 2020. Berturk – bert models for turkish. Zenodo.

Anna Siyanova-Chanturia, Kathy Conklin, and Walter J. B. van Heuven. 2011. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776–784.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Aygül Uçar. 2010. Light verb constructions in Turkish dictionaries: Are they submeanings of polysemous verbs? *Dil ve Edebiyat Dergisi / Journal of Linguistics and Literature*, 7(1):1–17.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Guanqun Yang, Mirazul Haque, Qiaochu Song, Wei Yang, and Xueqing Liu. 2022. TestAug: A framework for augmenting capability-based NLP tests. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3480–3495, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Raoyuan Zhao, Abdullatif Köksal, Yihong Liu, Leonie Weissweiler, Anna Korhonen, and Hinrich Schuetze. 2024. SynthEval: Hybrid behavioral testing of NLP models with synthetic CheckLists. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7017–7034, Miami, Florida, USA. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Yağmur Öztürk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the PARSEME Turkish corpus of verbal multiword expressions. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 100–104, Marseille, France. European Language Resources Association.

A Prompt Templates

A.1 Zero-shot

Zero-shot Görev: Aşağıdaki cümlede “light verb construction (LVC)” var mı?
Cümle: “sentence”
Tanım:

[1] LVC varsa: Yüklemin çekirdek anlamını bir isim taşır; fiil daha “hafif/yardımcı” rol oynar ve yapı tek bir yüklem gibi çalışır. Not: Bu deneyde kalıplaşmış/idiomatik isim+fiil yüklemlerini de [1] say.

[0] LVC yoksa: Fiil kendi temel (literal) anlamıyla kullanılıyordur (fiziksel eylem, nesne aktarma, gerçek görme/duyma vb.).

Eğer cümlede “light verb construction (LVC)” varsa sadece [1] yaz. Eğer cümlede “light verb construction (LVC)” yoksa sadece [0] yaz.
Başka açıklama yapma, sadece sonucu yaz.
Cevap:

A.2 One-shot

One-shot Görev: Aşağıdaki cümlede “light verb construction (LVC)” var mı?
Tanım:

[1] LVC varsa: Yüklemin çekirdek anlamını bir isim taşır; fiil daha “hafif/yardımcı” rol oynar ve yapı tek bir yüklem gibi çalışır. Not: Bu deneyde kalıplaşmış/idiomatik isim+fiil yüklemlerini de [1] say.

[0] LVC yoksa: Fiil kendi temel (literal) anlamıyla kullanılıyordur (fiziksel eylem, nesne aktarma, gerçek görme/duyma vb.).

Örnekler (fiile göre):
[VERB-*DUY*]
Cümle: “Ona büyük saygı duydu.” Cevap: [1]
Cümle: “Koridordan gelen sesi duydu.” Cevap: [0]

Şimdi sınıflandır:
Cümle: “sentence”
Yalnızca [0] ya da [1] yaz. Başka hiçbir şey yazma.
Cevap:

A.3 Master few-shot (example; abbreviated)

Master few-shot (example; abbreviated) Görev: Aşağıdaki cümlede “light verb construction (LVC)” var mı?
Tanım:

[1] LVC varsa: Yüklemin çekirdek anlamını bir isim taşır; fiil daha “hafif/yardımcı” rol oynar ve yapı tek bir yüklem gibi çalışır. Not: Bu deneyde kalıplaşmış/idiomatik isim+fiil yüklemlerini de [1] say.

[0] LVC yoksa: Fiil kendi temel (literal) anlamıyla kullanılıyordur (fiziksel eylem, nesne aktarma, gerçek görme/duyma vb.).

Örnekler (kısaltılmış; temsili örnek):
[*DUY-*]
Cümle: “Hayatında ilk defa birine güven duydu.” Cevap: [1]
Cümle: “Koridordan gelen sesleri duydu.” Cevap: [0]

[*VER-*]
Cümle: “Soruma hemen yanıt verdi.” Cevap: [1]
Cümle: “Öğretmen öğrencilere kâğıt verdi.” Cevap: [0]

[*GİR*-]
Cümle: “Sınav başlayınca paniğe girdi.” Cevap: [1]
Cümle: “Odaya hızlıca girdi.” Cevap: [0]

... (diğer fiiller ve örnek çiftleri aynı formatta devam eder)

Şimdi sınıflandır:
Cümle: “sentence”
Yalnızca [0] ya da [1] yaz. Başka hiçbir şey yazma.
Cevap:

B English Translations (instructions only)

B.1 Zero-shot (English)

Task: Does the sentence below contain a “light verb construction (LVC)”?
Definition:

[1] If there is an LVC: the core predicational meaning is carried by a noun; the verb plays a “light/auxiliary” role, and the whole expression behaves as a single predicate. Note: In this experiment, treat lexicalized/idiomatic noun+verb predicates as [1].

[0] If there is no LVC: the verb is used with its basic (literal) meaning (e.g., a physical action, transfer of an object, literal seeing/hearing, etc.).

Sentence: “sentence”
If the sentence contains an LVC, write only [1]. If the sentence does not contain an LVC, write only [0].
Do not provide any explanation; output only the label.
Answer:

B.2 One-shot

One-shot Task: Does the following sentence contain a “light verb construction (LVC)”?
Definition:

[1] If there is an LVC: A noun carries the core meaning of the predicate; the verb plays a more “light/auxiliary” role, and the construction functions as a single predicate. Note: In this experiment, also count conventionalized/idiomatic noun+verb predicates as [1].

[0] If there is no LVC: The verb is used with its basic literal meaning (physical action, object transfer, actual seeing/hearing, etc.).

Examples by verb:
[VERB-*DUY*]
Sentence: “He felt great respect for her/him.” Answer: [1]
Sentence: “He heard the sound coming from the corridor.” Answer: [0]

Now classify:
Sentence: “sentence”
Write only [0] or [1]. Do not write anything else.
Answer:

B.3 Master few-shot (example; abbreviated)

Master few-shot (example; abbreviated) Task: Does the following sentence contain a “light verb construction (LVC)”?
Definition:

[1] If there is an LVC: A noun carries the core meaning of the predicate; the verb plays a more "light/auxiliary" role, and the construction functions as a single predicate. Note: In this experiment, also count conventionalized/idiomatic noun+verb predicates as [1].

[0] If there is no LVC: The verb is used with its basic literal meaning (physical action, object transfer, actual seeing/hearing, etc.).

Examples (abbreviated; representative example):
[*DUY-*]
Sentence: "For the first time in his/her life, he/she felt trust in someone." Answer: [1]
Sentence: "He/she heard the sounds coming from the corridor." Answer: [0]

[*VER-*]
Sentence: "He/she immediately gave an answer to my question." Answer: [1]
Sentence: "The teacher gave paper to the students." Answer: [0]

[*GIR-*]
Sentence: "When the exam started, he/she panicked." Answer: [1]
Sentence: "He/she quickly entered the room." Answer: [0]

... (other verbs and example pairs continue in the same format)

Now classify:
Sentence: "sentence"
Write only [0] or [1]. Do not write anything else.
Answer: