Evaluation of Large Language Models on the Kazakh Word-in-Context Task: Dataset Creation and Analysis

Dissertation Project

Author: Nurseiit Bakkali
Student ID: 24092543
Degree: MSc Artificial Intelligence
Supervisor: Mohammad Taher Pilehvar
School of Computer Science and Informatics
Cardiff University
Date: 6th September 2025

Abstract

This dissertation explores how modern large language models (LLMs) can differentiate word sense in the Kazakh language, a low-resourced and morphologically rich language. To study it, the first Kazakh Word-in-Context (WiC) dataset has been created and it has 510 instance pairs where the target word appears in two different sentences. The task is simple by itself but challenging in practice: deciding whether the word has the same meaning in both sentences.

Several recent LLMs have been evaluated including GPT-4o, GPT-3.5-turbo, Gemini 1.5 Pro&Flash and Llama3 using this dataset. All models tested under different conditions: zero-shot versus few-shot prompting and English versus Kazakh instructions. Along with these experiments, human evaluation with native speakers was conducted to establish a reliable performance baseline.

Experiment results showed three clear patterns. First all LLMs underperformed human annotators: while humans reached over 93% accuracy, the best model achieved about 77%. Second, few-shot prompting steadily improved model performance which means the example in the prompt was beneficial. Third, the effect of prompt language was not significant: English instructions gave small gains in most cases, but results were not consistent across models. Moreover, Error analysis revealed systematic weaknesses, particularly with Kazakh morphology and figurative language.

Overall contributions of this study are: (i) introduction of the first open Kazakh WiC dataset, (ii) a systematic evaluation of LLMs on Kazakh semantic understanding, and (iii) insights into prompting strategies and error patterns which can guide future research in this field.

In total, this research highlights both the progress and limitations of LLMs for Kazakh. LLMs remain far from human-level understanding. The results emphasise that there is need for more resources, better morphological modelling, and cultural training data to make sure that advancement in NLP includes all languages.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Mohammad Taher Pilehvar, for his guidance, feedback and encouragement throughout the course of this dissertation.

I am sincerely thankful to the annotators who contributed their time and effort to the dataset creation process. Their careful judgments made this research possible.

My appreciation also goes to my family and close friends who emotionally supported me through many challenges of this project. Their encouragement and belief was a constant source of motivation.

Finally, I would like to acknowledge the broader academic community whose open-source resource, tool, and datasets helped for this study. This work is builded by using their efforts and would not have been possible without the collective spirit of collaboration in the NLP field.

Contents

Chapter 1

Introduction

1.1 Motivation

In the last decade there has been remarkable progress in natural language processing (NLP) with significant development in large language models (LLMs) trained on a vast multilingual corpora. These models demonstrated impressive capabilities in various tasks ranging from summarisation to translation (Brown et al., 2020; OpenAI, 2023; Gemini Team, 2024). However, most of this progress is concentrated in high-resource languages like English, Chinese and Spanish. Low-resource and morphologically rich languages like Kazakh remain underrepresented in both training corpora and evaluation benchmarks(Joshi et al., 2020; Haddow et al., 2022).

Kazakh presents unique challenges for NLP due to its morphological richness, agglutinative structure and relatively free word order (Makhambetov et al., 2013). These characteristics often increase complexity of lexical semantics making the word sense disambiguation more difficult for automated systems.

The Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019) is an effective benchmark for evaluating models ability to distinguish different meanings of the same word based on context. While WiC is extensively studied in high-resource languages, there is no publicly available extension for evaluating Kazakh WiC performance, especially for modern LLMs.

This dissertation is motivated by the need to close that gap. This is because evaluating LLMs on the Kazakh WiC task is not only important for understanding their cross-lingual generalisation capabilities, but for knowing how these model behave practically in applications such machine translation, educational tools or AI assistants in Kazakh language.

1.2 Research Context

Understanding the meaning of a word in context is one of the challenges in NLP. The WiC task was introduced as a benchmark for this purpose (Pilehvar and Camacho-Collados, 2019). The WiC task offers a sharper probe of contextual semantics than earlier datasets such as Stanford Contextual Word Similarity (SCWS) (Huang et al., 2012). WiC has become part of broader benchmarks like SuperGLUE (Wang et al., 2019; *SuperGLUE Benchmark* 2019) and has been extended to multiple languages in XL-WiC (Raganato et al., 2020). However, as told before, it lacks Kazakh and leaves an important gap in the evaluation of semantic understanding.

In parallel, LLMs have shown new ways to solve tasks without fine-tuning them. With in-context learning, they can perform tasks with only instructions (zero-shot) or with a small number of examples (few-shot) (Brown et al., 2020; Min et al., 2022; Dong et al., 2024).

In addition, recent evaluations confirm that current LLMs perform significantly worse on Kazakh than on English (Maxutov et al., 2024; Togmanov et al., 2025). In addition, studies show that performance can vary depending on the language of the prompt, and English instructions often do better for non-English tasks (Zhao et al., 2023; Ahuja et al., 2023). Yet these effects remain underexplored for Kazakh.

This raises three central questions: How well can LLMs handle sense disambiguation in Kazakh? Does a few-shot prompting improve their performance in such a low-resource setting? Does prompting in English give them an advantage over Kazakh? This dissertation addresses these questions by (i) manually curating the first WiC-style Kazakh dataset, (ii) systematically evaluating modern LLMs under different prompting strategies, and (iii) comparing the results of the model with human performance.

1.3 Structure of the Dissertation

The whole dissertation is structured as follows. Chapter 2 defines the research aims and objectives. Chapter 3 thoroughly reviews previous work on WiC, LLMs, low-resource NLP challenges, prompting strategies, and human baselines identifying research gaps in this area. Chapter 4 formulates the research questions and hypotheses. Chapter 5 explains the methodology, dataset creation, preprocessing, model selection, prompting design, and human evaluation. Chapter 6 presents the results of the whole experiment in zero-shot, few-shot, and cross-lingual prompting configurations. Chapter 7 discusses the findings with particular attention to error patterns and implications for low-resource NLP. Chapter 8 concludes the dissertation by summarizing key findings, limitations, and directions for future work. Finally, Chapter 9 reflects the author's personal learning and research process.

Chapter 2

Aims and Objectives

2.1 Aims

The primary aim of this research is to evaluate how well the state-of-the-art large language models (LLMs) understand the meaning in the Kazakh language when words are used in different contexts. To achieve this, this study introduces the first Kazakh WiC dataset and uses it to evaluate several LLMs under different experimental conditions like combining zero-shot vs few-shot and prompting languages.

This aim addresses the broader challenge of assessing semantic understanding in low-resource, morphologically rich languages and explores how LLMs generalise beyond high-resource settings.

In total, this research aims not only to test models abilities but also to provide the first reliable baseline for future work in Kazakh word semantic tasks in NLP.

2.2 Objectives

To achieve this aim, the dissertation follows these objectives:

1. **Kazakh WiC dataset.** Manually create high-quality sentence pairs in Kazakh which follows the WiC format. Ensure balanced labels and reliable annotation.

2. **LLMs evaluation.** Measure the ability of GPT-4o, Gemini 1.5 pro, GPT-3.5-turbo and Llama3 models on this Dataset.

3. **Prompting strategies comparison.** Investigate how zero-shot (just instructions) and few-shot (with examples) affect the performance.

4. **Prompting language effects.** Assess whether English prompting leads to higher accuracy than Kazakh prompting in WiC task.

5. **Human baseline.** Collect and analyse human performance on Kazakh WiC data to establish human baseline for comparison of LLMs results.

6. **Analyse errors.** Identify common error types by analysing models outputs and to understand where models struggle.

7. **Open resources.** Release the created Kazakh WiC dataset and supporting code openly so that other researchers and practitioners can continue on this work.

All these objectives provide a clear pathway from building a new resource to evaluating models and giving insights for both research and real-world applications of Kazakh NLP.

Chapter 3

Literature Review

Overview

This chapter surveys the background research done on the field of evaluation LLMs on the Kazakh language on the WiC task. Purpose of the chapter consist of three main parts: (i) explaining WiC and motivation of taking this task as the study on contextual semantics; (ii) reviewing the performance of the LLMs on the task depending on the prompt type like a zero-shot and few-shot; (iii) showing that Kazakh language is within low-resource NLP range, highlighting the dataset and evaluation challenges. The whole review is done so that it directly bears research questions (RQ1-RQ3) and hypotheses (H1-H3) stated in Chapter 4.

During background search, recent, peer-reviewed, and widely used benchmarks (e.g., WiC, SuperGlue) sources were prioritized. All used materials are recent preprints and state-of-the-art solutions. Studies on low-resource language resources rich in morphologically were included when they were considered relevant. §3.1 introduces WiC; §3.2 discusses LLMs and semantic understanding; §3.3 outlines low-resource/Kazakh challenges; §3.4 compares zero-shot vs. few-shot prompting; §3.5 analyzes prompt language effects; §3.6 contrasts LLM and human performance; and §3.7 summarizes and identifies the research gap that motivates our study.

Search strategy and scope: I searched for research articles on ACL Anthology, arXiv, Google Scholar, and ACM Digital Library (2012–2025) using terms such as 'Word-in-Context', 'WiC', 'word sense disambiguation', 'in-context learning', 'Kazakh', and 'low-resource NLP'. The materials chosen had the following criteria: peer reviewed or widely cited; benchmarks/datasets; multilingual or low-resource relevance. I mostly excluded materials without evaluation relevance or work related to other fields of NLP. All sources are listed in the References.

3.1 Word-in-Context (WiC) Task

The Word-in-Context (WiC) task for the first time was introduced on the research done by Pilehvar and Camacho-Collados Pilehvar and Camacho-Collados (2019). The WiC task asks a simple question: does the same word convey the same meaning in two sentences? The answer is binary - same or different. By design, the task probes context-sensitive semantics rather than surface overlap. If a model truly tracks meaning, it should tell when a polysemous word shifts sense across uses. The task is studied by creating the data set where every entry has a target word and two sentences and the label is 'True' or 'False'. In addition to this, words were restricted to nouns and verbs (Pilehvar and Camacho-Collados, 2019).

WiC was proposed to address weaknesses in earlier semantic evaluations such as the Stanford Contextual Word Similarity (SCWS) dataset (Huang et al., 2012), which scores word pairs in context on a graded similarity scale. However, SCWS has a low inter-rater agreement (IRA) - pairwise Spearman $\rho \approx 0.35$ and mean human-others $\rho \approx 0.52$ - indicating limited annotation reliability and a weak human upper bound (Pilehvar and Camacho-Collados, 2019). Under these conditions, the benchmark struggles to separate strong models from weaker ones. Moreover, because many SCWS instances compare different target words across the two contexts, context-independent embeddings can perform surprisingly well; for example, Google News word2vec has been reported to achieve $\rho \approx 0.65$, exceeding the dataset's optimistic human upper bound (Pilehvar and Camacho-Collados, 2019). WiC solves these issues by replacing graded similarity with a expert-curated binary decision dataset where in both sentences used same word and word has meaning depending on the context, yielding a sharper test of contextual meaning (Pilehvar and Camacho-Collados, 2019).

The whole dataset was made semi-automatically by using sources WordNet, Verbnet and Wiktionary. The authors removed pairs with subtle sense distinction to reduce annotator confusion, then built balanced positive/negative pairs so that shallow heuristics are unhelpful. In this experiment, human annotators reached about 80% accuracy while the best performed embedding model BERT got 65.5% accuracy(Pilehvar and Camacho-Collados, 2019).

The WiC task is currently part of the SuperGLUE benchmark (Wang et al., 2019), which is a harder version of the GLUE benchmark. These benchmarks are helpful for clear comparisons across model families. (*SuperGLUE Benchmark* 2019). However, when tasks are extended multilingual evaluations there is a big difference compared to English, especially in low-resource settings (Hu et al., 2020). As the multilingual version of the WiC-style task, the XL-WiC dataset was created by Raganato et al. (2020) across 12 languages. Even here the experiment shows a challenge on word semantic tasks in other languages (Raganato et al., 2020).

Kazakh language is totally absent from some benchmarks and if they contain it they are lacking from word semantic evaluation dataset. The WiC has a clear appeal to Kazakh language and constructed Kazakh dataset with this style will work well for evaluating LLMs on understanding word sense in Kazakh. In this dissertation, the main English WiC was the motivation to create the Kazakh version and evaluating various LLMs on different prompt methods.

> *We frame the task as binary classification. Each instance in WiC has a target word w, either a verb or a noun, for which two contexts, c1 and c2, are provided. Each of these contexts triggers a specific meaning of w. The task is to identify if the occur- rences of w in c1 and c2 correspond to the same meaning or not. .... We found the average human accuracy on the dataset to be 80.0%*

(Pilehvar and Camacho-Collados, 2019).

3.2 Large language models and In-context learning

Recent development and release of Large Language Models (LLMs) represent a major shift in natural language processing. LLMs built on the transformer architecture, they are trained on vast and diverse corpora covering whole internet data. This training on a large amount of texts enables them to capture contextual dependencies compared to earlier embedding models such as BERT and other recurrent models. This scaling effect allows LLMs to act like general-purpose learners: rather than training on specific tasks, they can adapt to the new problems through language instruction only.

A central mechanism which explains this adaptability is in-context learning (ICL) which is first demonstrated convincingly with GPT-3 (Brown et al., 2020). In ICL, a model is not trained and not updated with gradients like in real training of models but instead LLMs provided with a short context window which contains examples on the proposed task or instructions. Depending on the are examples provided or not it has namings like zero-shot and few-shot. With this prompting methods the same pretrained model can be reproduced for a wide range of the tasks without fine-tuning. Efficiency of this approach was highlighted in the report shown by OpenAI on the GPT-4 LLM, where the model achieved high performance in evaluation-only setup across diverse benchmarks (OpenAI, 2023). However, Min et al. (2022) also shows that ICL is highly sensitive to prompt format and to examples demonstrated in the prompt. This motivates for careful prompt design strategies to properly evaluate models in this dissertation.

For this study, five representative models were selected. GPT-4o, one of the latest versions in the GPTs from OpenAI, provides state-of-the-art reasoning and good multilingual coverage. GPT-3.5-turbo is a widely used, cost-efficient model in API settings of OpenAI. Gemini 1.5 (Flash and Pro) has comparative performance with extended context length and multilingual claims. Finally, Llama-3, model from Meta, was run locally on macOS hardware which represents an open-weight alternative that can be evaluated without API. This choice of various types of models show both commercial and open-source ecosystems, balancing accessibility with capability.

In the case of Kazakh, recent research shows that performance remains well below compared to English or Russian. Studies such as Maxutov et al. (2024) and the KazMMLU benchmark (Togmanov et al., 2025) report noticeable accuracy gaps for current LLMs, even when strongly settled prompts were used during evaluation. This highlights the need for more dedicated Kazakh evaluation resources and motivation for introducing a Word-in Context dataset in this dissertation.

Summary. This section motivates for evaluating models with a training-free evaluation: zero-show and few-shot. As the prompt design affects the outcome of the model, in the later chapter §3.4 methods will be discussed thoroughly and explained how the prompts should be constructed.

3.3 Low-resource language challenges

A lot of NLP research has shown that large language models tend to perform unevenly across languages, where performance is strongly correlated with the amount and quality of available training data. Languages which have limited digital presence are described as *low-resource*. These low-resource languages often lack both parallel data for translation, large-scale monolingual corpora and benchmarks which limits the development of robust pretrained models (Haddow et al., 2022). Although modern LLMs advertised that they can do multilingual tasks, their training data is dominated by English and other high-resource languages which makes gaps in semantic understanding for low-resource ones (Joshi et al., 2020).

Low-resource languages face several challenges. The scarcity of large, high-quality annotated datasets makes evaluation difficult and prevents fair comparison with high-resource languages. The second challenge is that many such languages are morphologically rich including Kazakh, which produce a wide variety of words from a single root. This property increases data sparsity and makes subword tokenisation harder as they are originally optimised for language with relatively simple morphology structure such as English (Ponti et al., 2019). Third, orthographic variation and inconsistent preprocessing standards add an extra problem for developing NLP pipelines on these languages.

In the case of Kazakh, orthography is standardised, but the language still faces a big difficulties due to the first two challenges. It is an agglutinative Turkic language with free word order and complex morphology, which makes the second challenge which we described above that a single lexical item can generate dozens of surface forms. By history, Kazakh language had Arabic, Latin and now it is in Cyrillic form which complicates the dataset unification and task of normalisation. While there are corpora sourses like Almaty corpus (Kazakh National Corpus, 2025), the Kazakh National Corpus (Qazcorpus, 2025), and Wiktionary (Wiktionary, 2025) which have some data source coverage, they are much smaller in size and limited than English equivalents. Also, as stated before KazMMLU benchmark (Togmanov et al., 2025) confirms that even state-of-the-art LLMs perform significantly worse on Kazakh language than on English. This shows that the challenges introduced above are a big problem in Kazakh. Along with these, evaluation resources remain limited. The multilingual benchmarks such as XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020) and other benchmarks do not include Word-in-Context (WiC) styled dataset in Kazakh language. This makes assessing of how well LLMs capture contextual meaning in Kazakh difficult. This presented work fills this gap by introducing a manually curated WiC-style test set in Kazakh. Where the main focus is on sense consistency rather than general classification. Therefore, the dataset offers a sharper probe of word semantic understanding in a language where evaluation resources are limited.

3.4 Prompting strategies: zero-shot vs few-shot

Prompting in the field of LLMS refers to guiding a pretrained language model by natural language instructions without updating its parameters. This is opposite of the traditional fine-tuning, where model weight is updated for each downstream task. As prompting does not need a time for fine-tuning and saves resources for training it has become a dominant way to evaluate LLMs and apply them across specific tasks. Evaluating language models with prompts exploits their ability to follow instructions and adapt in-context and sometimes for the new discoveries.

In *zero-shot* prompting, the model is given with only a task description and input. There are no other additional examples provided. In the case of this dissertation, while evaluating on the WiC task, the model will be simply asked whether the target word has the same sense in two sentences. The zero-shot approach is attractive because there is no additional data, but it is sensitive to style of instruction and can produce unstable results (Brown et al., 2020).

On other hand, there is *few-shot* prompting which involves adding examples of the task to the prompt. These demonstrations act as in-context examples which enables the model to align more closely with the task requirements. Previous work has shown that existence of the examples in the prompt and its format can matter more than the explicit task description itself (Min et al., 2022). In WiC, this could be in the form of two or three sample sentence pairs, where each has a target label 'True' or 'False', followed by a new pair for the model to classify. Few-shot prompting generally improves accuracy, but only up to a certain upperbound, as the prompts which contain more examples consume valuable context window space which will exceed input token size of models or make more cost (Dong et al., 2024).

In this report, both of these prompting methods are used during evaluation. This experiment will allow us to compare each prompt's effect to performance and it is connected to the second research question in Chapter 4. This comparison is important in the Kazakh context, where no training resources exist for WiC-style tasks. The analysis therefore helps to determine whether small, carefully chosen demonstrations can improve the model performance or whether zero-shot method is enough to capture semantic signals.

3.5 Prompt Language Effects

The language used during prompting can strongly influence large language model performance, especially for low-resource languages such as Kazakh. Previous research shows that English prompting yields better perfoemance, even when the target task is in another language (Ahuja et al., 2023). LLMs are trained on large corpora where English is dominant while Kazakh and other low-resource languages appear sparsely (Joshi et al., 2020; Hu et al., 2020). As a result models tend to perform better when prompted in English, because they can rely on richer learned representations and patterns acquired from high resource training (Liu et al., 2023).

At the same time, prompting in native language can be sometimes more beneficial for capturing the semantic cues and it will avoid translation problems. For example, word senses in Kazakh often map imperfectly with English equivalents, making fully English prompts a source of semantic distortion. This is important in the task of WiC, where subtle sense distinctions are important.

Empirical studies in cross-lingual prompting shows that hybrid strategies can sometimes fill this gap. For instance, few-shot demonstrations with English instructions combined with examples in Kazakh can yield stronger performance than using just one language in the prompt (Zhao et al., 2023). However, this effect is unstable across models and tasks and needs controlled evaluation.

For Kazakh language, evidence from multilingual evaluations such as KazMMLU (Togmanov et al., 2025) reports that model outputs significantly differs depending on whether instructions are in English or Kazakh, reflecting both training data imbalance and the model's ability to handle morphologically rich input.

The choice of prompt language presents a complex optimization problem.

English prompts may leverage the models' stronger English language understanding, potentially leading to better task comprehension and reasoning. However, Kazakh prompts might better align with the target language's linguistic features and cultural context, potentially improving performance on language-specific semantic phenomena.

In this dissertation, we systematically compare cross-lingual prompting (English instructions with Kazakh examples) against monolingual Kazakh prompting to quantify the extent of prompt language effects on WiC task performance.

3.6 LLM vs human performance

When doing evaluation of large language models, only comparing model results is not enough and it is important to have human performance. Human accuracy provides an upper bound for many NLP tasks. It indicates how trained annotators can solve the same problem without external tools. For semantic understanding tasks, human performance provides not only upper bound estimates, but it also helps identify observed models and system deficiencies (Nangia et al., 2019). For the original Word-in-Context (WiC) dataset, human annotators achieved around 80% accuracy (Pilehvar and Camacho-Collados, 2019). This result shows that even for humans, sense disambiguation is not easy. However, performance is much more than models' results.

In practice, LLMs performance on semantic understanding tasks is lower than human results, particularly in low-resource languages. For high-resource languages such as English, recent models like GPT4 can sometimes make closer performance or even match human performance on certain benchmarks (OpenAI, 2023). However, in cross-lingual benchmarks like XL-WiC, which is an extension of WiC to 12 languages, accuracy for multilingual models drops significantly compared to English (Raganato et al., 2020). All models used in XL-WiC experiment are BERT based, but gives some understanding of how will LLMs behave as they based on transformers architecture.

In the case of the Kazakh, recent work confirms these findings. The

KazMMLU benchmark (Togmanov et al., 2025) report shows that human performance on Kazakh multiple choice questions is significantly higher than model accuracy. These studies also found that LLMs show systematic weakness in both Kazakh language understanding and lackness in local knowledge. In total, these studies highlight that while humans can reliably interpret Kazakh text, LLMs stay behind, particularly on the task which require word semantic understanding.

This difference between human and model performances on the benchmarks is important for the present dissertation. Since WiC tasks involve understanding the word sense in two different contexts, it relies on fine-grained semantic distinctions which native speakers can do more reliably. On other hand, LLMs are more likely to have confusion in cases where subtle morphological changes or idiomatic usage affects the word sense. So, this motivates RQ3 where we compare and analyze human vs LLMs performance. By doing this experiment, it can be identified what is “good performance” for Kazakh WiC and what types of the instances are still outperformed by humans.

3.7 Summary and research gap

The review of existing literature gives several gaps. First, the Word-in-Context (WiC) task is a widely used benchmark for texting contextual word semantic understanding, which offers sharper probes than earlier similar datasets. Despite the development of multilingual variants of WiC as XL-WiC, no dedicated WiC dataset exists for Kazakh language. This shows that models’ ability to understand sense disambiguation in Kazakh has not been studied yet.

Secondly, studies on multilingual benchmarks confirm that large language models outperform on high-resource languages compared to low-resource ones which include morphologically rich languages. Kazakh language has agglutinative structure and relatively limited digital footprint which makes this language to have these challenges. However, recent works have evaluated general classification or questing answering and lack of evaluation on fine-grained semantic disambiguation.

Third, prompting strategies such as few-shot learning have been studied and shown that it improves results in English and other high-resource languages. However, their impact on Kazakh WiC has not yet been studied. Similarly, recent research demonstrates that prompt language (English versus the target language) influences the outcomes of the model. There exists research on prompt language impact on the results of the models on tasks except WiC in Kazakh. Also, a mix of English instruction and examples on Kazakh language during few-shot learning not studied yet. Therefore, these two questions remain open.

Finally, human performance is often used to show an upper bound in evaluation and there are no studies published which directly compares human and LLM performance on Kazakh semantic understanding. This gap makes difficult to interpret model results in relation to how native speakers can realistically achieve.

In total, these all gaps motivate the present dissertation research. By looking to Figure 3.1 we can see that this dissertation fills these gaps by (i) building the first Kazakh WiC-style dataset, (ii) systematically evaluating the modern LLMs (GPT-4, Gemini, GPT-3.5, and LLaMA-3) on this dataset, (iii) comparing zero-shot and few-shot prompting methods, and (iv) testing the impact of prompt language (English vs. Kazakh). Also, by comparing against human results, this study will provide the first picture where the contemporary LLMs succeed and where they have problems on contextual meaning in Kazakh.

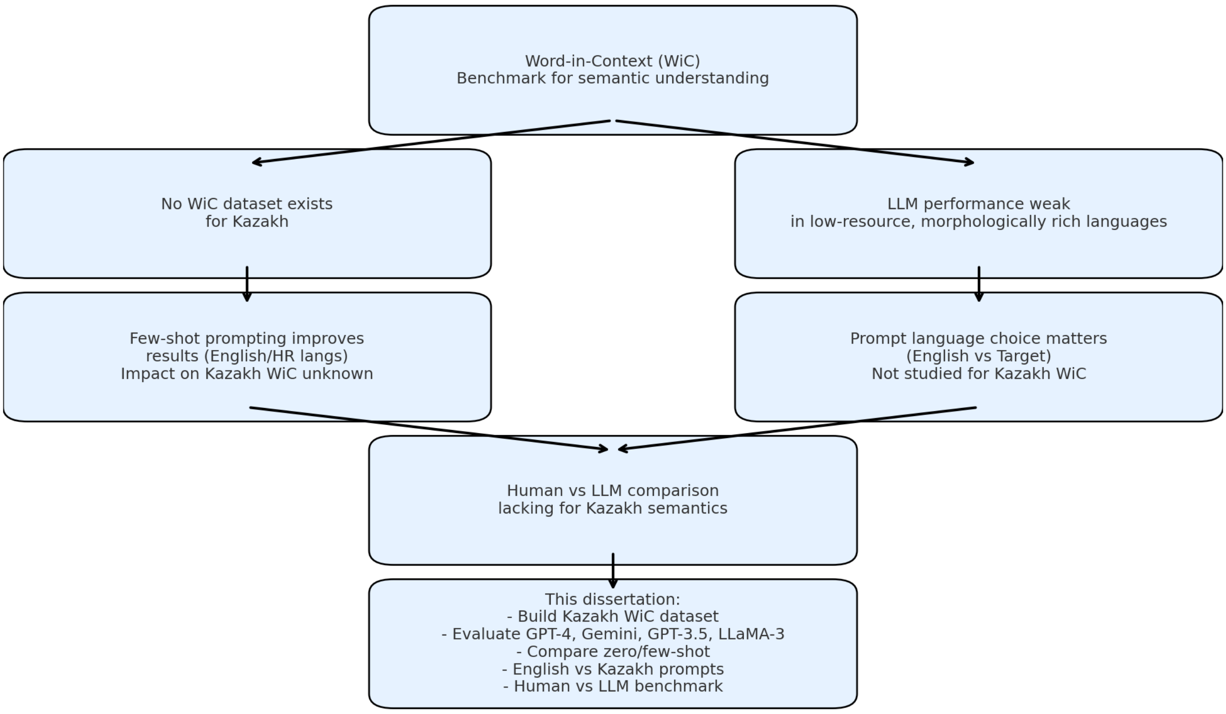

Figure 3.1: Flow chart summarising literature gaps and the contribution of this dissertation.

Chapter 4

Research questions and hypothesis

4.1 Introduction

This chapter builds the research questions and hypotheses that guide the rest of the dissertation. Analyzing the gaps in Chapter 3.7, it translates the literature review into a testable and focused experiment. The goal is to move from what is missing in previous work to how we will research and interpret it.

The main focus is the Kazakh Word-in-Context (WiC) task, a clear targeted probe of contextual meaning. The state-of-the-art models (LLMs) such as GPT-4, Gemini, and LLaMA3 will be evaluated and compared with the performance of human annotators. This human baseline is important, as it provides a reference point for judging LLMs and gives a clear view of where models diverge from native-speaker understanding.

More precisely, this chapter:

- shows three research questions (RQ1-RQ3) that come from the literature review;
- proposes three relevant, testable hypotheses (H1-H3) that link each question to measurable outcomes; and

- gives a clear outline of how these questions and hypotheses will be researched in later chapters through dataset creation, prompting methods, and human evaluation experiments.

The remaining part of the chapter will follow as follows: Section 4.2 states the research questions and briefly discusses the motivation of each one. Section 4.3 presents the hypotheses that will be tested using the Kazakh WiC data set and the evaluation of LLM. Section 4.4 summarizes the aim of the chapter and gives a short overview of how these elements will be the fundamental of the later experiments in the dissertation.

4.2 Research questions

Research gaps in Chapter 3 motivates to build the research questions in this chapter. This study addresses three central research questions. Each question is connected to the literature review on word semantic evaluation, prompting strategies, and instruction language effect.

RQ1: LLM performance on Kazakh WiC

To what extent can large language models (GPT-4, Gemini and Llama3) handle the word's meaning in context for the Kazakh language?

Lack of benchmarks to assess contextual semantic understanding in Kazakh was the foundation of this question. There exists Word-in-Context benchmark in English and other languages (Pilehvar and Camacho-Collados, 2019; Wang et al., 2019), but no equivalent to Kazakh. Evaluating LLMs on how they handle sense disambiguation in Kazakh therefore addresses a key gap in multilingual evaluation research and provides the first systematic baseline for this type of task.

RQ2: Prompting strategy

Does few-shot prompting outperform zero-shot prompting on the Kazakh WiC task?

The second research question came from evidence that in-context learning improves LLMs result, especially in low-resource languages (Brown et al., 2020; Min et al., 2022). A small number of examples in the prompt can help model to adapt to the structure of the task. However, this method is not tested yet for the Kazak WiC task, where the lack of huge training data and complex morphology might make the difference between the zero-shot and few-shot prompting.

RQ3: Prompt language choice

Does prompting in English yield better results than prompting in Kazakh for the Kazakh WiC task?

This last question addresses the huge difference in the available training data in English and Kazakh. Prior works have shown that multilingual LLMs often do better tasks with English instructions, even when the task itself has target language inputs (Zhao et al., 2023; Ahuja et al., 2023). Knowing that Kazakh is represented more rarely than English in the training corpora, it is important to test whether English instructions lead to stronger performance, or whether Kazakh instructions are better for models to have a clue about morphological structure of the language and semantic representation.

In total, these questions reflect the main challenges identified in the literature review: absence of the semantic evaluation benchmarks in Kazakh, unknown impact of few-shot on low-resource context, and the role of language of instruction in LLMs results. All of them provide the framework for the hypotheses presented in next section.

4.3 Hypotheses

Based on the research questions in the previous section, three testable hypotheses are formulated. The detailed rationales already have been discussed in Section 4.2. Here are the hypotheses in more precise form:

H1 - LLM vs Human Performance: Current large language models (GPT4o, Gemini, Llama3) will have significantly lower accuracy than human annotators on the Kazakh WiC task.

H2 - Zero-shot vs Few-shot: Few-shot prompting will achieve higher accuracy compared to zero-shot prompting for all tested LLMs.

H3 - Prompt Language Effects: Prompts in English will yield better performance compared to Kazakh prompts across all experiment cases.

4.4 Summary

The three hypotheses outlined in this chapter create a solid for the experimental work and results evaluations that follow in Chapter 5 and Chapter 6. Each hypothesis tackles a different piece of the puzzle.

The first hypothesis makes a foundation question: can modern large language models approach the same level as human performance when understanding the word's meaning in Kazakh? This motivates me to build crucial benchmarks to work with. The second hypothesis explores whether giving a few examples to the model before asking them to do the task makes a meaningful difference. This could be practically helpful and be used by anyone who will use these models with Kazakh text. The third hypothesis digs into whether the language used in instructions matters, which studies a broader issue about how these models handle low-resource languages beyond English.

These hypotheses are connected and do not exist in isolation. They fill the gaps of what I have discovered in the literature review and the actual experiments which I will conduct. More important is they will address real-world concerns about using large language models for Kazakh language. Also, it will contribute to a broader understanding of how these LLMs handle morphologically rich and under-resourced languages.

Testing these hypotheses will give me a concrete answer about the-state-of-the-art models limitations and provide practical insights for anyone thinking of deploying these systems for Kazakh or similar languages.

As Figure 4.1 shows, there is a clear path from research question to its corresponding hypothesis and to a specific experiment. This structure is helpful to understand that everything is connected: from the literature gaps identified earlier through the formulated research question to the experimental design which will provide answers.

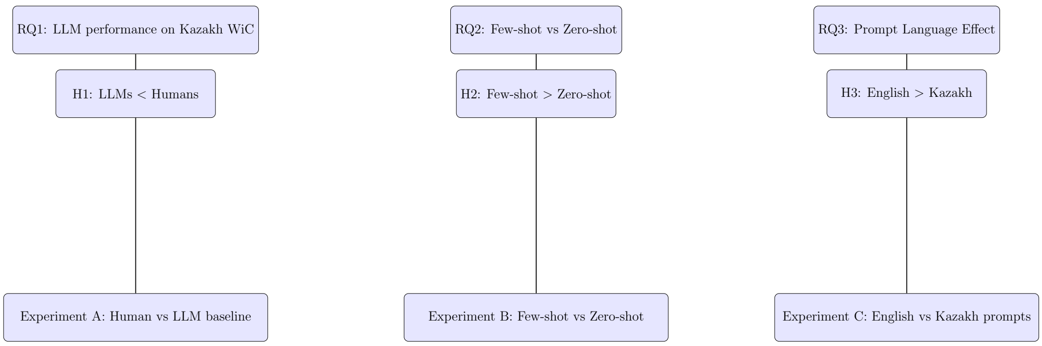

Figure 4.1: Mapping of research questions to hypotheses and experiments.

The next chapter outlines how actually these hypotheses will be tested. I will walk through my approach to creating the dataset, selecting and configuring the models, and designing evaluation methods that will give reliable answers to these questions.

Chapter 5

Methodology & Experimental Setup

5.1 Introduction

After creating research questions and hypotheses, it is time to explain exactly how these hypotheses will be tested. This chapter walks you through every step of the methodology: from creating the dataset to running the experiments to analyzing the results.

The goal of this chapter is transparency. It will give a clear understanding of what I did and why I made specific choices along the way. If someone wants to do some part of the work or run the experiment fully, they will have everything they need to do so.

The rest of this chapter will start by explaining how I built the Kazakh Word-in-Context dataset. This step involved collecting sentence pairs, cleaning them, and getting them ready for testing, where the process was seen as easy at first, but became more involved than I expected. Then we will discuss how language models have been selected and how the testing environment is set up. Then an approach of designing prompts in both English and Kazakh for different experimental conditions will be introduced. As comparing models with human performance is an important part of this dissertation, human annotators were selected to work on the same task. Here, the organization of processes and annotator agreements will be discussed.

The second part of this chapter will focus on the evaluation. The metrics used for performance measure, statistical tests on the results to get answers to hypotheses, and the approach of analyzing where models made mistakes will be discussed. Practical aspects of making this research reproducible and how I will share the data, code and other materials will be covered as well.

This chapter is a bridge between the theoretical framework built and the actual results which will be presented in the next chapter. Everything here is designed to be clear, systematic, and replicable, as good science should be something that others can build upon.

5.2 Dataset Creation

Sources and Scope

Creating the Kazakh WiC dataset was important for this study, as no dedicated benchmark was previously created for evaluating contextual semantic understanding in Kazakh. The dataset was designed by following the original format of the English WiC dataset (Pilehvar and Camacho-Collados, 2019), where each instance presents a target word appearing in two different sentences. The task is to determine whether the word has the same meaning in both contexts (True) or has different senses (False).

To ensure that the dataset contains authentic Kazakh language use, sentence pairs were collected from multiple available open sources. Kazakh language has limited sources and no similar equivalent to English's WordNet exists. So, it was a reason to collect sentences from various sources. The Almaty Corpus (Kazakh National Corpus, 2025) and Kazakh national corpus (Qazcorpus, 2025) are the substantial foundation of the benchmark, while contemporary news articles from major Kazakh media offered examples of current usage. Examples from digital lexicographic resources such as Wiktionary (Wiktionary, 2025) were also helpful to capture broader semantic contexts of the words. This multi-source approach ensured linguistic diversity and representation across different fields such as journalistic, literary and general usage, all adhering to standard Kazakh orthography.

Target Word Selection

Target words were restricted to **nouns, verbs and adjectives**, where in original WiC design there is only **nouns and verbs**, since these categories most frequently exhibit polysemy in Kazakh. The selection strategy was prioritising words that create realistic and challenging test cases. This included both high-frequency words that commonly encountered in daily language usage and in formal language use cases.

A particular challenge was with Kazakh verbs. Many verbs appear in citation form with the suffix -*у*, but we also considered truncated stems (e.g., *тұру* → *тұр*) to show morphological variation and ensure that the dataset reflect natural usage context.

Moreover, the main thing is that all candidate words were searched from digital dictionaries such as (Sozdik.kz contributors, 2025) and (Apple Inc., 2025). Even some sentences were extracted from these dictionaries as in some cases they contained example usages of the words in different senses. However, the main resources are Almaty Corpus and Kazakh National Corpus.

Selection Protocol

During dataset Creation, quality control is required to have both balance and authenticity. To keep them, the dataset is constructed so that it has equal distribution of True and False labels, while avoiding trivial or ambiguous cases. Sentence pairs with obvious overlaps or excessively subtle sense differences were removed to prevent ambiguous task and annotator disagreement.

Following the same practises as in multilingual WiC benchmarks (Raganato et al., 2020), the selection process was iterative and involved multiple rounds of review to ensure each instance is correctly constructed and has meaningful impact on the evaluation task.

Final Dataset Structure

The completed evaluation dataset contains *512* sentence pairs where 2 pairs are used for a few-shot prompt. The class distribution is nearly balanced between same-sense (true) and different-sense (False) cases which ensure that the dataset is meaningful for evaluation.

The dataset is stored in `JSONL` format, with each entry containing the following fields:

> `"word"`, `"sentence1"`, `"sentence2"`, `"idx"`, `"label"`, `"start1"`, `"end1"`, `"start2"`, `"end2"`, `"pos"`, `"version"`

Here, `label` is a boolean value (True/False); `start`/`end` indicate the character span of the first occurrence of the target word in each sentence (covering the full morphological form including suffixes); and `pos` shows the part-of-speech tag. Versioning metadata is included to support reproducibility and transparency.

In summary, the WiC-Kazakh dataset keeps the same principles as the original WiC benchmark by adapting it to low-resource, morphologically rich language. In the result, produced a dataset for robust evaluation of large language models on Kazakh contextual semantics.

5.3 Data Preprocessing

After data collection, preprocessing was needed to prepare raw sentence pairs for evaluation. The goal was to make sure that all instances of the dataset follow the same format, that target word is reliably identified in the context, and that the dataset could be used with prompts and evaluation scripts. The whole pipeline involved cleaning of text from empty spaces, span detection of target word and corpus statistics.

Text Cleaning

Manually taken raw sentences from various sources often contained irregular spacing or lacked terminal punctuations. To standardise sentence pairs, leading and trailing removed, excessive whitespaces were collapsed into single space. Also, it was checked for punctuations (".", "?" or "!") and if it was missing '.' punctuation added. All of these ensured uniform sentence structure and reduced variability in prompt inputs. Span indices for the target words added always after cleaning, to avoid misalignment caused by adding punctuation. All of these are done by python's regex library.

Target Word Span Detection

Kazakh verbs, nouns and adjectives appear with suffixes or with inflectional endings. To concretely find target word indices, it was implemented a Kazakh-aware regular expression which matches both words' initial form and its morphological variants. For example for verbs ending with suffix -*y*, truncated stems were also matched.

For each sentence, the pipeline returns the start and the end character offsets of the first occurrence of the target word(including its changed forms). These indices checked manually and corrected where the pipeline could not identify the spans. In most cases, the code returned valuable output for Kazakh. These spans are then stored in the dataset to facilitate error analysis and for other application cases. A simplifying assumption is that the first occurrence of the target word in a sentence is the relevant one which is documented in the dataset card. This is true as sentences collected to contain only one form of the target word. In the case of this study, these spans were not used. It was made to follow the same structure of the original WiC dataset.

POS and Data Format

Each entry of dataset has a single part-of-speech (POS) tag for the target word, which is identical in both provided sentences. The dataset is exported in `JSONL` format, with fields: `word`, `sentence1`, `sentence2`, `idx`, `label`, `start1`, `end1`, `start2`, `end2`, `pos`, and `version`. This structure provides both the text data and rich metadata required for evaluation and reproducibility.

Corpus Statistics

To explore and describe the dataset, we analysed it with the *Stanza* pipeline (Qi et al., 2020) for tokenisation and lemmatisation. The following statistics were extracted:

- Total number of instances (sentence pairs)
- Distribution of True vs. False labels
- Number of unique target words
- Number of unique word types in context
- Average sentence length (in tokens)
- Distribution of POS tags for target words
- Processing time for getting statisctics

These values are reported in Table 5.1, with a POS categories illustrated in Figure 5.1.

Table 5.1: Dataset statistics for WiC-Kazakh after preprocessing.

Statistic	**Value**
Instances (pairs)	510
True labels	258_{True}
False labels	252_{False}
Unique target words	147_{words}
Unique context tokens	5005_{types}
Average sentence length	10.54 tokens
Processing time	48.52 seconds

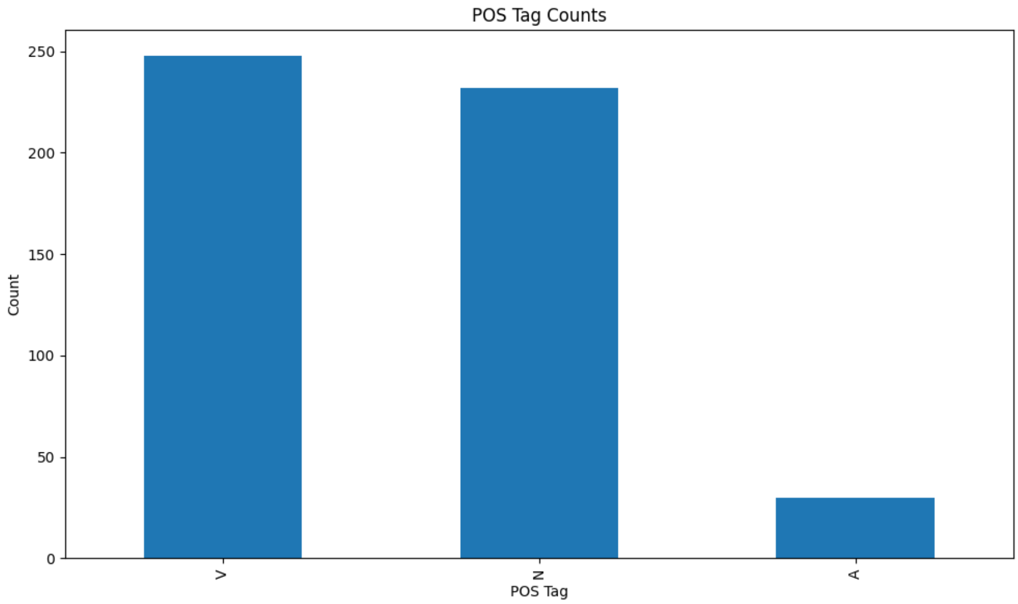

Figure 5.1: Distribution of POS tags among target words in WiC-Kazakh.

To conlude here, preprocessing made sure that the WiC-Kazakh dataset was clean, consistent, and well-documented. This process produced a morphologically sensitive and balanced evaluation set, which is suitable for testing large language models on contextual semantic understanding.

5.4 Model Selection and Runtime

Model Selection

For this experiment, three families of large language models (LLMs) were selected, reflecting a keeping of balance between closed source commercial systems and open-source alternatives. The main criteria when selecting the model were multilingual coverage, accessibility, and relevance to current research in natural language processing.

- **GPT-4o (OpenAI, 2024)**: One of the latest versions of OpenAI's GPT family. Optimised for reasoning and multilingual understanding

(OpenAI, 2023). OpenAI's GPT models are widely used across benchmarks as they have strong cross-lingual performance and experiments are reproducible through stable API endpoints. Also, GPT-3.5-turbo version was tested with same conditions.

- **Gemini 1.5 Pro and Flash (Google DeepMind, 2024)**: Google's Gemini is another alternative of commercial LLMs and it represents competitive modern system with more focus on the long-context reasoning and multilingual support (Gemini Team, 2024). Both the Pro and Flash variants were evaluated to compare runtime trade-offs between efficiency and accuracy.
- **Llama3 (Meta, 2024)**: An open-source alternative which can be run locally without depending on paid API. This model is included to know how open models perform in Kazakh compared to proprietary systems when there is limited resource available.

These choices of models show that both commercial and open-source LLMs are alternatively covered. This also ensures that the results of evaluation are informative for both research and applied contexts in low-resource Kazakh language evaluation.

Runtime Environment

API models. GPT and Gemini models were accessed by their paid APIs. All calls to API were made with selected parameters (`temperature` set to 0.0, `max_tokens` restricted to small values which are suitable for binary output) to ensure that it is reproducible and minimise variability.

Local runs. Llama3 was run locally on macOS hardware using Ollama (Ollama Team, 2025). It automatically applied quantisation as macOS RAM is only 8 GB. All model inference was made through calling a local API where the llama3 model runned on background. This ensured that results could be replicated independently of paid API availability.

To conclude, the combination of GPT-4o, Gemini 1.5 (Pro and Flash), and Llama3 covers various areas of LLMs: powerful closed-source APIs widely used in industry, and open-weight alternatives that can be deployed locally. This diversity ensures that the evaluation of Kazakh Word-in-Context performance is both comprehensive and reproducible.

5.5 Prompt Design

Strategy Development

Prompting design plays a crucial role during evaluation of large language models (LLMs), as even small changes in the prompt can substantially affect outputs (Brown et al., 2020; Dong et al., 2024). In this study, as introduced in Chapter 3, we implemented two widely used strategies: *zero-shot* and *few-shot* prompting.

In the **zero-shot** condition, the LLM receives only task instructions. The instructions contain a clearly specified evaluation goal, determining whether the same word has the same meaning in two different sentences and the expected output format. This method probes the models ability to use knowledge from pre-training to the Word-in-Context (WiC) task without demonstrated examples.

In the **few-shot** condition, the same instructions were used with additional labelled examples, by following the findings from the research that demonstrations can significantly improve the model accuracy in low-resource scenarios (Min et al., 2022). Each few-shot prompt includes two labelled examples, one for 'True' and one for 'False' case. The important thing is that few-shot examples were held fixed across models for fairness.

Language Variations

As said before during literature review, most LLMs demonstrate uneven multilingual ability (Hu et al., 2020; Liu et al., 2023), I used prompts in both *English* and *Kazakh*. The english prompt in a few-shot contained examples in Kazakh language and only instructions were in English. This experiment allows us to test the second hypothesis: does English prompts make LLMs do better than Kazakh prompts? This also means that whether instructions in English give a model advantage due to a large English corpus in pre-training and whether Kazakh instructions yield more natural alignments with the data itself.

English prompts have a following style: *"You are given two sentences that both contain the same target word. Your task is to determine whether the word has the same meaning in both sentences. If the meaning is the same, answer "True". If the meaning is different, answer "False". Respond with only "True" or "False"."*

Kazakh prompts were also carefully constructed and reviewed by native speakers to ensure naturalness and equivalence. Prompts in Kazakh followed the same structure as in English: *"Саған бірдей мақсатты сөзді қамтитын екі сөйлем берілген. Сенің тапсырмаң – осы сөздің екі сөйлемде де бірдей мағынада қолданылған-қолданылмағанын анықтау. Егер мағынасы бірдей болса, "True"деп жауап бер. Егер мағынасы әртүрлі болса, "False"деп жауап бер. Тек "True"немесе "False"деп жауап бер."*

Previous studies show that multilingual prompting can influence cross-lingual generalization (Raganato et al., 2020; Zhao et al., 2023), making it important to include both languages in this evaluation. Moreover, recent benchmarks like KazMMLU (Togmanov et al., 2025) and pilot evaluations of Kazakh LLMs (Maxutov et al., 2024) highlight the necessity of strict testing of instructions in underrepresented languages. Therefore, my variant of prompting will also influence testing prompts in low-resource languages.

Compliance Guard and Reproducibility

To prevent ambiguity, all prompts are strictly instructed to output either "True" or "False", even for Kazakh prompts. Responses were parsed using a strict boolean parser where variants of `true` and `false` were accepted, but all other outputs considered as invalid. This design choice ensures consistency and prevents partial-credit interpretations which meets the best practises for benchmark reproducibility (Wang et al., 2019).

For transparency, the full version of prompts including zero-shot and few-shot variants are shared in **Appendix**. Sharing exact prompt strings enables future researchers to replicate or extend this evaluation study for Kazakh or other low-resource languages.

5.6 Human Evaluation

Participants

Test items independently annotated by two native Kazakh speakers.[1] Both of them have university-level education and advanced literacy in standard Kazakh. Table Table 5.2 summarizes demographics.

Table 5.2: Annotator demographics.

Annotator	Education	Linguistic background
A	*B.A.*	Native Kazakh; fluent Russian
B	*B.A.*	Native Kazakh; fluent Russian

Materials and Interface

Annotators did the task on the same Kazakh Word-in-Context instances that models were evaluated on. Each instance consists of a target word and two sentences. Keeping the WiC protocol, the task was to judge whether the target word has the same sense in two contexts ('True') or different sense ('False'). All items were presented in a randomized way in Excel file to avoid position effects. Instructions given to annotators were the same as in Kazakh prompt used for models to ensure procedural parity with the evaluation of LLMs.

Procedure

Annotators worked independently and no communication between them was held. For each instance in the dataset, they selected one of two labels: `True`

[1] Annotators did not have prior exposure to the Kazakh WiC dataset or LLMs outputs. Participants were compensated with consent recorded. No personal information is reported.

(same sense) or `False` (different sense). After annotation of all items, I computed pre-adjudication agreement and then conducted a short adjudication session where two annotators reviewed only the disagreements and discussed their rationales. Also initial gold labels were considered as well and in total one label had three votes. Some labels where both annotators had disagreement with the initial gold label were changed by their rationales and some instances removed due to ambiguity. In total 5 instances changed their initial label, removed 10 instances and the final dataset contained 512 entries where two entries used for few-shot examples. Final adjudicated labels constitute the **gold standard** used for evaluating model predictions and as for human baseline initially before adjudication accuracy was taken.

Quality Control

To make consistent judgments, the instructions included brief positive/negative examples which were used in zero-shot prompts and the small FAQ session conducted with annotators to cover all the questions which could be raised during task completion. Also, no additional feedback provided during annotation. It was measured how long each annotator spent time on doing and none were flagged. As in prior WiC-style work, minimal guidance used and no additional lexicographic knowledge provided beyond the core task definition to avoid overfitting to annotation artifacts (Pilehvar and Camacho-Collados, 2019).

Agreement Metrics

I reported (i) **raw agreement** which is proportion of identical labels defined by two annotators and (ii) **Cohen's** κ for two raters (Cohen, 1960). Cohen's κ adjusts for chance of agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is observed agreement and p_e is the expected agreement if both annotators kept guessing randomly but kept their label distribution. I computed κ using the standard implementation (binary labels). To make fair annotators agreement was calculated before adjudication.

Table 5.3: Human agreement and accuracy summary (pre-adjudication).

Metric	Value
Raw agreement	**89.3%**
Cohen's κ	**78.6%**
Annotator A accuracy vs. gold	**97.4%**
Annotator B accuracy vs. gold	**89.5%**

For Cohen's Kappa value it is better if it is close to 1. If it is near to 0 this means that agreement is no better than random guessing. So, In this report case we have a 78.6% which shows that there is a substantial agreement and the dataset is valuable for evaluation.

Gold Standard and Human Baseline

Adjudicated final labels serve as the gold standard for the all models evaluation. For the human baseline, I took their average accuracy. This aligns the human baseline for future comparison of the models performance after evaluation.

Link to Hypotheses

H1 (LLM vs. Human). I will compare each LLMs accuracy to the human pre-adjudication accuracy (baseline). If the model's accuracy is lower than the annotator's accuracy, I will accept H1. If it is higher, then reject H1. The all results are reported in the Results Chapter 6 with appropriate metrics.

Limitations

Only two annotators used in this experiment and gold standards were finalised mostly by the author. On other hand, Cohen's kappa of two annotators is well-defined and interpretable which can show some reliability of data for evaluation. Additional annotators could further stress-test robustness and reliability of the dataset. However, in limited resources and no funding this is the resource-efficient method.

5.7 Evaluation Metrics

Primary Metric

Accuracy. As the main metric accuracy was chosen. This is because dataset is balanced across binary labels and it will be enough to analyze and compare results. Accuracy is a just proportion of correctly classified instances. Also, it is directly comparable to human baseline shown in Section 5.6 which is intuitive to interpret and follow same comparing metric in the original WiC benchmark (Pilehvar and Camacho-Collados, 2019; Wang et al., 2019). Given binary labels $y_i \in \{\texttt{True}, \texttt{False}\}$ and predictions $\hat{y}_i$, accuracy is

$$Acc \;=\; \frac{1}{N}\sum_{i=1}^{N} 1[\,\hat{y}_i = y_i\,] \;=\; \frac{TP + TN}{TP + TN + FP + FN}.$$

It was treated `True` ("same sense") as the positive class.

Secondary and Diagnostic Metrics

F1-score. To complement accuracy and to know is model doing asymmetric error pattern (like over-predicting `False` or `True`), I reported F1-score. This score balances precision and recall and very informative for showing how model is performing accross label classes. However, all the main analysis done by comparing accuracy metric. Below is the fornulas of F1-score, Recall and precision.

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \quad F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

It was additionally included **Macro-F1** (unweighted mean over classes) and **Weighted-F1** (class-frequency-weighted mean) to reflect potential skewed decisions. It is included just in case and not used during analyze of results. They are provided in Appendix.

Confusion matrix counts and Precision/Recall. To know how models performed for each class and see the numbers, I reported confusion counts (TP, TN, FP, FN). Precision and recall is important to know: precision reflects whether the model is aggressive in predicting positives(risking False cases), while recall reflects whether the model is conservative(risking True cases).

Reporting Conventions

All metrics provided are calculated by `compute_metrics (...)` in `utils.py` (accuracy; macro, and weighted F1;precision/recall/F1; TP/TN/FP/FN) file using the scikit-learn library (Pedregosa et al., 2011). In the main important result table, I will report **Accuracy** and **F1**, as these best summarize the overall performance of the models and are good for comparing with the human baseline. The confusion matrix will be reported in the error analysis section of Chapter 6. Other metrics results are provided in the appendix for completeness and other types of analysis for future.

5.8 Error Analysis

Error analysis is crucial for understanding why models make the errors (Belinkov and Glass, 2019). This includes not only numerical outputs, but also analysis of examples of outputs. In this study, we mix quantitative (numbers) and qualitative (examples) methods.

Quantitative Error Analysis

First I will show an error in terms of part-of-speech (POS). This is because my Kazakh WiC dataset includes three types of POS tags. I will calculate accuracy separately for nouns, verbs, and adjectives. This shows whether LLMs struggle more with certain word types. Along with this, verbs in Kazakh often have complex suffixes (tense, aspect, derivation) which may make them harder than nouns.

As stated in the Metrics section, confusion statistics will be reported. This will tell about models behavior like is it conservative (predicting more False cases) or aggressive (predicting more True cases). These numbers also show precision and recall metrics and will make model results more interpretable.

Qualitative Error Method

As numbers alone do not explain the reasons for failures, some misclassified cases were manually inspected. Kazakh WiC does not contain information like sense inventories like polysemy/homonomy labels, so I focused on linguistic and contextual factors that appeared several times. These observed categories are:

- **Morphological ambiguity** – when word meaning is changed subtly by suffixes.
- **Idiomatic or figurative usage** – when the meaning of a word is used in another sense than literal one.

Each type will be shown with examples in the Discussion section.

Cross-Model Comparison

Finally, I compare model performance for error patterns. Some errors are common for both models which will suggest that Kazakh linguistic information is underrepresented in training data. Other errors differ like one LLM will fail more on verbs, while another struggle on nouns. This comparison highlights strength and weaknesses of each model.

Summary

To conclude, error analysis in this dissertation will include:

1. **Quantitative profiling:** accuracy by POS and confusion statistics;
2. **Qualitative review:** reading and tagging misclassified examples;
3. **Cross-model comparison:** identifying shared and unique weaknesses.

This structured analysis of errors provides a precise picture of how modern LLMs succeed and where they fail in Kazakh semantic understanding.

5.9 Reproducibility & Artefacts

Ensuring that the results of the study can be reproduced afterwards and can be extended by other researchers is an important part of scientific practice (Drummond, 2009). Therefore, I made all code, dataset and experimental short documentation publicly available. This section describes what is released, how it is organised.

Code and Dataset

The Kazakh WiC evaluation dataset contains 510 entries and is saved in `.jsonl` format following the schema described in Section 5.2. This dataset is for evaluation purpose only and shared under appropriate research licensing to support further studies. Documentation includes data collection process, annotation guidelines, adjudication procedures and dataset statistics.

All code and dataset are publicly available in the Github repository: `https://github.com/Nedzhin/WiC_Kazakh_dataset`. The repository contains:

- **Preprocessing scripts** (text cleaning, tokenisation, and format conversion),
- **Evaluation utilities** (`utils.py`, metrics functions, annotator agreement computation scripts),
- **Prompt templates** (Kazakh and English variants).

In addition to the public repository, a complete archived version of the code and dataset is submitted in `.zip` format with this dissertation for assessment purposes.

Following the best practises in reproducible evaluation (OpenAI, 2023; Gemini Team, 2024), all API model calls were run with temperature fixed at 0 to minimise randomness. When the model is runned locally, seeds are fixed. When I runned each model three times they produced identical output with no variance across runs. This is because shuffling of the dataset does not have an effect on results as each time we are given an example at one try. Therefore, I reported a single result per model. Reproducibility is supported by providing library versions in `requirements.txt` file which specifies exact versions of tokenisers and inference frameworks used. All of these ensure that the experimental environment can be reused by future research.

Summary

In short, all artefacts (dataset, code and documentation) are openly released and archived. This enables independent verification studies and provides a foundation for future research in the field of Kazakh semantic understanding.

5.10 Ethics, Licensing, and Data Statement

Ethical Considerations

This dissertation research follows the ethical standards for computational linguistics (Bender and Friedman, 2018). Human participants were compensated fairly and their identities remained anonymised (see Section 5.6). No sensitive data was collected and the task was restricted to judgments on the linguistic meaning of words. Therefore, the project should not raise significant ethical risks.

Licensing and Usage Rights

All sentences in the dataset were collected from legitimate, non-restricted open online sources for academic research. All of these sources were cited and referenced fairly. The final Kazakh WiC dataset was released in open source, enabling broad use in academic and non-academic research. All code is released under open-source MIT license. Use of commercial APIs (OpenAI, Google Gemini) followed the providers' terms of service (OpenAI, 2023; Gemini Team, 2024).

Data Statement

Following the framework proposed by Bender and Friedman (2018), I provided structured description of dataset:

- **Curation rationale:** Created a benchmark for evaluating word semantic understanding in Kazakh which is a low-resource and morphologically rich language.
- **Language variety:** Standard Kazakh language used in contemporary Kazakhstan, in Cyrillic script.
- **Speaker demographics:** Annotators are two native Kazakh speakers with advanced literacy and university-level education (see Section 5.6).
- **Annotation process:** Native speakers annotated WiC independently. This is followed by adjudication of dataset and computation of inter-annotator agreement which are reported in Section 5.6.
- **Quality control:** Provided clear instruction with live FAQ session. Also, adjudication ensured consistent judgement. Along with this items with ambiguity labels were removed.

Summary

To conclude, the dataset is sourced in ethical form, documented and released under open licensing. This description is written to promote reuse and ensure future researchers can access the dataset's scope, limitations and ethical compliance.

Chapter 6

Results

6.1 Introduction

In this chapter, the results of the evaluation of large language models (LLMs) on the newly constructed Kazakh Word-in-Context (WiC) task are presented. The models which are evaluated include *Gemini-1.5-Flash*, *Gemini-1.5-Pro*, *GPT-4o*, *GPT-3.5-Turbo*, and *Llama3*. Different experiment configurations such as zero-shot and few-shot prompting and the impact of using English versus Kazakh prompts applied. Three hypotheses which guide the evaluation are:

- **H1:** The performance of LLMs on the Kazakh WiC task will be significantly lower than human annotators performance.
- **H2:** Few-shot prompting will have more accuracy over zero-shot prompting of LLMs.
- **H3:** English prompts will yield better performance than pure Kazakh prompts.

Therefore, this chapter follows this structure:

- Section 6.2 presents overall model performance across different configurations and compares the LLMs results to human baseline result.

- Section 6.3 compares the zero-shot and few-shot performance and provides an analysis of prompting strategies.
- Section 6.4 explores the effects of English versus Kazakh prompt language.
- Section 6.5 focuses on error analysis which include error patterns by part of speech and common error types.

6.2 Overall Model Performance (RQ1/H1)

The Table 6.1 shows the overall performance of the LLMs on the Kazakh WiC task. Accuracy and F1-score for each model across different configurations, such as prompt type and languages in prompts, included in the table.

Model	Prompt Mode	Language	Accuracy	F1-Score
gemini-1.5-flash	zero	kk	60.20%	36.76%
gemini-1.5-flash	few	kk	66.27%	51.69%
gemini-1.5-flash	zero	en	61.57%	40.61%
gemini-1.5-flash	few	en	76.27%	72.06%
gemini-1.5-pro	zero	kk	68.82%	57.14%
gemini-1.5-pro	few	kk	75.88%	70.78%
gemini-1.5-pro	zero	en	66.67%	52.24%
gemini-1.5-pro	few	en	**77.06**%	**76.27**%
gpt-4o	zero	kk	63.92%	44.91%
gpt-4o	few	kk	73.73%	65.46%
gpt-4o	zero	en	67.25%	53.22%
gpt-4o	few	en	75.49%	68.83%
gpt-3.5-turbo	zero	kk	53.33%	56.57%
gpt-3.5-turbo	few	kk	52.75%	41.36%
gpt-3.5-turbo	zero	en	**48.82**%	61.03%
gpt-3.5-turbo	few	en	53.14%	60.23%
llama3	zero	kk	52.55%	67.30%
llama3	few	kk	53.73%	64.35%
llama3	zero	en	54.51%	67.51%
llama3	few	en	53.53%	**21.26**%

Table 6.1: Overall Model Performance (RQ1)

The results show that, across all models, *Gemini-1.5-Pro* achieved the highest accuracy in the few-shot configuration with English prompt. It reached 77.06% accuracy and 76.27% F1-score. If we look at the table we can see that the same model achieved the second highest accuracy with a few-shot configuration, but in Kazakh language prompt. The closest model to *Gemini* was *GPT-4o* which achieved 75.49% accuracy and 68.83% F1-score.

When comparing these results to the human baseline, which achieved 93.45% accuracy, it can be understood that there is a significant gap in model performance. This evidence supports **H1**, which stated that LLMs would underperform than human annotators.

6.3 Zero-Shot vs Few-Shot (RQ2/H2)

Table 6.2 compares the performance of zero-shot and few-shot prompt configurations across all models with different prompting languages. Overall, the results show that few-shot prompting consistently improved performance compared to zero-shot, where average increase in accuracy 6.02%. The significant raise in accuracy ranges from 4% to 14%.

Model & Lang	**Zero**	**Few**	**Improvement**
gemini-1.5-flash kk	60.20%	66.27%	+6.07%
gemini-1.5-flash en	61.57%	76.27%	**+14.70%**
gemini-1.5-pro kk	68.82%	75.88%	+7.06%
gemini-1.5-pro en	66.67%	77.06%	+10.39%
gpt-4o kk	63.92%	73.73%	+9.81%
gpt-4o en	67.25%	75.49%	+8.24%
gpt-3.5-turbo kk	53.33%	52.75%	-0.58%
gpt-3.5-turbo en	48.82%	53.14%	+4.32%
llama3 kk	52.55%	53.73%	+1.18%
llama3 en	54.51%	53.53%	**-0.98%**

Table 6.2: Zero-Shot vs Few-Shot Performance (RQ2)

Provided results support the **H2**, indicating that few-shot prompting improves performance across most models. However, *gpt-3.5-turbo* in kazakh prompting and *llama3* in english prompting shows a slight decrease in accuracy when the prompting method switched to a few-shot configuration. This could be due to the model's limited capacity to handle additional examples effectively.

6.4 English vs Kazakh Prompting (RQ3/H3)

The table 6.3 presents the comparison between English and Kazakh prompts for all LLMs with different prompting method configurations. The results show that English instruction with Kazakh examples and just English instruction presented improvements in most cases.

Model & Prompt Method	**Kazakh**	**English**	**Difference**
gemini-1.5-flash zero	60.20%	61.57%	+1.37%
gemini-1.5-flash few	66.27%	76.27%	**+10.00%**
gemini-1.5-pro zero	68.82%	66.67%	-2.15%
gemini-1.5-pro few	75.88%	77.06%	+1.18%
gpt-4o zero	63.92%	67.25%	+3.33%
gpt-4o few	73.73%	75.49%	+1.76%
gpt-3.5-turbo zero	53.33%	48.82%	**-4.51%**
gpt-3.5-turbo few	52.75%	53.14%	+0.39%
llama3 zero	52.55%	54.51%	+1.96%
llama3 few	53.73%	53.53%	-0.20%

Table 6.3: English vs Kazakh Prompting Performance (RQ3)

Gemini-1.5-pro and *gpt-3.5-turbo* in zero-shot configuration, *llama 3* in few-shot configuration showed a slight decrease in performance with English prompts. The difference between languages is small, with the significant performance increase in *gemini-1.5-flash* with few shot method and a notable decrease in *gpt-3.5-turbo* with zero-shot prompting mehtod where overal average increase in accuracy was 1.31%. This is yet improvement, but not a significant evidence to support the **H3** for all models.

6.5 Error Analysis

6.5.1 POS-level accuracy

Below table 6.4 reports per-part-of-speech (POS) accuracies for each model under zero-shot and few-shot prompting, in both Kazakh and English language prompting. I reported three available POS categories in the dataset: adjectives (A), nouns (A), and verbs (V).

Model	Prompt	Lang	A (%)	N (%)	V (%)
Gemini 1.5 Flash					
gemini1.5flash	zero	kk	63.33	58.62	61.29
gemini1.5flash	zero	en	60.00	61.21	62.10
gemini1.5flash	few	kk	70.00	66.81	65.32
gemini1.5flash	few	en	**83.33**	75.00	**76.61**
Gemini 1.5 Pro					
gemini1.5pro	zero	kk	70.00	70.26	67.34
gemini1.5pro	zero	en	66.67	66.81	66.53
gemini1.5pro	few	kk	80.00	78.88	72.58
gemini1.5pro	few	en	76.67	**80.17**	74.19
GPT-3.5 Turbo					
gpt3.5turbo	zero	kk	53.33	51.72	54.84
gpt3.5turbo	zero	en	50.00	47.84	49.60
gpt3.5turbo	few	kk	**43.33**	50.43	56.05
gpt3.5turbo	few	en	53.33	56.90	**49.60**
GPT-4o					
gpt4o	zero	kk	66.67	64.22	63.31
gpt4o	zero	en	66.67	65.52	68.95
gpt4o	few	kk	63.33	73.71	75.00
gpt4o	few	en	76.67	75.43	75.40
Llama3					
llama3	zero	kk	56.67	51.29	53.23
llama3	zero	en	56.67	53.88	54.84
llama3	few	kk	66.67	**50.00**	55.65
llama3	few	en	53.33	53.45	53.63

Table 6.4: POS-level accuracy by model, prompt mode, and prompt language. A = adjectives, N = nouns, V = verbs.

There are 30 adjectives, 232 nouns and 248 verbs in the dataset. From the results of the models per POS it can be seen that:

- **Few-shot is also generally better across POS**, the most improved models are *gemini-1.5-flash* and *gemini-1.5-pro*. For example, *gemini-1.5-flash* improves from (A,N,V) = (60.0, 61.21, 62.10) in *zero/en* to (83.33, 75.00, 76.61) in *few/en*.

- **GPT-4o is consistently strong across all POS** under few-shot and english prompt: *few/en* (A,N,V) = (76.67, 75.43, 75.40).

- **Gemini models peaks on all POS**: *few/en pro* shows N=80.17%, while *few/en flash* has A=83.33% and V=76.61%.

- **GPT-3.5 Turbo and Lllama3 models lags on all POS categories**, with some improvements in few-shot settings. However, in general these models show the least performance near to guessing one.

The table 6.5 below shows metrics beyond the accuracy and F1 score. From these results we can know how models tend to perform: are they aggressively predicting True or conservatively predicting more False cases.

Observations:

- Both **Gemini models** shown balanced behaviour: *1.5-Flash* and *1.5-Pro* achieved relatively high TN while steadily improving TP with few-shot prompt. For instance, *gemini-1.5-pro few/en* got 188 TP but also yield a higher number of FP (47).

- Most conservative model was **GPT-4o**: in all configurations, FP were very low ranging from 1 to 5, while FN remained comparatively high. This means that model has bias towards predicting "False" unless there is strong evidence for "True".

- **GPT-3.5-turbo** shows unstable patterns. In *zero/en*, it predicts 205 true positives but also 208 false positives, showing models aggressive behaviour towards positive class. Its few-shot results reduce false positives but increase false negatives which reflects inconsistency.

Model	**Prompt Mode**	**Language**	**TP**	**TN**	**FP**	**FN**
gemini-1.5-flash	zero	kk	59	248	4	199
gemini-1.5-flash	few	kk	92	246	6	166
gemini-1.5-flash	zero	en	67	247	5	191
gemini-1.5-flash	few	en	156	233	19	102
gemini-1.5-pro	zero	kk	106	245	7	152
gemini-1.5-pro	few	kk	149	238	14	109
gemini-1.5-pro	zero	en	93	247	5	165
gemini-1.5-pro	few	en	188	205	47	70
gpt-4o	zero	kk	75	251	1	183
gpt-4o	few	kk	127	249	3	131
gpt-4o	zero	en	95	248	4	163
gpt-4o	few	en	138	247	5	120
gpt-3.5-turbo	zero	kk	155	117	135	103
gpt-3.5-turbo	few	kk	85	184	68	173
gpt-3.5-turbo	zero	en	205	44	208	53
gpt-3.5-turbo	few	en	181	90	162	77
llama3	zero	kk	249	19	233	9
llama3	few	kk	213	61	191	45
llama3	zero	en	241	37	215	17
llama3	few	en	32	241	11	226

Table 6.5: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)

- **Llama3** performs poorly overall: in zero-shot configuration it has high false positives, but in *few/en*, the opposite happens: true positives drops to 32 while false negatives rises to 226. This suggests unstable behaviour across prompting conditions.
- Across all models, **few-shot method generally reduces false negatives** Gemini-1.5-flash went from 199 to 166 in kk, GPT-4o went from 183 to 131 in kk setting. However, it often raises false positives slightly.

These observed patterns show how each model balances conservative versus aggressive. Gemini and GPT-4o models have relatively balanced tradeoff, while GPT-3.5-turbo and Llama3 models show unstable behaviour, either over-predicting positives or failing to capture true cases.

6.5.2 Common Errors

After manual review of misclassified instances, I revealed several recurring error patterns. The most common category which could lead to misclassification could be morphological confusion where models struggled to identify the same sense because of different inflected or derived forms of the same root. Another one is figurative use of the same word in two different contexts. Below, given examples of the most frequent categories observed across all models. The given examples were the most disagreed ones with the gold label by the results of the models. Overall 16 models configurations out of 20 had the another label than gold one.

Illustrative Examples of common errors

- **Morphological Confusion:** "сүю" (kiss) word used in different morphological form in sentences and it will be the main cause of major misclassification. The llama3 model with kazakh prompts and english instruction in zero-shot and gpt-3.5-turbo in zero en configuration handled this case.

- **Figurative way of use:** "шаң" (dust) used in figurative way in both sentences while meaning the same thing (which mean not allowing even the dust of that thing). Only llama3 with its kazakh prompts across zero and few shot method and gpt-turbo-3.5 with few kk configuration could predict it correctly.

Here, I provided possible causes of the error accurred by analysing manualy the results of each configuration. More discussion of it is in Chapter 7.

6.6 Summary of Hypotheses

H1: Supported. LLMs had low accuracy compared to human baseline which achieved 93.45% accuracy

H2: Supported. Few-shot prompting method improved performance across most models except some small accuracy drop in *gpt-3.5-turbo kk* and *llama3 en* cases.

H3: Not fully supported. While English prompts improved performance nearly to all models, the improvement was not significant and it is hard to support it in general.

6.7 Chapter Summary

This chapter provided results from the evaluation of LLMs on different experiment configurations and compared the results. These results highlighted the challenges faced by LLMs in low-resource language and provided insights into benefits of few-shot prompting and language choice during prompting. Detailed error analysis showed each model's key areas for improvement and gave a full description of how each model performed on task.

Chapter 7

Discussion

7.1 Introduction

This chapter explains the results reported in Chapter 6 in the context of the research questions (RQ1-RQ3) and hypotheses (H1-H3) outlined in Chapter 4. The discussion draws connection to existing literature reviewed in Chapter 3, evaluates whether each hypothesis supported or refuted and interprets why some models worked better than others. It also considered how these findings could be applied for other Kazakh and low-resource language research and Large Language Model (LLM) evaluation.

7.2 LLM Performance on Kazakh WiC

The important finding was that all models performed much lower than the human baseline. While human annotators reached over 93% accuracy, the best model achieved about 77%. This shows that even the strongest LLM across chosen ones is not fully reliable for Kazakh semantic tasks. These findings are consistent with prior studies which report performance gap between humans and LLMs on semantic disambiguation in low-resource language (Raganato et al., 2020; Joshi et al., 2020). The main reason for this big gap is the limited coverage of Kazakh in the training data of most models. Morphological complexity of Kazakh also causes tokenisation and representation challenges (Makhambetov et al., 2013). If a model has seen little Kazakh data during training, it will be difficult to separate various meanings of the same root word. This could be an explanation of why models often confuse literal and metaphorical use cases or different grammatical forms of the same word. These similar patterns have also been reported in other low-resource languages, so the problem is not unique for Kazakh. When comparing models, *Gemini 1.5 Pro* and *GPT-4o* were consistently stronger than *GPT-3.5-Turbo* and *Llama3*. This could be because of stronger models trained with more multilingual data and improved instruction-tuning which helps them perform better on unseen languages. The weaker models showed inconsistent behaviour: either over-predicting similarity, either being too cautious. In general, H1 (LLMs will perform significantly worse than humans) is supported by results of this study.

7.3 Few-Shot vs Zero-Shot Prompting

Few-shot prompting almost in all cases improved the LLMs results. On average, accuracy of models increased by around 6 percent, where in some cases it showed larger improvements. This findings support the idea that in-context learning enhances the LLMs performance, particularly in morphological rich, low-resource languages (Brown et al., 2020; Liu et al., 2023).

The one cause of in-context learning enhancement may be that Kazakh has many surface variations and examples in the prompt show that different endings and different forms can still be semantically same. In other words, examples give the model a clear picture of what "same sense" and "different sense" looks like.

As it can be seen, not all models benefited from few-shots. *GPT-3.5-Turbo* in Kazakh prompts and *Llama3* in English prompts performed a little worse in few-shot mode. This suggests that smaller or less robust models sometimes can not handle much information in the prompts, especially if they are not trained to handle long in-context examples effectively. On the other hand, *Gemini flash/pro* and *GPT-4o* models in both language prompting benefited more, showing significant difference between the prompting methods. This shows that state-of -the-art and private models are more robust to in-context learning and that they are trained well to handle longer context with examples.

In total, it can be concluded that H2 (Few-shot prompting outperforms zero-shot) is valid and supported. The practical insight of this experiment was that even a small set of provided examples can meaningfully boost performance of models without fine-tuning which is important in low-resource contexts where labeled data is scarce.

7.4 English vs Kazakh Prompting

The effect of prompt language was more mixed. In most cases it showed slight improvement in accuracy and *Gemini flash* with few-shot settings presented 10 percent improvement in accuracy. However, *Gemini flash* and *GPT-3.5-turbo* with zero-shot settings showed noticeable accuracy drops. On average, English instructions gave a small improvement (around 1 percent), but the differences were not large and consistent across all models.

The reason why English instructions should help in this task in Kazakh was that most instruction tuning of these LLMs done in English (Zhao et al., 2023). This means models may understand task instruction more clearly in English. However, in this experiment examples themselves given in Kazakh and models still need to process Kazakh sentences either way. This could be the explanation for why there is not a big gap between English and Kazakh prompts. So, results show that the choice of language for instruction matters less than provided good examples in most LLM cases.

Therefore, it is hard to support H3 (English prompts outperform Kazakh prompts) as evidence provided is not enough in this study case.

7.5 Error Analysis

The error analysis itself will help us to see *what kinds* of mistake models tend to make. Here quantitative and qualitative error analysis results will be discussed.

7.5.1 By Part of Speech

Nouns were generally easier for models to handle, especially in few-shot settings compared to adjectives and verbs. This could be because nouns often have more concrete meanings and clearer contextual cues. In this experiment only 30 adjective examples were used and from this I got that adjectives were the hardest to handle. This is likely because in Kazakh adjectives have figurative or evaluative meanings (like "cold" as the temperature vs as unfriendly behaviour). Verbs were in the middle, but also challenging as they have different endings and metaphorical uses that require deeper understanding.

7.5.2 Confusion Patterns

From the confusion matrix tables we can see the models' personalities. *GPT-4o* was very conservative and it rarely made false positive predictions. However, this cost of missing many true cases which lead to high false negatives. This shows that GPT-4o needs strong evidence before predicting the "Ture" label. In contrast, *GPT-3.5-Turbo* and *Llama3* were unstable. The *Gemini* models were balanced, improving recall in few-shot while increasing slightly false positives.

7.5.3 Common Error Types

The two common qualitative error types identified in Chapter 6 will be interpreted here in light of how LLMs are trained and what challenges appear for Kazakh language.

Morphological confusion. Models often struggle when the same root word appears in different inflected or derived forms. This could be because most LLMs are trained on relatively simple morphology languages such as English. In the case of Kazakh, a single root word *бас* ("head") can have many surface forms with case markers, possessive endings or derivational suffixes. If the training data contains limited coverage of such forms, models will tend to treat them as identical ignoring morphological changes. This will lead to more false positives where humans would easily identify the distinction. In other words, it could suggest that Kazakh models are not "morphology-aware" and they collapse subtle differences made by suffixes.

Figurative vs. literal usage. Another common error was figurative usage of the words in context. Models often trained on the more default literal meaning of the word missing idiomatic or cultural uses. For example, adjective like *суық* ("cold") are interpreted literally as temperature, rather than figurative "unfriendly". This behaviour could be from frequency bias in model training. Moreover, figurative senses of the words are often culture-specific and if the model has not seen enough Kazakh idioms in pre-training, it cannot easily generalise. Therefore, these explain why figurative cases were the hardest ones for all model families.

Summary. Taken together all these error patterns show that models not only struggle by the limited amount of Kazakh data in their training corpora, but they also have problems with structural factors like morphological modelling and bias from training corpora. These weaknesses explain systematic mistakes observed in Chapter 6 and show a clear direction for improvement.

7.6 Implications for Low-Resource NLP

The results of this dissertation highlights these three broader lessons:

1. **Few-shot prompting works.** Even without fine-tuning, providing examples enhances the performance of the LLMs. For low-resource languages, careful prompt design is a practical way to improve the accuracy.

2. **Morphology challenge.** Agglutinative languages like Kazakh tend to have many forms of the same root which makes it challenging for models

to disambiguate them. Without explicit training or other methods to solve this, LLMs will collapse these forms together.

3. **Instruction language matters, but only slightly.** English instruction is helpful sometimes because of instruction-tuning bias, but Kazakh examples remain essential during prompting.

4. **LLMs are not yet human-competitive.** After seeing results of evaluation, it is hard to fully rely on them and use them for downstream applications like translation or education tools.

For practitioners, all of these means that LLMs can be used for Kazakh tasks, but only with careful prompt design and with an awareness of their limitations. Also, to test them thoroughly before releasing the product. For researchers, the results suggest that better morphological processing, more Kazakh corpora during training with cultural information and exposure to figurative language are key directions for future research and model development.

7.7 Chapter Summary

Concluding this chapter, LLMs showed a clear weakness on the Kazakh WiC task, giving performance well below humans. The Few-shot prompting method was the best way to boost performance, while English instructed prompts presented only a small average rise in the accuracy. Also, error analysis revealed that models struggled with morphology and figurative expressions. These findings confirm the difficulties of using LLMs on low-resource languages and point to both practical strategies and research direction for future improvements.

Chapter 8

Conclusion

8.1 Summary of the Study

This dissertation research is done to evaluate large language models (LLMs) on the Kazakh Word-in-Context (WiC) task. This study started by exploring a clear gap in existing resources. Despite the popularity of WiC in English and its multilingual versions, no equivalent dataset existed for Kazakh language. To address this, a new WiC-style dataset was manually curated, focusing on nouns, verbs and adjectives. The size of the dataset is 510 entries. With this dataset, the state-of-the-art LLMs (GPT-4o, Gemini flashand Llama3) are systematically evaluated under different prompting configurations (zero-shot vs few-shot, English vs Kazakh). Also, human evaluation was conducted to establish a realistic performance baseline. The research questions addressed by this experiment are:

1. **Can LLMs interpret word meanings in context for Kazakh as accurately as humans?**

2. **Does few-shot prompting outperform zero-shot prompting for Kazakh WiC?**

3. **Does English prompting improve performance compared to Kazakh prompting?**

All of these are carefully tested and analysed. Then, respective answers were given.

8.2 Key Findings

The results of the study answered research questions and hypotheses and highlighted three main insights.

- **LLMs lag behind humans.** The best performance of LLMs reached around 77% (Gemini 1.5 Pro) while human annotators achieved over 93% accuracy. This shows that there is a substantial gap between human semantic understanding and modern LLMs abilities to do it in Kazakh.
- **Few-shot prompting works.** Across most modes, providing examples during prompting showed an average 6% accuracy improvement which tells that in-context learning is a practical strategy for low-resource languages.
- **Prompt language has mixed effects.** English instructions showed a slight overall improvement. Which suggests that there is no big difference in language choice and rather examples play a crucial role in increasing of the accuracy of the models.

8.3 Contributions

This dissertation makes three contributions:

1. Created first Kazakh WiC dataset and released openly for future research.
2. A systematic evaluation of different types of LLMs (open and closed) on Kazakh word sense distinction which provides a much-needed baseline.

3. Prompting strategy insight and error patterns like morphological complexity of Kazakh words and their figurative usage which can be a future guide for both researchers and practitioners.

8.4 Limitations

Like all studies this research has several limitations that should be acknowledged. First, dataset size (510 instances) is relatively small compared to high-resource benchmarks. Second, only two human annotators were involved during annotation, which is sufficient for establishing reliable dataset, but limits the diversity of perspectives. Third, not all modern LLMs are used during evaluation. Furthermore, the study was restricted to evaluation without fine-tuning models, meaning the results capture only instruction-following capabilities.

8.5 Future Work

Despite findings of this study, future research can expand this work in several directions. Increasing the size of the Kazakh WiC dataset will improve robustness and allow fine-grained linguistic analysis. Involving more annotators would strengthen the human baseline. Also, fine-tuning models on Kazakh corpora or integrating morphological-aware tokenisation could reduce errors. Moreover, adding more examples in a few-shot prompting could lead to another performance of the models. Evaluating on other open source models like DeepSeek, Mistral and closed sourced models like Claude Sonnet 4, GPT-5, Gemini-2.5-pro/flash could be done as the extension of this study. Using span indices during prompting could be tested for its effect on results. Finally, future research could be done on extending evaluation beyond WiC to related tasks such as idiom detection would provide a better semantic understanding in Kazakh.

8.6 Closing Remarks

This study illustrates limits of the recent LLMs for low-resource, morphologically rich languages by doing experiments on sense understanding tasks. While LLMs have shown good progress and performed well on the WiC task, they remain far from human-level performance in Kazakh. Along with this, contribution of this dissertation is introduction of the new Kazakh WiC dataset and providing a foundation for future research in this direction. More generally, these findings remind all of us that advanced LLMs must be inclusive of all languages. Therefore, by creating resources and evaluating models, we can come closer to a future where AI technologies support all languages equally. This will vanish inequality in languages and scarcity of materials in low-resource languages.

Chapter 9

Reflection and Learning

9.1 Skills and Personal Development

This project was really challenging, but the reward of finishing after cost it. While writing this project, I strengthened not only my technical skills, I developed analytical, personal and communications skills also. All of this will serve me in the future research and professional work.

From technical and analysis skills more precisely I gained first the deeper expertise in designing controlled experiments for evaluating LLMs on low-resource tasks. Secondly, learned building the Kazakh WiC dataset from scratch. I developed an understanding of the data selection criteria and learned how to validate it after collection. Third, this experiment enhanced my prompt engineering skills by learning and analysing various zero-shot vs few-shot methods. Finally, analysing errors and results of the experiment taught me the way of writing with evidence and increased my analytical skills.

Another important thing which I learned was managing time effectively. When I started, the creation and annotation of the dataset seemed like it would be finished in a short time, but it took more than expected. To have a clear vision and manage the research more effectively, I started using the *Notion* tool which helped me to organise tasks better by making to-do lists.

Reading and analysing already existing works increased my ability to find research gaps and evaluate methods carefully. Writing each chapter and applying my supervisor's feedback strengthened my confidence in expressing complex ideas clearly. I feel that after finishing this project I had not only deeper technical knowledge but also stronger habits of clear thinking and academic writing.

9.2 Project Management Insights

Here, I will talk thoroughly about project management. Managing this dissertation taught me how important is planning and structuring the work. As I said before, I used a Notion tool for writing my plans. Also, I used Github for managing my code/dataset and to make it open-source. I divided the project into main phases like literature review, dataset creation, experiments and writing. This gave me direction and milestones. Regular meetings and communication through email with my supervisor helped me to apply feedback and motivated me to improve my work every time. Preparing for these meetings and writing proper email ensured they were productive and helped me to focus on more important things.

At the same time, I learned that plans should not always be kept strictly and it can be flexible. In some stages like cleaning and balancing the dataset took longer than I expected. From this I learned that adjusting timeline and priorities is also important. This balance between structure and adaptability was one of the important project management skills I gained.

9.3 Research Process Reflections

Working on a low-resource language evaluation showed me how to be a real researcher. I realised how limited resources exist for the Kazakh language and how much effort is required to build a reliable dataset. This research taught me that research in low-resource almost in all cases requires creation of a new benchmark or new tool to test the modern LLMs.

Also, precisely stating the research questions and hypotheses in the beginning helped me to clearly organise the structure of the report. Along with this, it was beneficial for connecting results and giving more throughout discussion. I also understood the importance of feedback and iteration. After every comment from my supervisor, whether it was about research questions, methodology, or writing style, I always improved my work. Although it sometimes was demanding, I saw how much better the final dissertation became through the process.

Perhaps the most important insight was learning persistence. There were times when results did not meet the expectations or when progress was so slow. Instead of being discouraged, I learned to keep moving, adapt and focus on improvement. These all things are new insights which I will apply for future projects.

In conclusion, this dissertation not only gave me academic expertise but also influenced personal growth. I learned how to manage time and project effectively, how to research and write so that it is understandable for all readers and how to be patient when there are challenges. Also, knowledge gained during Masters study was very helpful to use all necessary tools and achieve these results. These lessons go beyond this work and will shape my approach for future research and professional work in the future.

References

Ahuja, K., H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Aiyer, N. Mohamed, et al. (2023). "MEGA: Multilingual Evaluation of Generative AI". In: *arXiv preprint arXiv:2303.12528*. Available at: `https://aclanthology.org/2023.emnlp-main.258.pdf`.

Apple Inc. (2025). *Oxford English Kazakh Dictionary*. Version 15.5. Accessed via the macOS Dictionary application. Cupertino, CA.

Belinkov, Y. and J. Glass (2019). "Analysis methods in neural language processing: A survey". In: *Transactions of the Association for Computational Linguistics* 7, pp. 49–72. Available at: `https://arxiv.org/abs/1812.08951`.

Bender, E. M. and B. Friedman (2018). "Data statements for natural language processing: Toward mitigating system bias and enabling better science". In: *Transactions of the Association for Computational Linguistics* 6, pp. 587–604.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33, pp. 1877–1901. Available at: `https://arxiv.org/abs/2005.14165`.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46.

Dong, Q., L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, and Z. Sui (2024). "A Survey on In-Context Learning". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Miami, Florida, USA: Association for Computational Linguistics, pp. 1107–1128. Available at: https://aclanthology.org/2024.emnlp-main.64/.

Drummond, C. (2009). "Replicability is not reproducibility: nor is it good science". In: *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th International Conference on Machine Learning (ICML)*. Montreal, Canada: Citeseer.

Gemini Team (2024). "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context". In: *arXiv preprint arXiv:2403.05530*. Available at: https://arxiv.org/abs/2403.05530.

Haddow, B., R. Bawden, A. V. Miceli Barone, J. Helcl, and A. Birch (2022). "Survey of Low-Resource Machine Translation". In: *Computational Linguistics* 48.3, pp. 673–719. Available at: https://direct.mit.edu/coli/article/48/3/673/111479/Survey-of-Low-Resource-Machine-Translation.

Hu, J., S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson (2020). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR. Available at: https://proceedings.mlr.press/v119/hu20b.html.

Huang, E. H., R. Socher, C. D. Manning, and A. Y. Ng (2012). "Improving Word Representations via Global Context and Multiple Word Prototypes". In: *Proceedings of ACL*. Jeju, Korea: Association for Computational Linguistics, pp. 873–882. Available at: https://nlp.stanford.edu/pubs/HuangACL12.pdf.

Joshi, P., S. Santy, A. Budhiraja, K. Bali, and M. Choudhury (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of ACL*, pp. 6282–6293. Available at: https://aclanthology.org/2020.acl-main.560/.

Kazakh National Corpus (2025). *Almaty Corpus of Kazakh Language (NCKL)*. http://web-corpora.net/KazakhCorpus/search/. Accessed: 22 August 2025.

Liang, Y., N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, X. Fan, R. Zhang, and R. M. Agrawal (2020). "XGLUE:

A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 6008–6018. Available at: https://aclanthology.org/2020.emnlp-main.484.

Liu, N. F. et al. (2023). “Languages are not equally represented in multilingual models: Improving equitable language modeling via adaptive sampling”. In: *Transactions of the Association for Computational Linguistics* 11, pp. 1–16. Available at: https://aclanthology.org/2023.tacl-1.1/.

Makhambetov, O., A. Makazhanov, Z. Yessenbayev, I. Sabyrgaliyev, and I. Sharafudinov (2013). “Assembling the Kazakh Language Corpus”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1022–1031. Available at: https://aclanthology.org/D13-1104.pdf.

Maxutov, A., A. Makazhanov, Y. Issabekov, and et al. (2024). “Do LLMs Speak Kazakh? A Pilot Evaluation of Seven Models”. In: *Proceedings of the First Workshop on Turkic and Languages in Contact with Turkic (SIGTURK 2024)*. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.sigturk-1.8/.

Min, S., M. Lewis, L. Zettlemoyer, and H. Hajishirzi (2022). “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?” In: *Proceedings of EMNLP*. Available at: https://arxiv.org/abs/2202.12837.

Nangia, N., C. Vania, R. Bhalerao, and S. R. Bowman (2019). “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1953–1967.

Ollama Team (2025). *Ollama: Run Large Language Models Locally*. https://ollama.com/. Accessed: 25 August 2025.

OpenAI (2023). *GPT-4 Technical Report*. https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2025-08-21.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pilehvar, M. T. and J. Camacho-Collados (2019). "WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations". In: *Proceedings of NAACL-HLT*. Minneapolis, MN: Association for Computational Linguistics, pp. 1267–1273. Available at: `https://aclanthology.org/N19-1128/`.

Ponti, E. M., H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen (2019). "Modeling Language Variation and Universals: A Survey on Typological Linguistics for NLP". In: *Proceedings of ACL*. Available at: `https://aclanthology.org/J19-3005/`.

Qazcorpus (2025). *Qazcorpus: National Corpus of the Kazakh Language*. `https://qazcorpus.kz/`. Accessed: 22 August 2025.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning (2020). "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pp. 101–108.

Raganato, A., T. Pasini, J. Camacho-Collados, and M. T. Pilehvar (2020). "XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization". In: *Proceedings of EMNLP*. Association for Computational Linguistics, pp. 7193–7206. Available at: `https://aclanthology.org/2020.emnlp-main.584/`.

Sozdik.kz contributors (2025). *Sozdik.kz: Online Kazakh-Russian Dictionary*. `https://sozdik.kz`. Accessed: August 2025.

SuperGLUE Benchmark (2019). Available at: `https://super.gluebenchmark.com/` (Accessed:

Togmanov, M., N. Mukhituly, D. Turmakhan, J. Mansurov, M. Goloburda, A. Sakip, Z. Xie, Y. Wang, B. Syzdykov, N. Laiyk, A. F. Aji, E. Kochmar, P. Nakov, and F. Koto [2025]. "KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), Long Papers*. Available at: `https://aclanthology.org/2025.acl-long.701/` [Accessed:

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman (2019). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *arXiv preprint arXiv:1905.00537*. Available at: `https://arxiv.org/abs/1905.00537` (Accessed:

Wiktionary ■[2025■]. *Wiktionary, the free dictionary.* Accessed: 22 August 2025. Available at: `https://www.wiktionary.org/` ■[Accessed:

Zhao, T., J. Li, D. Guo, H. Ma, J. Sun, and J. He ■(2023■). “Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual Thought Prompting”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Accessed: 22 August 2025. Association for Computational Linguistics, pp. 11142–11154. Available at: `https://aclanthology.org/2023.findings-emnlp.826.pdf` ■(Accessed:

Appendix A

Prompt templates in English

Kazakh sentences are replaced with description.

Prompt template: zero-shot & English

```
You are given two sentences that both contain the same target word.
Your task is to determine whether the word has the same meaning in both sente
If the meaning is the same, answer "True".
If the meaning is different, answer "False".
Respond with only "True" or "False".

Sentence 1: {sentence1}
Sentence 2: {sentence2}
Word: {target_word}
Answer:
```

Prompt template: zero-shot & English

```
You are given two sentences that both contain the same target word.
Your task is to determine whether the word has the same meaning in both sente
If the meaning is the same, answer "True".
If the meaning is different, answer "False".
```

```
Respond with only "True" or "False".

Examples:

Sentence 1: {First sentence in Kazakh}
Sentence 2: {Second sentence in Kazakh}
Word: {kazakh target word}
Answer: False

Sentence 1: {First sentence in Kazakh}
Sentence 2: {Second sentence in Kazakh}
Word: {kazakh target word}
Answer: True

Now your task:
Sentence 1: {sentence1}
Sentence 2: {sentence2}
Word: {target_word}
Answer:
```

Appendix B

Other results tables

Here I included additional results tables which are not used during analysis for future research
Macro and Weighted F1 scores.
Precision and Recall results for each model:

Model	**Prompt Mode**	**Language**	**Weighted F1**	**Macro F1**
gemini-1.5-flash	zero	kk	53.66%	53.86%
gemini-1.5-flash	few	kk	62.76%	62.89%
gemini-1.5-flash	zero	en	55.92%	56.10%
gemini-1.5-flash	few	en	75.68%	75.72%
gemini-1.5-pro	zero	kk	66.21%	66.32%
gemini-1.5-pro	few	kk	75.12%	75.07%
gemini-1.5-pro	zero	en	63.19%	63.32%
gemini-1.5-pro	few	en	77.02%	77.03%
gpt-4o	zero	kk	58.88%	59.04%
gpt-4o	few	kk	72.05%	72.13%
gpt-4o	zero	en	63.89%	64.02%
gpt-4o	few	en	74.25%	74.32%
gpt-3.5-turbo	zero	kk	53.11%	53.07%
gpt-3.5-turbo	few	kk	50.78%	50.89%
gpt-3.5-turbo	zero	en	43.37%	43.16%
gpt-3.5-turbo	few	en	51.70%	51.60%
llama3	zero	kk	40.75%	40.43%
llama3	few	kk	49.39%	49.21%
llama3	zero	en	46.10%	45.85%
llama3	few	en	43.88%	44.15%

Table B.1: Weighted F1 and Macro F1 for Models

Model	**Prompt Mode**	**Language**	**Precision**	**Recall**
gemini-1.5-flash	zero	kk	93.65%	22.87%
gemini-1.5-flash	few	kk	93.88%	35.66%
gemini-1.5-flash	zero	en	93.06%	25.97%
gemini-1.5-flash	few	en	89.14%	60.47%
gemini-1.5-pro	zero	kk	93.81%	41.09%
gemini-1.5-pro	few	kk	91.41%	57.75%
gemini-1.5-pro	zero	en	94.90%	36.02%
gemini-1.5-pro	few	en	80.00%	72.87%
gpt-4o	zero	kk	98.68%	29.07%
gpt-4o	few	kk	97.69%	49.22%
gpt-4o	zero	en	95.96%	36.82%
gpt-4o	few	en	96.50%	53.49%
gpt-3.5-turbo	zero	kk	53.45%	60.08%
gpt-3.5-turbo	few	kk	55.56%	33.95%
gpt-3.5-turbo	zero	en	49.64%	79.46%
gpt-3.5-turbo	few	en	52.77%	70.16%
llama3	zero	kk	51.66%	96.51%
llama3	few	kk	52.72%	82.56%
llama3	zero	en	52.85%	93.41%
llama3	few	en	74.42%	12.40%

Table B.2: Precision and Recall for Models

Appendix C

Annotation guidelines

During annotation each annotator asked to do:

1. Read the pair of sentences very carefully
2. Decide whether target word has the same meaning in both sentences or not
3. Label "True" if has the same meaning, "False" in other case

Instruction were simple to avoid bias during judgment. Annotators prohibited from using external tools like dictionaries. Datset was adjudicated after annotation and very ambiguous cases removed.
Dataset follows ethical and licensing guidelines in Chapter 5 and is released as open-source for academic and practical use.