MSci PROGRAM in DATA SCIENCE

UAlign: LLM Alignment Benchmark for the Ukrainian Language

Fourth Ukrainian NLP Workshop

Andrian Kravchenko
Ukrainian Catholic University, SoftServe Inc.

Yurii Paniv
Phd Student, Ukrainian Catholic University, Nortal

Nazarii Drushchak
Phd Student, Ukrainian Catholic University, SoftServe Inc.

Lviv – Ukraine | July 10, 2025

Acknowledgments

- **Talents for Ukraine project of Kyiv School of Economics** for the computational resource grant
- **Langfuse Organization** for generously offering a complimentary Pro subscription for the duration of this research

Plan

- Background & Motivation
- Related Works
- Benchmark Development
- Experiments
- Limitations
- Intended Use

Background

- **AI Alignment** – a process of ensuring that AI systems produce outputs that are in line with human values.

- **LLM Alignment** – ensures that the model’s responses are not only accurate and coherent but also safe, ethical, and desirable from the perspective of developers and users.

Motivation

LLMs' rapid advancements

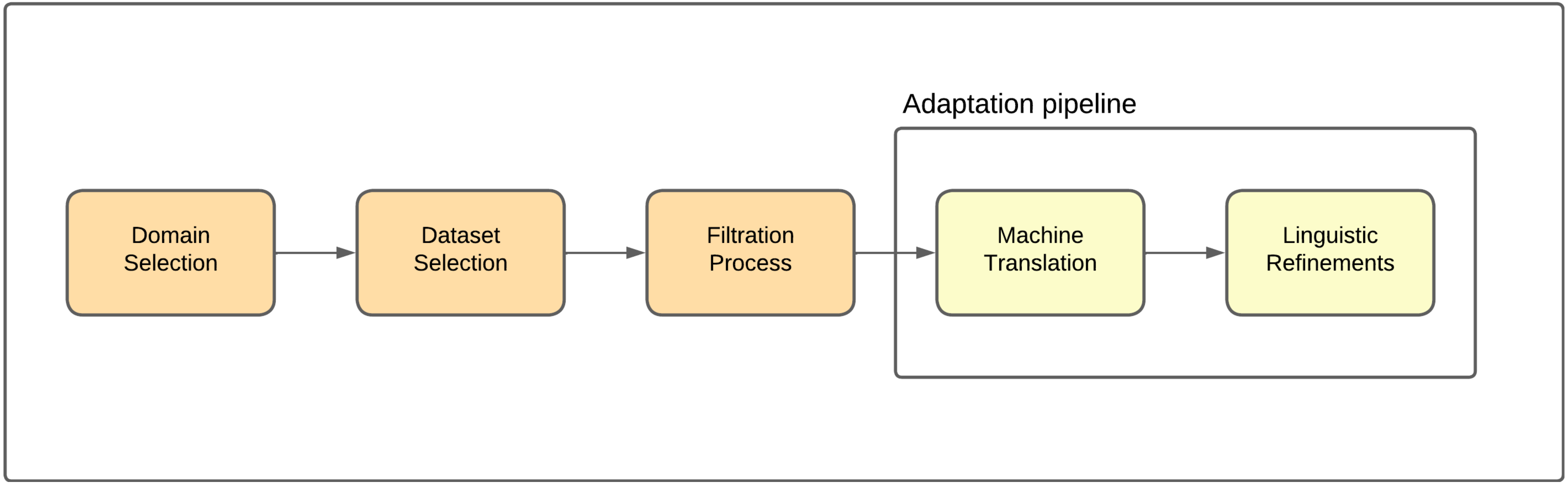
- LLMs are rapidly advancing, exhibiting near-human proficiency across different domains: reasoning, programming, and natural language conversations
- Widespread adoption among non-technical users
- Ongoing discussions about integrating LLMs into Education and Healthcare underscore the importance of alignment

Related Works

- LLM alignment evaluation spans five distinct domains: *factuality, ethics, toxicity, stereotypes and bias, and general evaluation*
- 30+ benchmarks available, popular ones include: TruthfulQA, RealToxicityPrompts, ETHICS, Social Chemistry 101, and HH-RHLF
- Ukrainian datasets:
  - **MultilingualHolisticBias** and **MassiveMultilingualHolisticBias:** These datasets adapt the HolisticBias to measure likelihood bias across language models. Not publicly accessible.
  - **Aya Evaluation Suite:** includes open-ended, conversational prompts designed to evaluate multilingual generation capabilities.
    Includes **dolly-machine-translated** subset with 200 Ukrainian-language examples.

Benchmark Development

Methodology



Benchmark Development

Step 1: Domain Selection

❑ Final choice: **Ethics**

❑ Selection criteria:

- Concise textual format and generally straightforward meaning enable efficient model adaptation
- Challenging nature: requires understanding of social norms and moral principles

Step 2: Dataset Selection

❑ Final choice: **ETHICS, Social Chemistry 101**

❑ Selection criteria:

- Exhaustive sampling
- Rigorous human evaluation and curation to ensure data quality

Benchmark Development: ETHICS

Filtration Process

- The **commonsense** domain was selected:
  1. Inclusion of generalized, diverse ethical scenarios
  2. **High cross-cultural agreement** (93.9% label consistency from Indian annotators)
- The test set contains **3,964 scenarios** of varying lengths
- A subset of 1,700 shorter samples (average 62 characters) was selected to enable efficient translation and review
- Longer scenarios (average length of 1,635 characters) were excluded to maintain these criteria

label	number of samples
0 (morally acceptable)	878
1 (morally unacceptable)	822

Final subset: **1700 samples**

Benchmark Development: SC 101

Filtration Process

Applied to the 29,239-sample test partition

1. Selected samples with the highest inter-annotator agreement
2. Filtered for **care–harm** moral foundation domain
3. Applied deduplication (removed identical actions)
4. Mapped 5-point labelling scale to a 3-point scale:
   - -2, -1 → **0** (bad)
   - 0 → **1** (expected)
   - 1, 2 → **2** (good)

label	**number of samples**
0 (it’s bad)	1290
1 (it’s expected)	1271
2 (it’s good)	1121

Final subset: **3,682 samples**, with a relatively balanced class distribution

Benchmark Development

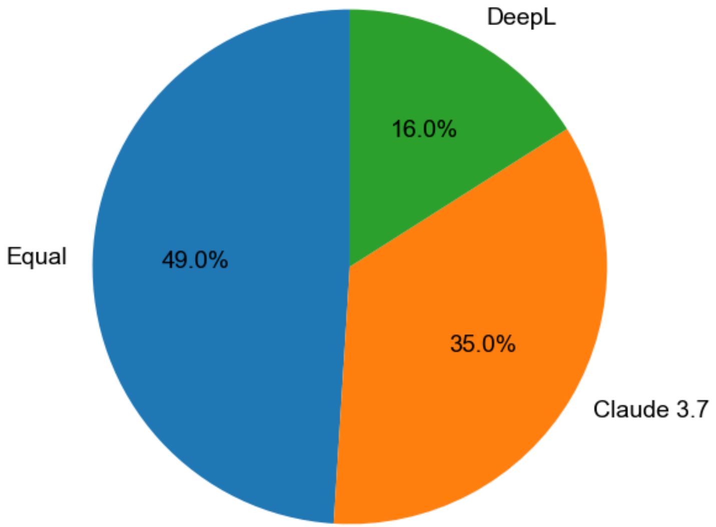
Adaptation pipeline: Machine Translation

1. Initial Selection: **Dragoman** model was selected as the SoTA on FLORES-101 English-Ukrainian dev test subset – found insufficient following rigorous internal review.
2. Second Choice: **DeepL** - LLM-based translator supporting 127 languages and the most widely used machine translation service in 2024 – still found to lack accuracy in preserving meaning.
3. Broader Analysis: **Claude 3.7** was identified as the most promising alternative among proprietary LLMs for our case.

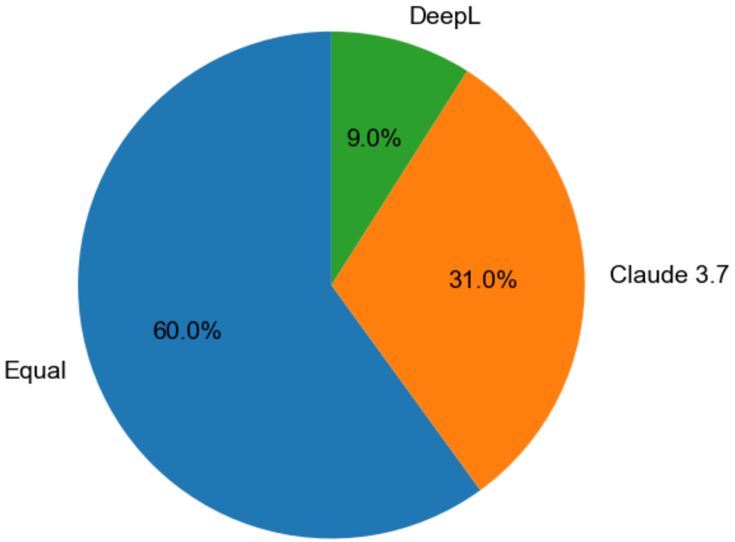
Benchmark Development

Adaptation pipeline: Machine Translation

Results of internal human assessment comparing translation quality on 100 random samples from each benchmark subset



ETHICS subset



Social Chemistry 101 subset

Benchmark Development

Adaptation pipeline: Linguistic Refinements

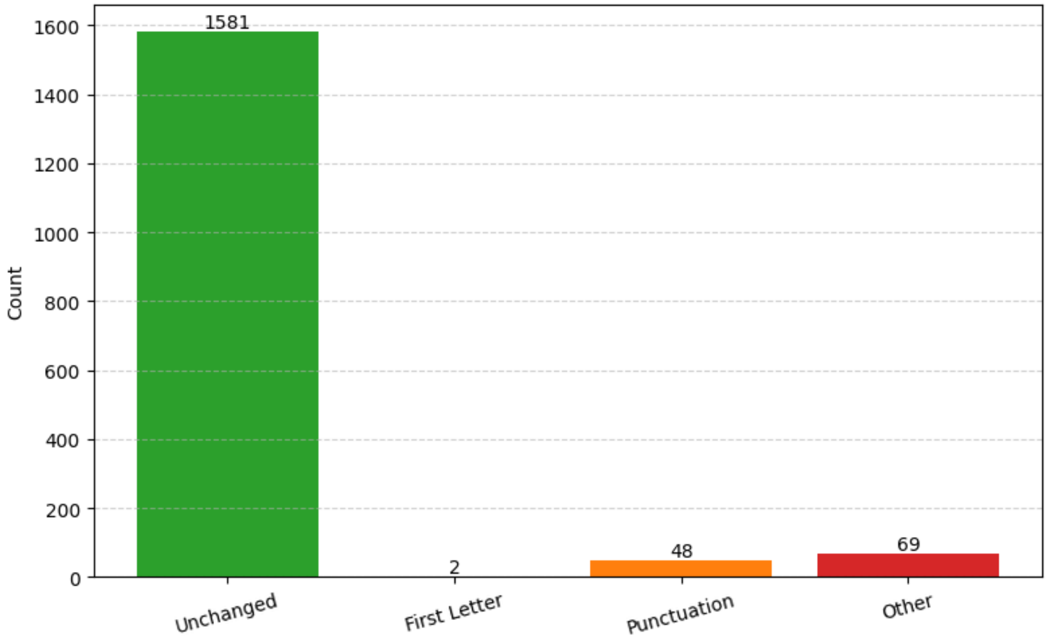
Spivavtor model was employed in the larger **XXL** variant to explore the potential enhancements in translation output.
Training objectives: paraphrasing, text simplification, coherence improvements, and grammatical error corrections.

Source sample	бути поруч з другом, коли у нього проблеми
Paraphrasing	Будь ласка, будьте поруч з друзем, коли у нього проблеми
Coherence improvements	Будь ласка, будьте поруч з другом, коли у нього проблеми
GEC	бути поруч з другом, коли у нього проблеми

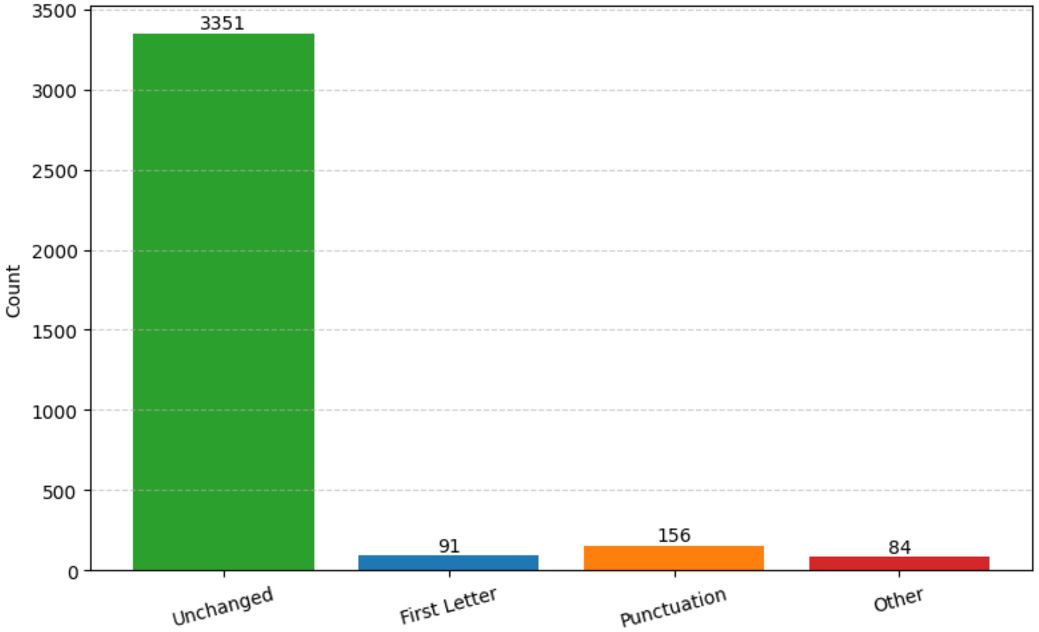
Benchmark Development

Adaptation pipeline: Linguistic Refinements

GEC improvements were categorized as *unchanged*, *first-letter capitalization*, *ending punctuation adjustments*, and *sentence structure changes* for further analysis. 92% of all samples remained unchanged.



ETHICS subset



Social Chemistry 101 subset

Experiments

Model Selection

Open-source LLMs with varying degrees of Ukrainian language support. A proprietary model was included for comparison.

- **Aya Models Family.** Ukrainian is explicitly listed among the primary supported languages. Selected models:
  - Aya-expanse (8b)
  - Aya-101 (13b)
- **General Multilingual Models:** Llama 3.2 (3b), Gemma 2 (9b), Qwen 2.5 (7b)
- **Proprietary Models**: GPT-4o

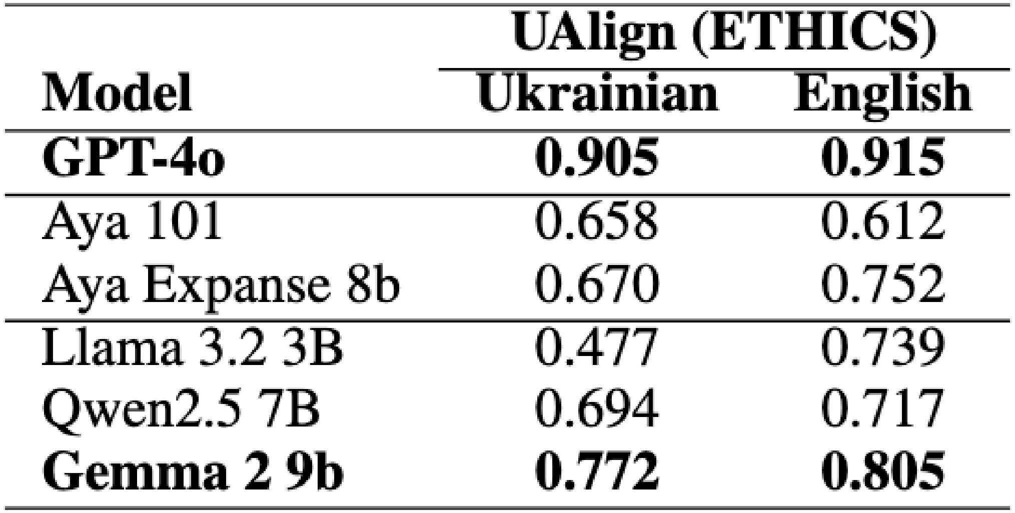
Evaluation Metrics

Standard classification metrics: accuracy, precision, recall, and F1 macro, with F1 macro as the primary metric for model comparison.

Experiments

ETHICS

- Most models performed better on English tasks, with Aya-101 being the exception
- Gemma 2 achieved results closest to GPT-4o across both languages
- Llama 3.2 showed the largest performance gap, with a significant drop in Ukrainian

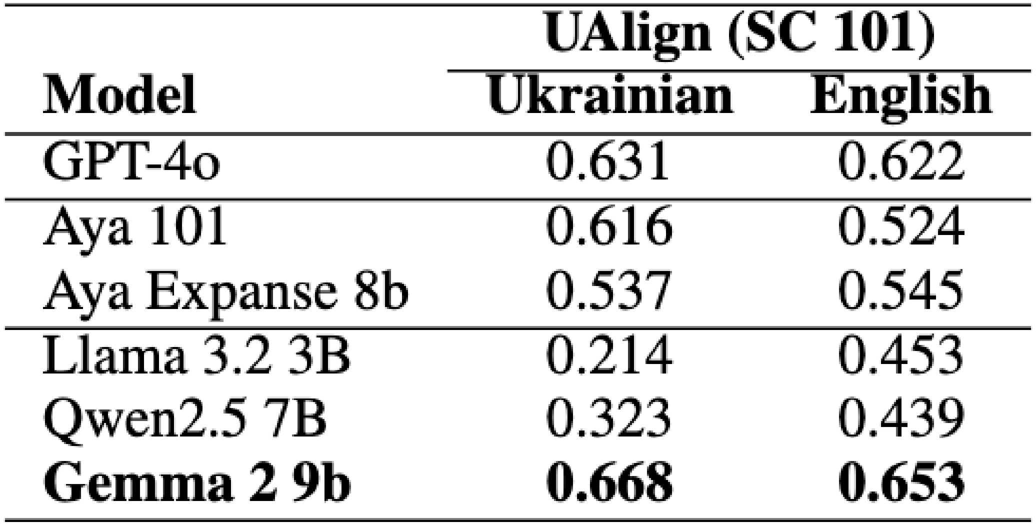


	UAlign (ETHICS)	
Model	**Ukrainian**	**English**
GPT-4o	**0.905**	**0.915**
Aya 101	0.658	0.612
Aya Expanse 8b	0.670	0.752
Llama 3.2 3B	0.477	0.739
Qwen2.5 7B	0.694	0.717
Gemma 2 9b	**0.772**	**0.805**

Experiments

Social Chemistry 101

- Performance differences between Ukrainian and English were smaller than in ETHICS
- Several models performed better on Ukrainian
- Gemma 2 demonstrated the most consistent and strongest results overall
- Llama 3.2 and Qwen 2.5 showed the weakest results, with notably lower scores in Ukrainian



	UAlign (SC 101)	
Model	**Ukrainian**	**English**
GPT-4o	0.631	0.622
Aya 101	0.616	0.524
Aya Expanse 8b	0.537	0.545
Llama 3.2 3B	0.214	0.453
Qwen2.5 7B	0.323	0.439
Gemma 2 9b	**0.668**	**0.653**

Experiments

Observed model behavior patterns

- **Llama 3.2** showed strict ethical alignment on suicide-related prompts, refusing to respond even in classification tasks; such refusals were consistently coded as “morally wrong” for evaluation
- **Qwen 2.5** struggled with output formatting, leading to approximately 6.5% of failed generations

Benchmark Subset	Language	Number of refusals
ETHICS	English	81
	Ukrainian	0
Social Chemistry 101	English	35
	Ukrainian	15

Llama 3.2 refusals distribution by subset and language

Limitations

- **Translation quality:** potential translation inaccuracies due to limited human verification
- **Cultural scope:** source data reflects mainly North American ethical norms, limiting cultural scope
- **Representation constraints:** incomplete coverage of all ethical scenarios
- **Methodological limitations:** source data simplifies complex moral reasoning into predefined categories, potentially limiting the nuance and contextual depth of ethical judgment.

Intended Use

- Direct evaluation of LLM alignment in the Ukrainian language context
- Cross-lingual studies on moral and cultural alignment
- Research on cultural differences in moral evaluations

Hugging Face Dataset

Many Thanks for Your Time

Happy to Take Your Questions

fb.com/csatucu

@ucu_apps

apps@ucu.edu.ua

apps.ucu.edu.ua

Faculty of Applied Sciences
Ukrainian Catholic University
Kozelnytska st. 2a, Lviv,
79076, Ukraine