

Магістерська робота

Автоматичне транскрибування та ідентифікація у змішаному українсько-англійському мовленні

Нікітін Руслан Валерій Умарович

Керівник: д.ф.-м.н., доц. Ігнатенко О.П.

НаУКМА · Факультет інформатики · 2026

Рецензент: Солопова Вероніка Максимівна
доктор філософії з комп'ютерних наук

Проблема: code-switching в IT-середовищі

"Треба пофіксити bug у pipeline перед деплоєм на staging"



Асиметрія алфавітів

Кирилиця ↔ Латиниця:
модель вирішує ЯКОЮ
графікою записати слово



Відсутній датасет

Жодного якісного
відкритого корпусу
для UK-EN пари



Неоптимізовані ASR

Існуючі системи не
адаптовані до внутрішньо-
реченнєвого CS

Існуючі корпуси code-switching: прогалина для UK-EN

Датасет	Мови	Год.	Мовців	Тип
ASCEND	ZH-EN	10,6	23	спонтанне
SEAME	ZH/MS-EN	192	156	спонтанне
Miami Bangor	ES-EN	35	84	спонтанне
FAME!	FR-NL	18,5	~541	broadcast
uk-en-code-mixed-asr-2h ★	UK-EN	2,1	1	начитане

uk-en-code-mixed-asr-2h

448

аудіозаписів

16 IT-категорій

2,1 год

чистого мовлення

~17 с / запис

99,6%

intra-sentential CS

446 з 448 зразків

18,9%

англійських слів

21,9% латиниці

1 251

унікальних EN-токенів

3 207 входжень

HF

публічний доступ

HuggingFace

Формат read speech → абсолютно точна еталонна транскрипція → коректний підрахунок CS-WER на межах перемикання

Методологія оцінювання: CS-WER та ITN



Метрика CS-WER

$$CS-WER = (Sen + Den + Ien) / Nen$$

Ізольована оцінка лише **англійських tokenів** — WER на UK та EN сегментах окремо

Проблема: 1 хибний EN-термін у реченні з 15 слів = WER 6,7%, але CS-WER = 100%



ITN-фільтр

1 240

англійських термінів

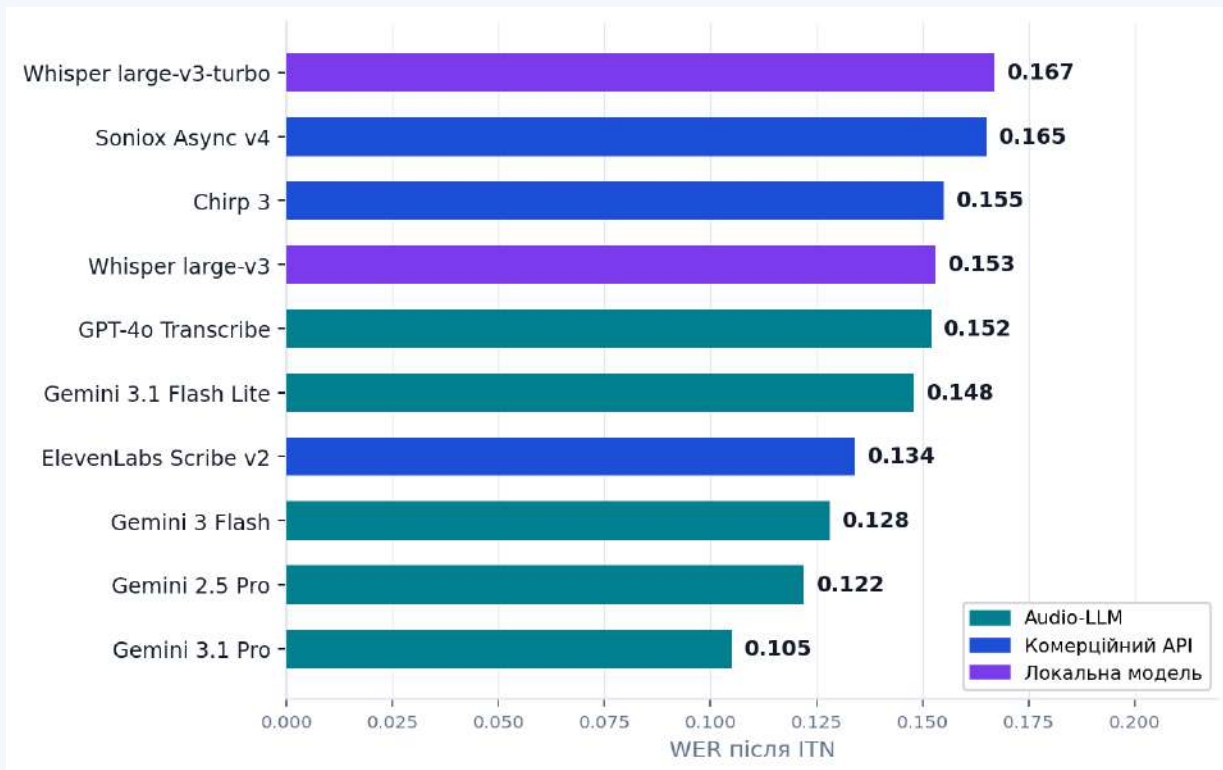
4 838

унікальних пар відповідності

"деплой" → *deploy* | "пайплайн" → *pipeline*

Детермінований алгоритм (regex + інвертований словник)
Формується апіорно — без витоку даних у бенчмарк

Результати: 49 конфігурацій, 3 класи архітектур



UA моделі: WER > 0.47
ридатні без fine-tuning

Асиметрія CS-WER: природа помилок

Модель	EN-WER	EN ITN	Δ EN	UK-WER	UK ITN	Δ UK	WER	WER ITN	Δ WER	CER	CER ITN	Δ CER
Gemini 3.1 Pro	0,4256	0,2442	-0,1815	0,0742	0,0737	-0,0005	0,1359	0,1053	-0,0306	0,0784	0,0443	-0,0342
ElevenLabs Scribe v1	0,7490	0,4528	-0,2962	0,1349	0,1321	-0,0028	0,2427	0,1943	-0,0484	0,1508	0,0941	-0,0567
Whisper large-v3-turbo	0,4487	0,3172	-0,1315	0,1042	0,1039	-0,0003	0,1781	0,1529	-0,0252	0,0894	0,0635	-0,0258
Whisper Large v3 (force_uk)	0,4487	0,3172	-0,1315	0,1042	0,1039	-0,0003	0,1781	0,1529	-0,0252	0,0894	0,0635	-0,0258
Whisper Large v3 (auto)	0,4487	0,3185	-0,1303	0,1073	0,1071	-0,0003	0,1877	0,1633	-0,0244	0,0919	0,0669	-0,0250

Система транскрибування повного циклу: Voice Diary



VAD dual-threshold hysteresis

Onset 0,60 / Offset 0,45
Усуває стрибки пам'яті
Макс. сегмент 13 с



ECAPA-TDNN: попір 0,5

Крос-сесійна ідентифікація
Адаптація до зміни акустики
під час перемикання мов



Захист від галюцинацій Whisper

Blocklist + n-gram penalty
Білий список мов: uk, en
Force-flush при помилці VAD

ВИСНОВКИ

01

Створено відкритий датасет uk-en-code-mixed-asr-2h (448 записів, 2,1 год, HuggingFace)

02

Бенчмарк 49 конфігурацій ASR - перше порівняння для UK-EN (not TTS) пари

03

EN-WER перевищує UK-WER у 3-10 разів для всіх архітектур

04

ITN-фільтр знижує помилки на англіцизмах на 13-30%: орфографічна природа збоїв

05

Voice Diary: VAD-гістерезис + ECAPA-TDNN крос-сесійна ідентифікація