

N8CIR SUMMER INTERNSHIP 2026



Looking at the Right Words

Interpreting Readability Evaluation with
Explainable AI

Owais Mahmood

University of Leeds

jpdj5670@leeds.ac.uk



What Simplification Looks Like

ENGLISH • ORIGINAL

Make sure that we recognise and value the care workforce and unpaid carers



SIMPLIFIED

Help the workforce to feel valued

FRANÇ AIS • ORIGINAL

Faire un CV et le modifier seul



SIMPLIFIED

Faire un CV et le mettre à jour

ARABIC • ORIGINAL

أما في ما بعد، أعترف أنني لم أحتمل، لأحملك، بكل ضعف، المسؤولية
بالكامل



SIMPLIFIED

أما بعد ذلك، أعترف بأنني لم أستطيع

Is the AI looking at the same words we would?

The Gold Standard



1,930

aligned sentence pairs



6

languages — Arabic, Catalan,
English, French, Italian, Spanish



281

pairs in the held-out test set

WHAT A WORD-LEVEL ANNOTATION LOOKS LIKE

ORIGINAL make sure that we recognise and value the care workforce and unpaid carers

SIMPLIFIED help the workforce to feel valued

The Highlighter Pen

PER-WORD
IMPORTANCE

make

sure

that

we

recognise

an

value

the

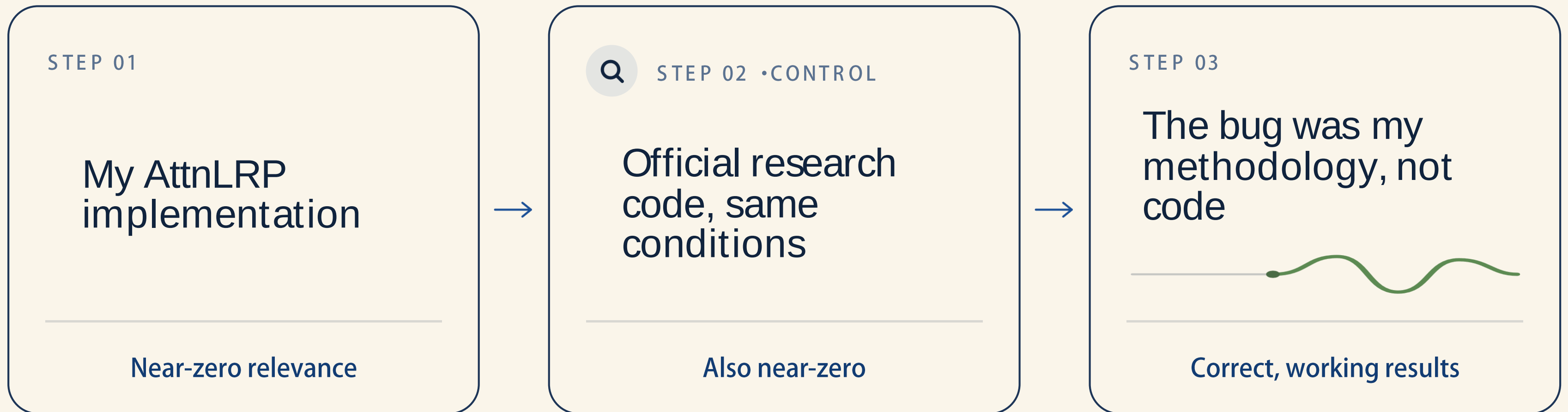
care

workforce

	INTEGRATED GRADIENTS	GRADIENT SHAP	RAW ATTENTION	ATTN LRP
mBERT	✓	✓	✓	—
XLM-R	✓	✓	✓	✓
Multilingual-E5	✓	✓	✓	✓
Qwen	✓ ADAPTE D	—	—	—

12 COMBINATIONS TESTED

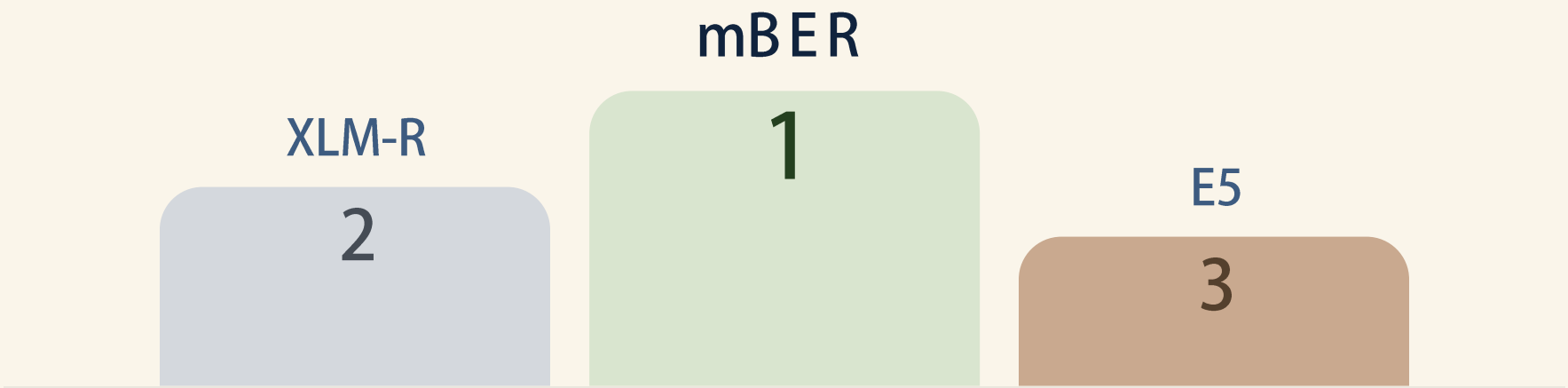
Debugging the Method



	BEFORE	AFTER
TWO LINES, TWO MISTAKES	<pre>target.backward(target) relevance = gradient</pre>	<pre>target.backward() relevance = gradient * input</pre>

ISOLATE BEFORE YOU REWRITE

The Ranking Flips



NAIVE RANKING EACH MODEL AT ITS OWN BEST-SCORING METHOD

1	XLM-R	0.291
2	Multilingual-E5	0.275
3	mBER	0.239

AttnLRP



FAIR RANKING ONE SHARED METHOD — INTEGRATED GRADIENTS

1	mBERT	0.232
2	XLM-R	0.226
3	E5	0.214

Where This Leaves Us

THE QUESTION

Did these models' pay attention to the same things a human editor would?

THE APPROACH

4 models, 4 explainability techniques, tested across 6 languages.

THE KEY FINDING

Comparing fairly matters, for which technique wins, and for which model wins.

NEXT STEPS

Extending to more models and looking closer at specific edit types.

Thank You

SUPERVISED BY

Dr. Nouran Khallaf

Tom Hardcastle

WITH THANKS TO

Sorrel Harriet

Andrew Harvie

**Patricia
Ternes**

N8 CIR UNDERGRADUATE
UNIVERSITY OF LEEDS