AI-ENABLED DIGITAL ACCESSIBILITY

Can Language Models Flag Inaccessible Text?

Developing approaches to automatically **identify and label** the segments of administrative web content that violate cognitive accessibility guidelines to:

- support the professionals who revise public-facing administrative information;
- explore the automation of cognitive accessibility assessment.

CENTRE FOR TRANSLATION STUDIES
UNIVERSITY OF SURREY

1 Background

Technical aspects of web accessibility can be tested automatically, but **cognitive accessibility of the text cannot**.

Readability scores indicate *that* a text is complex, not *what* makes it hard or *where*.

2 The Task

- **The span.** We work at the document level to find and label the exact stretch of text that creates a barrier – the *complex span*.
- **What we are looking for.** Agreement with expert human annotators on both the boundaries and the labels.
- **Why this is hard.** A new task: no no baseline, no established annotation framework, and no labelled (or unlabelled) dataset.

3 Real-world Example

GOV.UK GUIDANCE

Higher education

All universities and higher education colleges UNEXPL should have a person in charge of UNCOM disability issues that you can talk to about the support they AMB offer. The policies are regularly reviewed GRAM by the administration UNEXPL.

LABELS BASED ON ACCESSIBLE LANGUAGE GUIDELINES

HPLU	High-Proficiency Lexical Unit
UNEXPL	Unexplained Term or Entity
AMB	Ambiguous Language
GRAM	Punctuation and Grammar
ABBREV	Abbreviations
UNCOM	Unnecessary Complication

4 Approach

DATA COLLECTION
GOV.UK web content

↓

GROUND TRUTH
Human annotation

↓

BASELINE + SILVER
Prompt proprietary LLMs

↓

DISTIL
Fine-tune open-weight models

5 Results So Far

How aligned are LLM annotations with human labels?

Method	Precision	Recall	F1 Score
zero-shot	0.314	0.166	0.217
zero-shot (fuzzy matches)	0.323	0.170	0.223
three-shot	0.376	0.273	0.317
three-shot (fuzzy matches)	0.374	0.272	0.315
three-shot+description	0.369	0.272	0.313
three-shot+description (fuzzy matches)	0.500	0.368	0.424

22-doc / 560-span ground-truth set.

YES / NO HUMAN VALIDATION

~90%
of flagged spans were valid

K=0.61
inter-annotator agreement for for span validity

DATASET

716 WEB PAGES

12.6K PARAGRAPHS

500K WORDS

ACKNOWLEDGEMENTS
This research is carried out as part of a PhD project within the ADA Leverhulme Doctoral Scholarships Network funded by the Leverhulme Trust.

LEVERHULME TRUST

Supervisory Team:
Constantin Orăsan, Centre for Translation Studies
Diptesh Kanojia, Institute for People-Centred AI

CONTACT
Daria Sokova