Engaging Content
Engaging People

DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Taighde Éireann
Research Ireland

LLMs as Assistants for Data Annotation: Addressing Disagreement and Supporting Expert Processes

Mark Andrade*,[1] Bláithín Heffernan*,[1] Abigail Walsh,[1] Sheila Castilho[2]

*These two authors contributed equally to this work
[1]firstname.lastname@adaptcentre.ie
[2]firstname.lastname@dcu.ie

HOST INSTITUTIONS

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

PARTNER INSTITUTIONS

MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

Maynooth University
National University of Ireland Maynooth

TU Dublin

TUS

University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

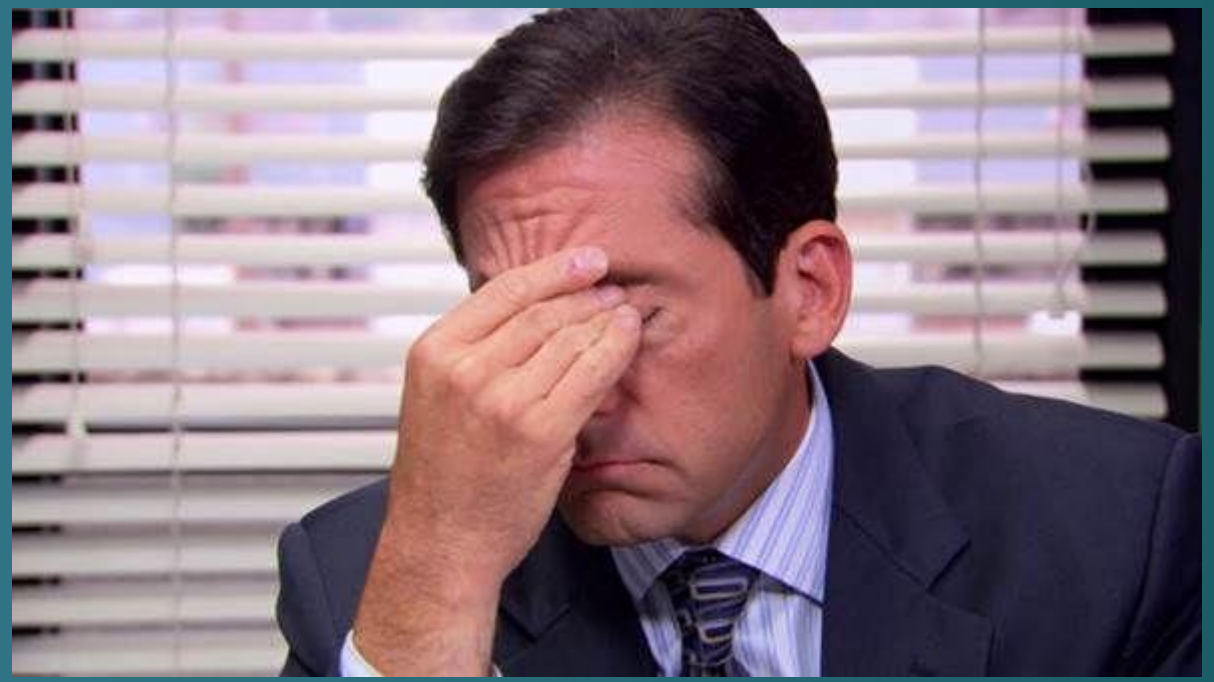
Ollscoil na Gaillimhe
University of Galway

Background & Research Questions

DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

ADAPT
Engaging Content
Engaging People

Data Annotation is a stressful job…



Can LLMs help??

Background & Research Questions

DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

ADAPT
Engaging Content
Engaging People

RQ1: Can LLMs create annotation rules with no guidance from human-generated annotations? If so, how helpful are these rules?

RQ2: How helpful are LLMs in resolving disagreements between human annotators?

RQ3: Can LLMs help with analysing expert human annotators' disagreements and their discussion?

Methodology

Annotation framework

Text can be annotated with multiple dimensions (Castilho and O'Brien 2026):

- **Content**: purpose of end-users and how content is delivered
- **Genre**: conventional structures used to construct a complete text within that variety
- **Text Type**: intratextual or linguistic features
- **Topic**: the thematic content of the document

Methodology

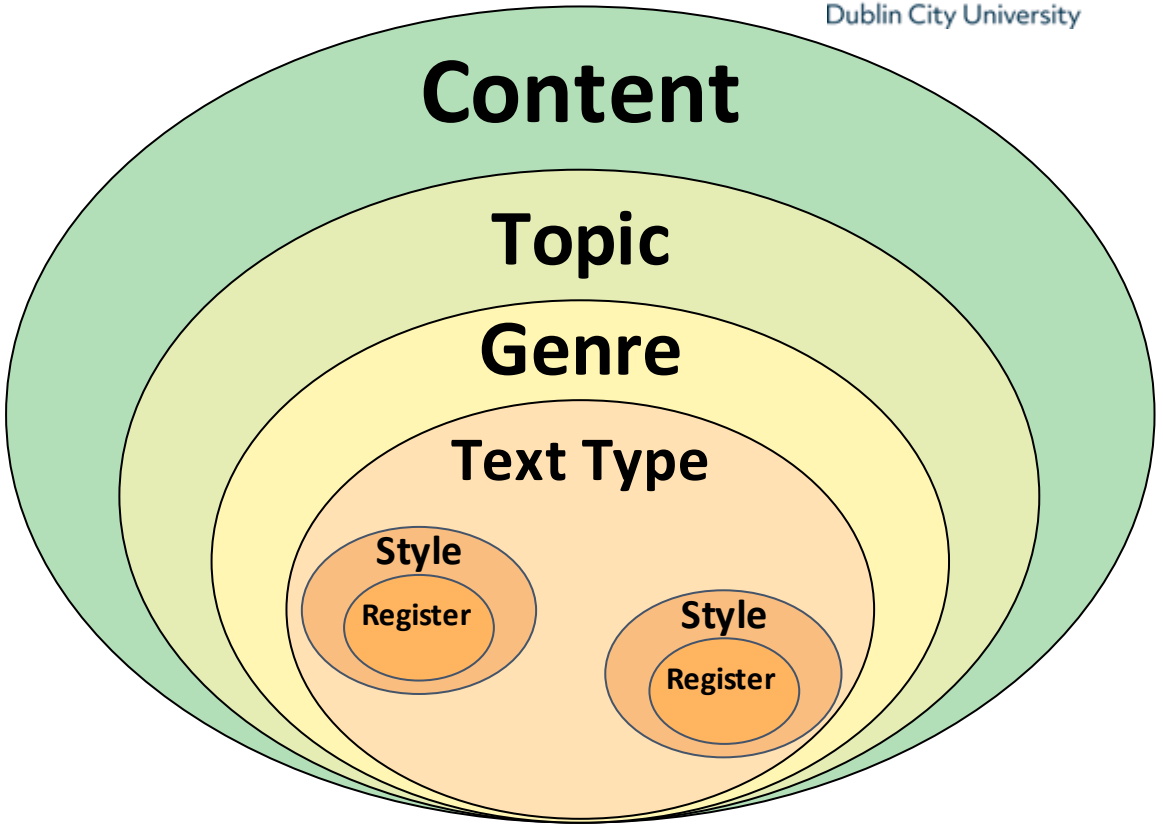


Figure adapted from Castilho and O'Brien (2026)

Methodology

Evaluation

Gius et al. (2019) propose three metrics for annotation guideline evaluation:

- **Coverage** of the theoretical basis
- **Applicability** or how well the guideline prepares annotators to do actual annotations
- **Usefulness** or potential insight provided by the guidelines
    - Inter-annotator agreement (IAA)

Methodology

Mini-Reference & Training

- Four linguists collaborated to create a *mini-reference* to annotate across four labels (**Content**, **Genre**, **Text Type**, **Topic**)
- Two annotators practised applying these initial guidelines to consecutive test datasets, cumulating in Pilot Annotation Task
    - More details in Experiment 3

DCU
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

ADAPT
Engaging Content
Engaging People

Mini-Reference & Training

B	C	D	E	F
Text	**Content**	**Genre**	**Topic**	**Text Type**
The sample text	Who needs and who is using this content? How is it created, managed and delivered? Who is going to read the content? For what purposes?	In which format was the information delivered?	If the segment related to a sub-topic is removed, is the topic still relevant? What domain of expertise is relevant to understanding this?	What is the text trying to do?
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.	Literary	Fiction	Society and Culture	Narrative

Experiment 1: LLMs as Annotators

GPT-4.5 and Claude Sonnet 4 were prompted to annotate texts with **Content** and **Genre** labels

Category definitions for Content and Genre

Content is "Who needs and who is using this content? How is it created, managed and delivered? Who is going to read the content? For what purposes?"
Genre is "In which format was the information delivered?"

LLMs were then prompted to construct **annotation guidelines** using their annotations

- Converted from text to decision tree format

Experiment 2: LLMs as Domain Experts

GPT-4.5 and Claude Sonnet 4 were given human-annotated texts (Content and Genre) and prompted to construct annotation guidelines

Initial prompt

You are deriving annotation guidelines. Here are XX annotated examples. Each has two annotators.
Your task:
1. List all implicit rules that explain how labels are chosen.
2. Note disagreements and hypothesize the rule conflict.
3. Do NOT generalize beyond evidence.
Return structured rules.

As with Experiment 1, LLMs were prompted to create a decision tree

Results of Experiments 1 and 2

Coverage:

Experiment	GPT 4.5	Sonnet 4
Experiment 1	143 (75%)	157 (82%)
Experiment 2	180 (94%)	181 (95%)

Number and % of labels (of 191) from *mini-reference* covered by guidelines generated in Experiment 1 and 2.

Results of Experiments 1 and 2

Applicability:

- **Survey** presented to four expert linguists comparing annotation guidelines
- **Scores (0-3)** indicate how applicable each guideline was deemed to be

Applicability measures for Experiment 1

Score (0-3)	*Average*	Linguist 1	Linguist 2	Linguist 3	Linguist 4
Sonnet 4	1	2	1	0	1
GPT-4.5	2	3	2	1	2
Overall preference	Sonnet 4	Sonnet 4	Sonnet 4	Sonnet 4	Sonnet 4

Applicability measures for Experiment 2

Score (0-3)	*Average*	Linguist 1	Linguist 2	Linguist 3	Linguist 4
Sonnet 4	1.5	2	2	1	1
GPT-4.5	2	3	1	2	2
Overall preference	Sonnet 4	Sonnet 4	Sonnet 4	GPT-4.5	Sonnet 4

Usefulness:

N/A (no downstream tasks or human IAA)

Experiment 3: LLMs as Analysts

Pilot Study 1

- Documents were annotated with four labels (**Content**, **Genre**, **Text Type**, **Topic**) by two trained human annotators using mini-reference
- Discussion of annotations was recorded using Zoom's transcription tool
- Claude Sonnet 4 was prompted to modify mini-reference based on discussion and disagreements in transcription

Experiment 3: LLMs as Analysts

Pilot Study 2

- Three external human annotators (briefly trained on 10 samples) were provided 40 samples and revised guidelines for annotating the four labels
- Tasked with annotating for four categories (**Content**, **Genre**, **Text Type**, **Topic**)

Results of Experiment 3

- **Coverage:**
  - Pilot 1: 100% coverage over mini-reference
  - Pilot 2: perceived lack of coverage determined by expert annotators to be lack of clarity
- **Applicability:**
  - Pilot 1: Ambiguity between digital and online content; a tagset should be used for **Topic** instead of decision tree
- **Usefulness:**
  - Pilot 1: High level of agreement scores between trained annotators
  - Pilot 2: Untrained annotators disagreed more

Results of Experiment 3

Category	Pilot Study 1	Pilot Study 2	
		Warm-up	Main Task
Content	0.72	0.36	0.27
Genre	0.61	0.34	0.15
Text Type	0.46	0.43	0.19
Topic	0.6	0.61	0.69

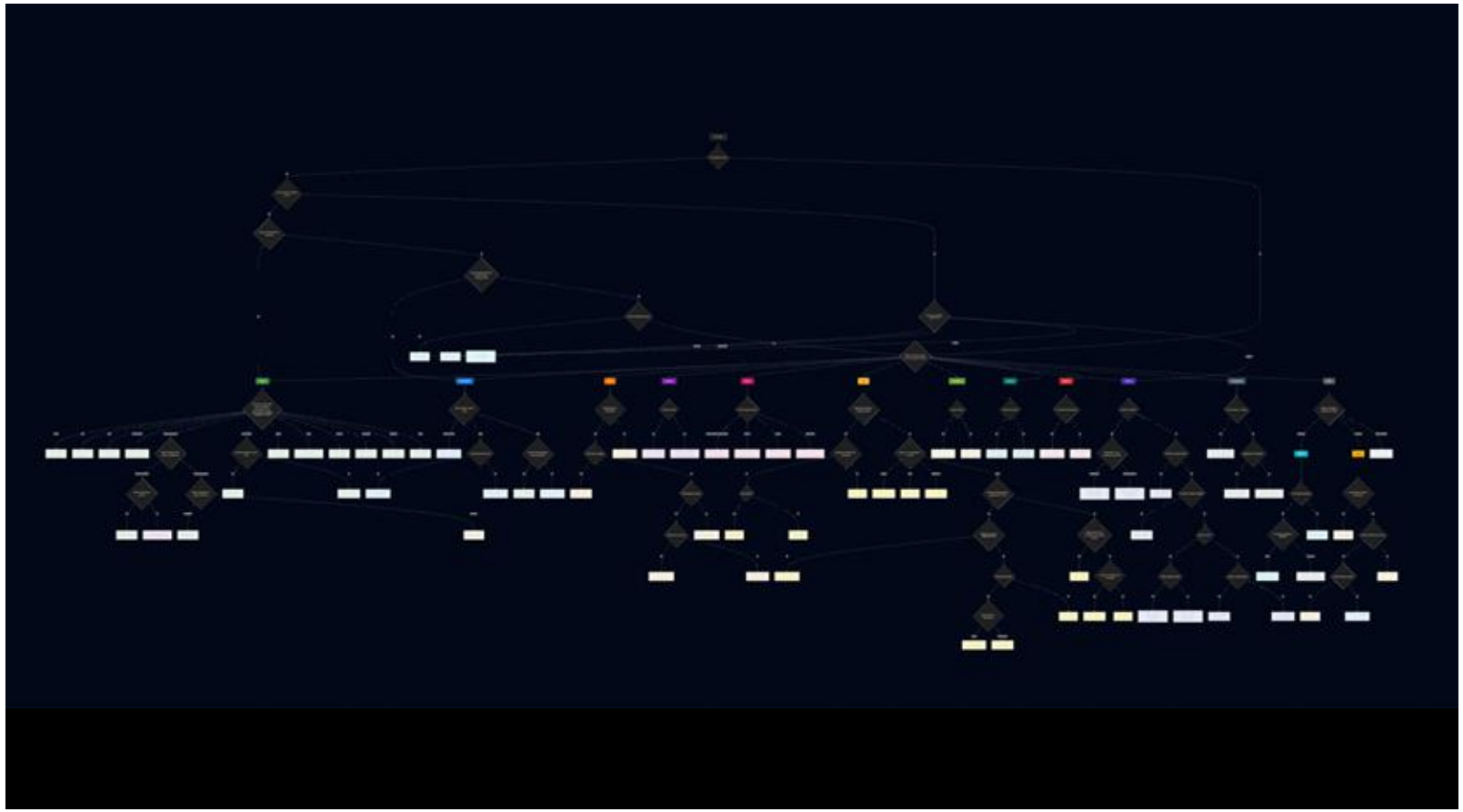
Inter-annotator Krippendorff's α scores from the Pilot Study annotations.

Summary of results

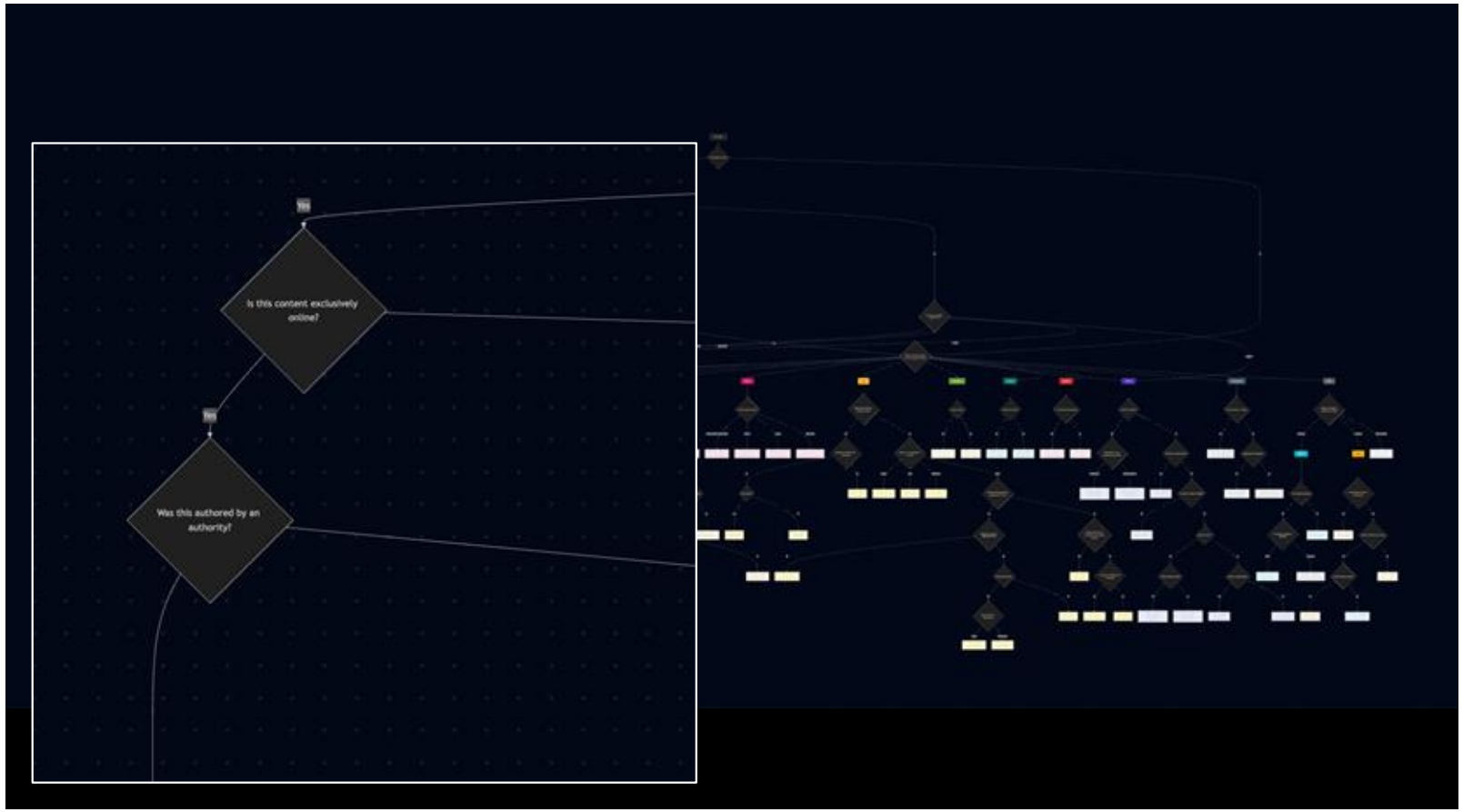
Mini-Reference & Training

- The set-up in Experiment 3 - Pilot Study 1 had the greatest:
  - Coverage
  - Applicability
  - Usefulness
    - when provided with adequate training

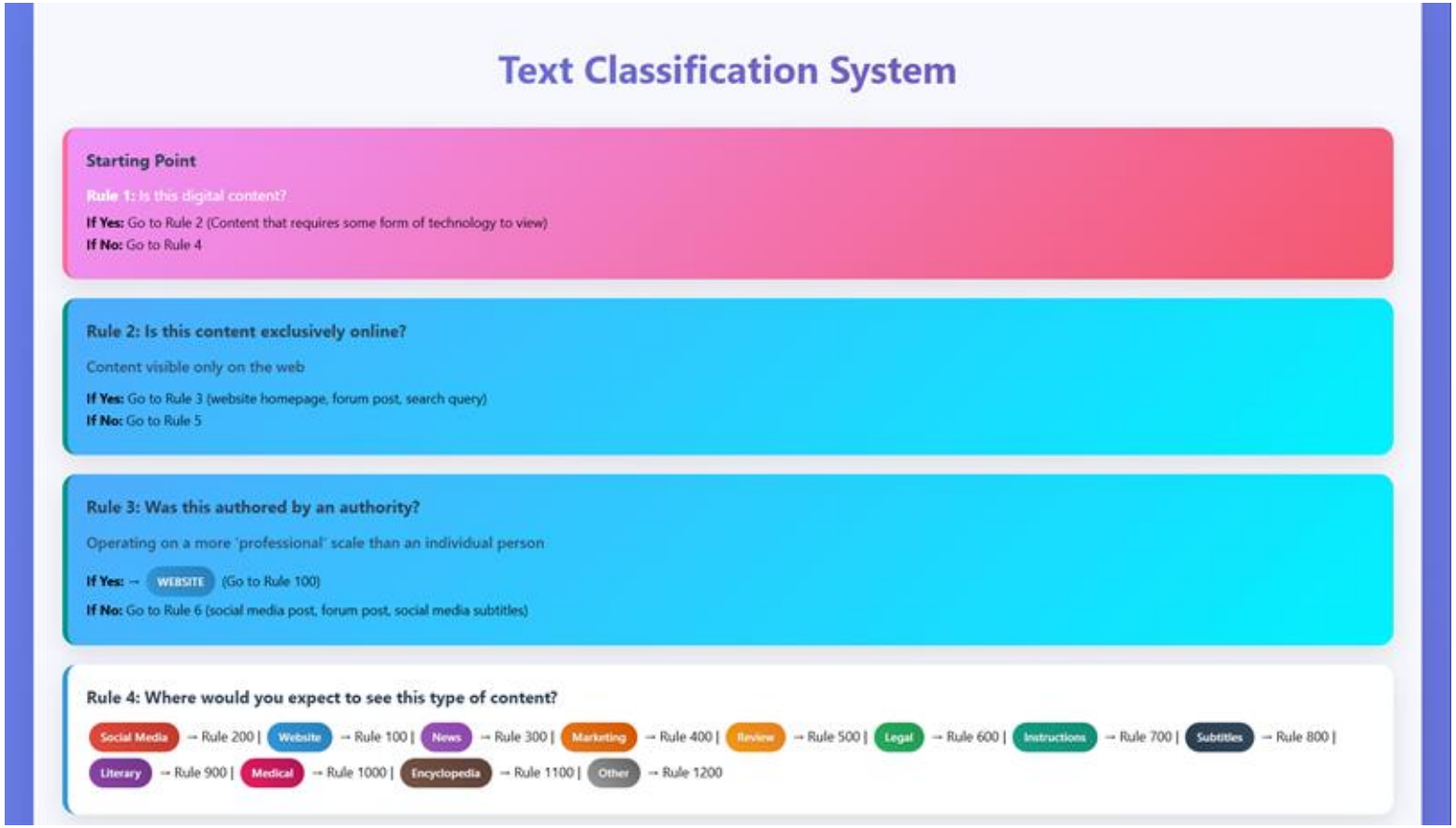
Annotation Guidelines from Experiment 3



Annotation Guidelines from Experiment 3



Annotation Guidelines from Experiment 3



Future Work

- Expanding to other languages (initial steps taken for Irish and Spanish)
  - Focus should move to more typologically diverse languages
- Repeating the experiment with more refined prompts
- Move to annotations of the two innermost classification categories
  - **Style** & **Register**
- Can LLMs help with annotator training?

Takeaways

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Engaging Content
Engaging People

Can LLMs help?

Yes, with the right prompts and proper annotator training!

Essential Reading

- Sheila Castilho and Sharon O'Brien. 2026. Content, Genre, and Domain: Are they all the same? A profiling investigation. In *Proceedings of the 56th Linguistics Colloquium, Switzerland.* Peter Lang. (Forthcoming).
- Gius, E., Reiter, N., & Wielland, M. (2019). A Shared Task for the Digital Humanities Chapter 2: Evaluating Annotation Guidelines. *Journal of Cultural Analytics*, 4(3).

Thank you for your attention!

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

Engaging Content
Engaging People

Questions?
We'd love to hear from you!

QR Code Link to current Annotation Guidelines