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

Mark Andrade, Bláithín Heffernan, Abigail Walsh, Sheila Castilho
ADAPT Centre, Dublin City University
mark.andrade@adaptcentre.ie, heffernanblaithin@gmail.com,
abigail.walsh@adaptcentre.ie, sheila.castilho@dcu.ie

Abstract

This paper investigates the potential of Large Language Models to assist human annotation pipelines, with a particular focus on supporting the development of expert-informed annotation guidelines for document-level content categorisation. We present three experiments exploring distinct roles for LLMs in annotation: as annotators, as domain experts assisting in disagreement resolution, and as analysts of annotator discussions. Using GPT-4.5 and Claude Sonnet 4, we evaluate LLM-generated annotation guidelines for a document-level classification tasks in terms of coverage, applicability, and usefulness. Preliminary results are mixed-to-positive, with evidence that LLMs can provide useful support across different stages of the annotation pipeline, particularly when supplied with rich contextual information such as prior human annotations and annotator discussions.

Keywords: LLM-assisted annotation, annotation guidelines, annotation pipeline

1. Introduction

Data annotation is a central component of Natural Language Processing (NLP), providing the labelled data necessary for training, evaluating, and analysing computational models (Fort, 2016). Across tasks such as classification, information extraction, and discourse analysis, annotation enables the operationalisation of linguistic and semantic phenomena into structured representations that can be processed computationally. Annotation remains a resource-intensive and complex process, often requiring expert knowledge, continuous discussion among annotators to resolve disagreements and align interpretations, careful guideline development, and several rounds of validation to ensure consistency and reliability.

In this context, Large Language Models (LLMs) are rapidly reshaping the landscape of the NLP field, by offering scalable, cost-effective, and accessible solutions. Their adoption across research domains has accelerated in recent years, particularly for text classification, evaluation, sentiment analysis, and other annotation-heavy tasks. On the one hand, LLMs can outperform or complement human annotators in speed and accuracy (Gilardi et al., 2023; Goel et al., 2023), and fine-tuned open-source models have shown promising performance across diverse applications. On the other hand, their outputs are highly sensitive to implementation choices such as model selection, prompting strategy, and parameter settings, leading to risks of bias, reproducibility issues, and "LLM hacking" (Pavlovic and Poesio, 2024; Baumann et al., 2025). Furthermore, the absence of established standards and best practices has raised concerns about the overall quality and validity of LLM-based annotation research (Törnberg, 2024). Annotation outputs are also prone to format inconsistencies, invalid labels, or deviations from instructions, which can undermine downstream analysis if not properly managed.

Given these opportunities and challenges, there is a growing need to critically examine the role of LLMs in data annotation. This work is motivated by the need to empirically test a content profiling framework proposed by Castilho and O'Brien (2026), which encompasses multiple dimensions, including content, topic, genre, text type, style, register, and domain (section 3.1). We explore the potential of LLMs in assisting the human annotation when annotating a subset of this framework (content, genre, text type, and topic) in English. We aim to investigate whether LLMs can further reduce cognitive and labour load, allowing for more productive annotation iterations, while still maintaining control over the decisions at each iteration. To this end, we address the following research questions:
RQ1: Can LLMs generate annotation rules with no previous human annotation as guidance? If so, how helpful are they?
RQ2: How helpful are LLMs in resolving disagreements with human annotators?
RQ3: Can LLMs help with analysing expert human annotators' disagreements and discussion?

In order to evaluate the helpfulness of LLMs, we adopt Gius et al. (2019) assessment model for annotation guidelines (section 2.1.1).

2. Related Work

2.1. Effective Annotation Pipeline

Corpus annotation is defined by Fort (2016, pg. 11) as "adding a note to a source signal". Manu-

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 62–72 May 11, 2026.

ally annotated corpora, in turn, provide the "food for our hungry machine learning algorithms and reference for evaluation" (Fort, 2016, pg. 9), highlighting their central role in the development and assessment of NLP systems. For this reason, annotation guidelines play a crucial role in ensuring that these collaboratively produced annotations are consistent, interpretable, and useful. One of the key challenges in this process is determining the appropriate level of granularity, which directly impacts both annotation quality and usability. Developing effective guidelines therefore requires balancing generality and precision, ensuring they are "as generic as possible but as precise as necessary" (Reiter, 2017). Both authors highlight two key principles of a good annotation pipeline: (i) it must be **iterative**, operating in an agile fashion (Fort, 2016), and (ii) should involve **multiple annotators** working independently on each iteration, and collaborating to discuss disagreements and improvements to the methodology.

Hovy and Lavid (2010) outline six steps[1] for an annotation pipeline: (1) preparing representative texts (the training corpus), (2) instantiating a linguistic concept to define a tagset and the conditions for application, (3) annotating some of the data with multiple annotators, (4) comparing annotators' decisions and unifying divergent annotations, (5) determining a level of satisfactory agreement, and iterating the process until this level is reached, and finally (6) annotating the rest of the corpus. In the first (pilot) iteration of this process, annotators will have to rely on the linguistic concept and features of the training corpus to initiate annotation decisions. Fort (2016) recommends a minimum of two expert annotators with domain and task knowledge to build this mini-reference.

This approach is not universally approved, however, as both Aroyo and Welty (2015) and Deng et al. (2023) challenge the notion of single truth annotations, proposing instead either adopting a crowdtruth approach, or modelling annotator idiosyncrasy arising from differing interpretations, edge cases, and divergent world views and expertise.

Despite these criticisms, the "expert annotators" iterative approach offers a fast and efficient method for development of workable annotation guidelines; particularly of interest for low-resource languages, where crowd-sourced annotation campaigns may be challenged by a lack of fluent speakers, or in low-resource settings, where modelling annotator idiosyncrasies may be prohibitively expensive. Our approach in this paper is motivated by an interest in including LLMs as assistants in this pipeline to further reduce cognitive and labour load, allowing for more productive annotation iterations, while still maintaining control over the decisions at each iteration.

[1] In fact, seven steps are outlined, but the final step specifically relates to training an automatic annotation system, which is not a necessary step for manual annotation pipelines.

Document-level annotations: In Fort (2016)'s definition of annotation referenced above, the source signal can range from a single point or a span of text, up to the entire document. Castilho (2021) found context and the importance of context to be the most salient difference between document-level annotations and their sentence-level counterparts. While it is possible to annotate sentences without knowing its origin, the author found that a lack of context leads to ambiguity.The author also notes that context made annotations easier, including recognising the adequacy and fluency for each. These findings were incorporated into the design of our annotation task, by providing longer samples.

2.1.1. Evaluation Principles

Gius et al. (2019) propose a model to assess annotation guidelines, made up of three dimensions: the **coverage** of a guideline, its **applicability**, and its **usefulness**. Coverage refers to the proportion of the theoretical basis that is covered by a guideline, as different guidelines will cover different samples. Gius et al. (2019) defines applicability as "how well the guideline prepares annotators to do actual annotations", which we interpret as the degree of clarity the guidelines provide. Usefulness reflects how much insight can be provided by a guideline, once correctly employed by an annotator. We understand this final metric to be linked to the application of the annotated data to a downstream task, as the insight provided by annotations depends on what information is valuable in a given context.

2.1.2. Inter-Annotator Agreement (IAA)

While there exist several different measures to calculate IAA, two stand out as providing sufficiently comprehensive coverage to understand the IAA in a given set of annotations.

Krippendorff's alpha (α) can be used on any number of annotators and categories (Krippendorff, 2011). It gives a score from 0 to 1, using the formula

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is the observed disagreement and D_e is the expected agreement.

3. Methodology

3.1. Categorisation Framework

This research approach is grounded in the content profiling framework proposed by Castilho and

O'Brien (2026), which encompasses multiple dimensions, including content, topic, genre, text type, style, register, and domain. This present work is originally motivated by the need to empirically test and operationalise this framework. As shown by the authors, annotating these dimensions is inherently challenging, particularly due to the lack of consistency in how such categories are defined and applied across fields, institutional and industrial contexts.

In line with this, the present study adopts a document-level perspective on annotation, where labels are assigned based on the interpretation of the text as a whole rather than isolated segments. While the broader framework of our annotation includes the dimensions proposed by Castilho and O'Brien (2026), this present paper focuses on four concepts:

- **Content**: defined by the purposes of the end-user and can overlap - or not - with each other depending on the purpose of the end-users and how the content is delivered (e.g. *Technical*, *Creative*, *Legal*, or *Marketing*).
- **Genre**: defined by the conventional structures used to construct a complete text within the variety (e.g. *Press reportage*, *Blog*, or *Letters*).
- **Text Type**: intratextual or linguistic features (e.g. *Narrative*, *Exhortive*, or *Persuasive*).
- **Topic**: the thematic content of the document (e.g. *Politics*, *Economy*, or *Medicine*).

For a detailed discussion of each dimension, readers are referred to Castilho and O'Brien (2026) for comprehensive definitions and examples. The annotation guidelines developed for this work were constructed iteratively, based on exposure to a subset of web-based content. In this context, the distinction between web and non-web content often proved to be a decisive factor in informing annotation decisions and refining category boundaries.

3.2. Experiments

This section describes three approaches to the creation of annotation guidelines with the assistance of LLMs, for the purpose of categorising text with **Content** and **Genre** labels. Our most successful approach also included coverage for **Text Type** and **Topic**, however, we focus primarily on the first two categories as these were higher-levels in the model proposed by Castilho and O'Brien (2026). In order to maximise the reliability and performance of LLM assistance, these experiments focus on English language, however, the process was also shown to be replicable in part with two other languages (see Section 7). The LLMs selected for this task were Claude Sonnet 4 (Anthropic, 2024) and GPT-4.5 (OpenAI, 2023). These models were selected to reflect both widespread research practices and complementary technical capabilities. GPT-based models are among the most commonly used in NLP research and provide a strong baseline for comparison. Claude Sonnet 4, in contrast, offers extended context windows and strong performance in handling longer documents, which is particularly relevant for document-level annotation tasks. In addition, its demonstrated capabilities in Irish were important for further related experiments conducted as part of this study. A table of the data used can be found in the Appendix.

3.2.1. Building a Mini-Reference

Before the creation of the initial annotation guidelines, two annotators and two domain experts, all fluent English speakers, instantiated examples and labels for **Content**, **Genre**, **Topic** and **Text type** in an iterative process, with reference to literature explored in the relevant work section, and focusing on web content in particular.

Five different multilingual language datasets were selected as initial representative datasets: DELA (Castilho et al., 2021), Common Crawl (Common Crawl, n.d.), HPLT (Aulamo et al., 2023), OpenSubtitles (Lison and Tiedemann, 2024), and the WMT 2024 General Task (WMT, 2024). For the purposes of this study, only the English portion of each dataset was used. A total of 280 randomly-selected documents were used from these corpora.

The DELA and CC datasets were first annotated separately by the two annotators for all four categories. The annotators then discussed their annotations, focusing on resolving disagreements. These discussions were integrated into a schema, consisting of notes and general questions to consider, for use in subsequent annotation tasks.

The remaining three corpora (HPLT, OpenSubtitles, WMT) were annotated differently. Using Zoom (Zoom Video Communications, Inc., 2024) to record the discussions of annotation decisions, a transcription was automatically generated representing human expertise and analysis of each annotation decision. In total, 191 label combinations across **Content** and **Genre** were generated for these three corpora. This discussion and the annotated data from all five corpora represent a mini-reference of 280 documents, which are integrated into LLM prompts described in experiments below. During this process, it was found that **Content** and **Genre** labels, typically demonstrated a one-to-many relationship when grouped together, as each Content type had multiple Genre types associated with it.

3.2.2. Experiment 1: LLMs as Annotators

Experiment 1 aims to answer RQ1 by investigating LLMs' capacity as data annotators, with no human guidance. Both Claude's Sonnet 4 and Open AI's GPT-4.5 were tasked with annotating the same documents in the mini-reference. Models were provided the entire dataset consisting of 280 documents, which were collated into subsets for efficiency and manually pasted into the chat interface, along with a prompt to annotate each document for **Content** and **Genre**, and provided with the following definitions:

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?"

Both models generated annotations for every document in the corpus, and were then prompted to create a series of annotation guidelines based on their annotations, producing two textual lists of instructions to follow to arrive at the annotations generated by each model. Both LLMs were then further prompted to create decision tree representations of those guidelines, using Javascript code for a Mermaid tree diagram (Sveidqvist et al., 2025).

3.2.3. Experiment 2: LLMs as Domain Experts

Experiment 2 aims to answer RQ2 by investigating whether LLMs can assist as domain experts in resolving disagreements between annotators, and generate structured rules based on human annotations.

Both LLMs were provided with the entire collection of annotated documents from the mini-reference in blocks through the chat interface, followed by this prompt:

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.

Note that the number of annotated examples given at one time (20, 25, 30) varied for each dataset, depending on context limits. The rules were edited by the LLM according to the samples provided. Both LLMs were asked to implement these structured rules into decision trees using Javascript code.

3.2.4. Experiment 3: LLMs as Analysts

The experimental methodology for Experiment 1 and Experiment 2 was designed to retroactively explore questions that emerged throughout the process of generating annotation guidelines. Experiment 3, was intended to investigate whether LLMs could assist with generation of annotation guidelines when provided with human judgements and discussions of annotation decisions (RQ3).

Annotation guidelines were generated by Claude's Sonnet 4, due to higher input length capacity, and perceived higher quality, compared to GPT-4.5. following a similar methodology to the previous experiments. The model was prompted to generate the Mermaid tree diagram in Javascript after the first transcript was provided, as the reasoning behind the labels applied would give greater context. Following each subsequent transcript, the model was prompted to modify the existing tree design to incorporate additional annotation examples and decisions. This method mimics the iterative annotation pipeline described in Section 2.1.

Experiment 3 additionally included annotator discussions and decisions for **Text type** and **Topic**, with Sonnet 4 prompted to produce decision trees for each of these categories also. Unlike the latter pair of categories, text type and topic were not deemed to be conceptually linked, and annotation guidelines for these two categories were independent of the other.

3.2.5. Human Pilot Annotations

Two pilot annotation tasks were conducted in order to test the efficacy of the annotation guidelines produced in Experiment 3 (LLMs as Analysts), and further improve on the output of Sonnet 4. In each case data was annotated according to the guidelines, and suggestions were made with the aim of improving the existing guidelines.

Pilot 1: The first study was organised following the generation of annotation guidelines produced in Experiment 3, involving the same two annotators who were involved in building the mini-reference. Three different sources of English data were selected: the WMT 2024 General Task (WMT, 2024) data, the UD English Web Treebank (EWT) (Universal Dependencies, 2025), and OpenSubtitles (Lison and Tiedemann, 2024). Once again, post-annotation discussion was transcribed and used

as input, in combination with the document annotations and Experiment 3 annotation guidelines, to Sonnet 4, with a prompt to modify the guidelines to incorporate annotator decisions and feedback. Inter-annotator agreement between the two annotators was calculated with Krippendorff's α, the results of which are discussed in Section 4. Immediately following this task, the two annotators and two domain experts reviewed these guidelines and proposed changes, which were applied in a second pilot annotation with external testers.

Pilot 2: In order to test the developed guidelines on a sufficiently varied dataset, a subset of fifty English documents were gathered from the English side of fifty different multilingual corpora from the OPUS collection (Tiedemann, 2025).

To establish whether the guidelines created from the mini-reference and first pilot study could be used by annotators with less training, three external testers were gathered and tasked with annotating a new dataset of documents, with all four category labels: **Content**, **Genre**, **Topic** and **Text type**.

A short training session was provided to explain the task and function of the annotations guidelines. Ten documents were first annotated in a warm-up task, followed by a discussion to record any ambiguities or confusion that the testers had encountered. Some slight modifications were made to the annotation guidelines in order to clarify ambiguous language, however no structural changes were made. Testers then completed the annotation of the following forty documents.

4. Results

The three metrics proposed by Gius et al. (2019) were adopted in our evaluation of the annotation guidelines produced by the two LLMs and modified by the human experts.

Coverage was assessed by manually reviewing how many of the 191 labels (including near duplicates) produced in the mini-reference were covered by labels created by LLMs in output annotation guidelines. In cases where LLM-generated labels were deemed more specific than the human-generated annotations, the reviewer referred to the original document to assess if the LLM-generated label was consistent with the original text.

Applicability was assessed by examining the annotation trees and assessing their clarity, through surveying the two annotators and two domain experts, or through feedback gathered during the two pilot studies.

Usefulness was the most challenging metric to measure. According to Gius et al. (2019), this metric relates to the application (including subsequent analysis steps) and understanding (including hermeneutic interpretation) of the annotated data within the field of digital humanities. In the absence of a relevant downstream application for this annotated data, we instead use annotator consistency and accuracy measures as an approximate measure of usefulness, understanding that this alone does not provide a complete understanding of annotation guideline usefulness. Our IAA studies are limited to Experiment 3 (Section 3.2.5). Pilot studies assessing annotator consistency for Experiments 1 and 2 are relegated to future work (Section 7).

4.1. Exp 1: LLMs as Annotators & Exp 2: LLMs as Domain Experts

4.1.1. Coverage

Table 1 shows that all four annotation trees had a high level of coverage over the mini-reference annotations for **Content** and **Genre**. While Sonnet 4 does outperform GPT 4.5 when creating the tree from scratch (Experiment 1), the two are almost equivalent in terms of the coverage generated when given the human annotations (Experiment 2). It would seem intuitive that LLMs would incorporate the human-generated annotations provided as input, however, neither model had 100% coverage of these annotation labels.

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

Table 1: The number of labels from the mini-reference that were deemed to be covered by the labels created by the LLM-generated annotation guidelines in Experiments 1 and 2.

4.1.2. Applicability

To assess the applicability of the annotation guidelines (i.e. how well the annotation guidelines prepare annotators to perform the annotation task (Gius et al., 2019)), the four annotators and domain experts were asked to review each of the four **Content/Genre** annotation guidelines from Experiment 1 and Experiment 2, and provide a rating between 0 to 3, indicating how clear the guideline was, with 0 indicating that no additional explanation was necessary, and 3 indicating that the tree was not useable without additional explanation. Additionally, as each model tended to produce guidelines of a particular style, each participant was asked to choose between both models for each experiment, selecting the guideline they would prefer to use for the annotation task. The results are presented in Table 2 and Table 3.

Across both experiments, Sonnet 4 models were generally judged to be clearer and more applicable than GPT 4.5 models. Interestingly, the guideline

Score (0-3)	*Average*	Annotator 1	Annotator 2	Annotator 3	Annotator 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

Table 2: Applicability measures for Experiment 1.

Score (0-3)	*Average*	Annotator 1	Annotator 2	Annotator 3	Annotator 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

Table 3: Applicability measures for Experiment 2.

created by Sonnet 4 from Experiment 1 data (i.e. no human annotations) was deemed to be the most clear and understandable model, with an average applicability score of 1. The results require further investigation through additional pilot tasks.

4.2. Exp 3: LLMs as Analysts

4.2.1. Coverage

As the annotation guidelines in both pilot studies were derived from the mini-reference annotations, and were manually checked at each iteration of their generation, the final iteration of guidelines produced in Exp. 3 had complete coverage over the data used in its creation.

In Pilot 3, testers felt that certain documents were not covered by the **Content** and **Genre** decision tree, including religious texts and song lyrics. As this was not the case for the expert annotator i.e. the annotator from the mini-reference, it appears that the issue here is one of clarity rather than coverage.

For both **Text Type** and **Topic**, the testers felt that the ability to select multiple labels for the same category would have been more appropriate, as it create a more comprehensive understanding of the text. This in turn would increase the amount of insight the annotations would give, and represents an avenue for further research.

4.2.2. Applicability

Feedback from Annotators in Pilot Study 1: Annotators agreed that rules relating to *Subtitles* and *Reviews* were clearest in the **Content/Genre** decision tree, while noting difficulties in differentiating between the various texts within *Social Media* content, and between *Social Media* and other forms of *Online/Web* content. Annotators also had difficulty annotating documents containing elements from a mixture of **Content** or **Genre** labels. To improve clarity, additional questions were added to enhance the guidelines for annotation of *Online/Web* content. In order to flatten the tree and reduce label bias (i.e. creating a more symmetrical tree shape), a top-level content determination question was added.

The decision tree produced for annotating **Text Type** initially appeared simpler in structure to the **Content/Genre** tree, with fewer possible labels; in practise, however, annotators found it difficult to decide on which features of a document were most pertinent for deciding on a final **Text Type** label. Rules relating to *Expository* and *Narrative* branches of the tree were found to be the clearest, while rules pertaining to the *Persuasive* and *Interrogative* branches were less so. Annotators again struggled with documents containing mixed text. Occasionally, the annotation guidelines prioritised an annotation based on textual features in the documents that were contrary to what annotators believed was the primary purpose of the text (e.g. *How-to Guides* contain elements of *Expository* text, but should be classed as *Instructional*).

Rules produced for annotating **Topic** were deemed the least clear out of all four categories. As selection of **Topic** label typically correlates with lexical choices in the document (rather than surface level or contextual features associated with the other three categories), Sonnet 4 consistently produced very flat **Topic** annotation guidelines with minimal-to-no application rules, and a tagset incorporating generalised labels covered in the mini-reference. Additionally, annotators found the labels were often too broad, acting as a 'catch-all' for documents where the topic was more ambiguous (e.g. '*Society and Culture*' and '*Lifestyle and Recreation*' as separate **Topic** labels). To improve clarity, annotators determined a new set of subcategories, improving the granularity of **Topic** labels.

Feedback from Testers in Pilot Study 2: Discussions during Pilot 2 revealed that testers had the most difficulty differentiating between *Digital* and *Website* **Content**. One rule in particular required determining whether the content was digital-only, or available in both digital and non-digital formats, which was a source of tester disagreements and

confusion. Annotating the **Content** of text originating from Wikipedia (Wales and Sanger, 2001) caused confusion as to whether the *Encyclopedia* or *Website* was more appropriate. Testers found that rules pertaining to *Legal* and *Subtitle* **Content** were easy to apply, similar to annotators from Pilot 1. Within *Legal* **Content**, there was high agreement applying the label for the *Legislation* **Genre**.

Contrary to feedback from annotators in Pilot 1, testers in Pilot 2 found the annotation guidelines for annotating **Text Type** were the clearest overall, with the *Expository* label being the easiest to apply. The only reported issue was uncertainty when two labels seemed equally valid for a document. Mixed texts, therefore, continue to present the greatest challenge in terms of **Text Type**.

Testers reported that labels for annotating **Topic** were generally appropriate for the documents, noting that a small number of texts could reasonably be categorised under two different **Topic** labels. Testers found *Technology*, *Politics* and *Religion* were easily applicable labels, while *'Nature & Environment'* was deemed a vague label.

4.2.3. Usefulness

Annotator consistency and accuracy scores were calculated to provide an approximate measure of the usefulness of the annotation guideline produced in Experiment 3.

Consistency: The α scores in the first column of Table 4 show that annotating the **Content** category in Pilot 1 produced the highest average IAA score (0.72), followed by **Genre** and **Text Type** in the same task at 0.61 and 0.6 respectively. IAA scores notably dropped in Pilot 2, consistent with addition of new testers who were not expert in this task. Counter-intuitively, IAA scores do not increase from the warm-up task to the main task for annotating **Content**, **Genre**, and **Text Type**, which is discussed in Section 5.2. Interestingly, IAA scores of **Topic** annotations improve throughout Pilot 1 and Pilot 2, however the increase is slight.

Category	Pilot Study 1	Pilot Study 2 Warm-up	Pilot Study 2 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

Table 4: Inter-annotator Krippendorff's α scores from the Pilot Study annotations.

Accuracy (Pilot Study 2): Table 5 provides an approximate accuracy score for annotation guidelines produced in Experiment 3. These scores represent an IAA measure comparing labels produced by each tester in Pilot 2 with a gold standard label produced by the expert annotator, indicating how closely new testers aligned with ground truth annotations. Comparing these figures with accuracy scores generated for the warm-up task (Table 6), we see that accuracy scores greatly improved following the warm-up and discussion phase, (average of +0.41 per annotator, or +0.42 per category), indicating the importance of annotator training.

Category	1 vs G	2 vs G	3 vs G	*Average*
Content	0.40	0.68	0.60	*0.56*
Genre	0.23	0.50	0.40	*0.38*
Text Type	0.43	0.58	0.53	*0.51*
Topic	0.65	0.73	0.75	*0.71*
Average	*0.43*	*0.62*	*0.57*	

Table 5: Accuracy scores from the **main task** (40 documents) of the second Pilot Study, comparing IAA scores of a tester (1, 2, or 3) and a gold-standard annotation provided by an expert (G).

Category	1 vs G	2 vs G	3 vs G	*Average*
Content	0.1	0.1	0.3	*0.17*
Genre	0	0	0.3	*0.1*
Text Type	0.2	0.1	0.2	*0.17*
Topic	0	0.1	0.1	*0.04*
Average	*0.08*	*0.08*	*0.23*	

Table 6: Accuracy scores from the **warm-up task** (10 documents) of the second Pilot Study, comparing IAA scores of a tester (1, 2, or 3) and a gold-standard annotation provided by an expert (G).

5. Discussion

5.1. Exp 1: LLMs as Annotators & Exp 2: LLMs as Domain Experts

Both guidelines exhibited adequate levels of coverage in Experiment 1. GPT-4.5 tended towards broader labels, while Sonnet 4 simply eliminated near duplicates and some labels deemed to be once-off instances. Some of Sonnet 4's labels were more specific than those provided in the mini-reference, particularly for *Subtitles* content.

The guidelines in Experiment 2 unsurprisingly had even higher levels of coverage. However, GPT-4.5 failed to generate labels to cover *Instructions* content, while Sonnet 4 failed to provide Genre labels for general *Websites* like *Homepages* and *Indexes*.

Across both experiments, Sonnet 4 annotation trees showed greater clarity than their GPT-4.5 equivalents. It seems Sonnet 4 models were better able to provide specific questions to help annotators find the correct **Content** and **Genre** labels. ChatGPT gave simplified trees, with smaller structures, but failed to be as specific as was needed for this task, resulting in lower applicability scores.

5.2. Experiment 3: LLMs as Analysts

As mentioned, the resulting guidelines from Experiment 3 had complete coverage, which was checked manually. As well as that, it was possible to add in any changes that were necessary from the Pilot Studies to the existing model. In order to make similar changes to the trees in the first two experiments would have been more difficult, as the samples would have to be re-annotated in the case of Experiment 1, and the disagreement would have to be solved by the LLMs in the case of Experiment 2.

As seen in both Pilot Studies, the trees for **Content/Genre** and **Text Type** were found to be quite clear. The same was the case for the **Topic** sublistings. It it worth noting this was the case both for internal annotators, those familiar with the initial data, and external annotators, who were only presented with the data they had to annotate.

The usefulness of this set of guidelines is not completely clear. IAA results show fairly high agreement between annotators in Pilot 2, and fairly low levels of agreements between participants in Pilot 2 for **Content**, **Genre** and **Text Type**. **Topic**, on the other hand, improved slightly compared to the Pilot Study. In contrast, accuracy scores reported for Pilot 2 showed moderate-to-high IAA between each new tester and an expert annotator from Pilot 1, indicating that the low IAA scores between testers may not be entirely due to lack of clarity in the guidelines.

A strong possibility for the perceived decrease in IAA from the warm-up task to the main task in Pilot 2 is that easier texts were intentionally selected for the warm-up task, meaning texts in the main task were deemed more complex. Other possibilities for low IAA scores in Pilot 2 include the briefness of the training session, and the fact that it was not possible to hold all the feedback sessions at the same time. Another possible explanation for the decrease in agreement of the first three categories may be the use of a list format for Topic, as this may have been easier to comprehend than the trees.

5.3. Final guidelines

Following Experiment 3 and Pilots 1 and 2, we arrived at a model made up of 16 Content types, 65 Genres, 14 Topics and 6 Text Types. As evidenced from the Methodology section, these categories were assembled empirically from the data examined. While we cannot claim to account for all domains present in online corpora, we hope this serves as a somewhat comprehensive document that can be applied moving forward.

One **Content** label (*Historical*), and four **Genre** labels (*Website-Search Results*, *Website-Social Media-Feed/Post*, *Website-Prompt/Answer* and *Website-Prompt*) were removed during the Pilot studies. The *Historical* **Content** label came from a biography, but these types of texts were subsumed into the *Encyclopedia* label. On the **Genre** side, *Search Results* was similarly subsumed into *Search Query*. *Feed* and *Post* were split into two genres. *Prompt* and *Prompt/Answer* referred to similar texts found in the WMT data, designed to test data quality metrics. These were subsumed by *Didactic-Exam* and *Didactic-Study Notes*, as it was felt these types of documents could occur elsewhere.

6. Conclusion

This paper explores the potential of LLMs to support human annotation pipelines, with a particular focus on their usefulness in assisting the development of expert-informed annotation guidelines for document-level content categorisation(Castilho and O'Brien, 2026). To this end, we designed three experimental setups to address our RQs, using GPT and Claude Sonnet: (RQ1) LLMs as Annotators, where models were provided with the full dataset and prompted to assign Content and Genre labels based on predefined definitions; (RQ2) LLMs as Domain Experts, where models were exposed to a collection of human-annotated examples from a curated mini-reference; and (RQ3) LLMs as Analysts, where models were additionally given access to annotator discussions and decision-making processes, extending the task to Text Type and Topic.

Our results indicate that LLMs can produce usable guidelines in all three settings, when evaluated on coverage, and applicability. In particular, our results indicate performance improves when models are provided not only with definitions or annotated examples, but also with access to annotator reasoning and disagreement, suggesting that LLMs benefit from richer representations of the annotation process rather than static guidelines alone. Among the models tested, Claude Sonnet shows stronger performance, which may be attributed in part to its ability to process longer input contexts. The pilot studies conducted in Experiment 3 further suggest that the final guidelines produced in this iterative human-LLM pipeline show moderate-to-high effectiveness for the task of content-categorisation, when evaluated on coverage, applicability, and usefulness. These findings point towards the importance of modelling annotation as a dynamic and

context-sensitive activity, rather than a purely label-assignment task.

Overall, this work contributes to the field in several ways. First, it provides empirical evidence that LLMs can assist not only in annotation itself, but also in the development and refinement of annotation guidelines. Second, it demonstrates the value of incorporating annotator discussions and disagreement into LLM-supported pipelines. Third, it proposes a structured framework for integrating LLMs at different stages of the annotation pipeline, from labelling to analysis. Finally, it opens new avenues for research into context-aware and human-centred approaches to annotation, particularly in settings characterised by ambiguity and low inter-annotator agreement.

7. Limitations & Future Work

This study has several limitations that open avenues for future research. First, while the experiments reported here focus primarily on English-language data to maximise LLM performance and controllability, the extent to which these findings generalise across languages remains an open question. Ongoing work applying the same framework to both Irish and Spanish provides an initial step in this direction, offering insights into how LLM-assisted annotation performs in a lower-resourced language setting. However, broader validation across typologically diverse languages is still needed. Future work will therefore extend this framework to multilingual scenarios, with particular attention to languages where annotation categories such as genre, text type, and register may not map cleanly across linguistic and cultural contexts.

A second limitation relates to the experimental design of Experiment 3, where models were exposed to annotated data and annotator discussions. While this setup was intentional in order to simulate LLMs as analysts, it introduces a potential risk of data leakage. In principle, models could reproduce or approximate previously seen annotations when applied in earlier experimental conditions, thus inflating performance estimates. Although no direct memorisation effects were observed in our analysis, this possibility cannot be fully ruled out. Future work should therefore explore stricter data separation protocols, as well as controlled evaluations designed to explicitly test for memorisation and leakage effects in LLM-assisted annotation workflows.

Finally, the research presented in this paper represents initial results exploring this question. Although register and style were included in our preliminary annotations, the decision was made to set these categories aside until our knowledge of, and accuracy in the first four categories had increased. In the future, it is our intention to incorporate these categories, to provide a richer understanding of texts. Additionally, we intend to expand the evaluation of LLM-generated annotation guidelines through additional pilot studies and downstream application of the data, allowing for more reliable measures of usefulness.

8. Acknowledgements

This research is partially funded by the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

9. Appendix

Dataset	Usage
DELA, Common Crawl, HPLT	Mini-reference
OpenSubtitles	Mini-reference, Experiments 1, 2 and 3 (both studies)
WMT 2024 General Task	Mini-reference and Experiment 3: Pilot Study 1
UD English Web Treebank	Experiment 3: Pilot Study 1
ada83, bible-uedin,giga_fren, hrenWaC, infopankki, Joshua-IPC, KDE4, KDEdoc, KFTT, liv4ever, MDN_Web_Docs, memat, MIZAN, MultiUN, NeuLab-TedTalks, News-Commentary, NLLB, OpenOffice, OpenSubtitles, Paracrawl, Parlce, PHP, pmindia, QED, RF, Salome, sardware, SciELO, SETIMES, SPC, StanfordNLP-NMT, Tanzil, TED2020, TedTalk, TEP, Tilde-MODEL, tldr-pages, Ubuntu, UNPC, wikimedia, Wikipedia, Wikisource, WMT24++, WMT-News, and Xhosa Navy	Experiment 1 and 2, and Experiment 3: Pilot Study 2

Table 7: The datasets used in the creation of the mini-reference and each of the experiments.

9.1. Existing typology at the end of Experiment 3

Content Digital, Social Media, Website, Marketing, News, Review, Legal, Instructions, Notice, Subtitles, Fan Subtitles, Literary, Medical, Encyclopedia, Didactic, Other

Genre Archive, Software, Video Game Interface, Search Results, Profile, Feed, Forum, Post, FAQ, Blog, Search Query, Catalogue, Product Description, Index, Brochure, How-to, Online Help, Fiction, Creative Nonfiction, Homepage, Boilerplate, Article, Opinion Piece, Hard News, Feature Article, Press Report, Interview, Media Review, Product Review, Service Review, Experience Review, Quiz, PSA, Proceedings, Form, Legislation, Journal Article, Press Release, Newsletter, Guide, Programme, Report, Recipe, Visual Entertainment, Social Media, Talk, Audio Entertainment, Nonfiction, Review Article, Object Biography, Biography, Reference Material, Exam, Study Notes, Address, and Obituary

Text Type Interrogative, Instructional, Narrative, Argumentative/Persuasive, Expository, Descriptive, Other

Topic History, Finance, Politics, Religion, Personal Relationships, Science, Technology, Culture & Entertainment, Health, Education, Lifestyle & Recreation, Nature & Environment, Society & Demographics, Industry & Employment, Other

10. Bibliographical References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation.

Sheila Castilho. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45. Association for Computational Linguistics.

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).

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. Wiley-ISTE.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Evelyn Gius, Nils Reiter, and Marcus Wielland. 2019. A shared task for the digital humanities chapter 2: Evaluating annotation guidelines. *Journal of Cultural Analytics*, 4(3).

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. LLMs Accelerate Annotation for Medical Information Extraction.

Eduard Hovy and Julia Lavid. 2010. Towards a science of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Nils Reiter. 2017. How to develop annotation guidelines. Last accessed: 23 February 2026.

Knut Sveidqvist, Sidharth Vinod, Ashish Jain, Neil Cuzon, Tyler Liu, Alois Klink, Reda Al Sulais, Nikolay Rozhkov, Justin Greywolf, Steph Huynh, Matthieu Morel, Marc Faber, Yash Singh, Nacho Orlandoni, Per Brolin, and Mindaugas Laganeckas. 2025. Mermaid: Diagramming and charting tool.

Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. *Sociologica*, 18(2):67–85.

11. Language Resource References

Anthropic. 2024. *Claude Sonnet 4*. Anthropic.

Mikko Aulamo and Nikolay Bogoychev and Shaoxiong Ji and Graeme Nail and Gema Ramírez-Sánchez and Jörg Tiedemann and Jelmer van der Linde and Jaume Zaragoza. 2023. *HPLT Corpus v1.2*. European Association for Machine Translation.

Sheila Castilho and João L. Cavalheiro Camargo and Miguel Menezes and Andy Way. 2021. *DELA corpus*. Sheila Castilho.

Common Crawl. n.d. *Common Crawl Corpus*. Common Crawl Foundation.

Pierre Lison and Jörg Tiedemann. 2024. *OpenSubtitles parallel corpora v2024*. OPUS – The Open Parallel Corpus, University of Helsinki.

OpenAI. 2023. *ChatGPT (GPT-4.5)*. OpenAI.

Jörg Tiedemann. 2025. *The OPUS collection*. University of Helsinki.

Universal Dependencies. 2025. *Universal Dependencies English Web Treebank v2.16*. Universal Dependencies.

Jimmy Wales and Larry Sanger. 2001. *Wikipedia — Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia. [Online; accessed 23-February-2026].

WMT. 2024. *Shared Task: General Machine Translation (WMT24)*. WMT Shared Task Organizers. Retrieved 15 September 2024.

Zoom Video Communications, Inc. 2024. *Zoom*. Zoom Communications, Inc.