What happens on the fringes, stays on the fringes?

Information flows in the contemporary media system

Simon, M.

Publication date
2024

Chapter 3

Validating methods tracing information flows

Testing An LLM-assisted human validation scheme

Abstract

Scholars of (computational) communication science, political science, and information science have developed a wealth of approaches to detect whether a connection exists across various platforms, agendas, or individual actors, through the information they share. However, not only is the tracing of such connections difficult and computationally expensive, the validation of information flows is even more challenging. High quality manually-labeled human annotations are costly, time-consuming and cognitively taxing even for expert annotators due to the dynamic nature of flows. We harness the power of recently emerged more open, light-weight, and self-hosted Large Language Models (LLMs) LlaMA 7B-chat and LlaMA 13B-chat and an human expert annotator to propose and test to what extent can we employ an *LLM-assisted human validation scheme* for document-comparison-based annotation tasks such as the detection of information flows across documents. We apply this novel scheme to a corpus of news article pairs published around the Dutch General elections of 2021. Although the LLMs in our proposed validation scheme did not prove very successful on neither the topic-level, nor the news events-level matching tasks, our analyses provide a series of valuable insights that scholars should take into account when working with LLMs and when validating flow-based methods. We also publish a high-quality, expert evaluated gold standards dataset for information flows on topic-level as well as news-events-level that can be used for evaluation and training purposes.

keywords: information flows, validation, LLMs, annotation, gold standard

Introduction

Chapter 2 and Chapter 3, studied information flows using state-of-the-art computational methods to provide a deeper understanding of the role of dark platforms in the broader media system as well as to explore new ways of studying influence dynamics in the contemporary information ecosystem. After uncovering new ways of tracing information flows in the contemporary media system, it has become clear that methodological progress in tracing information flows is held back by the absence of rigorous validation approaches. In fact, our knowledge is extremely limited regarding the challenges that affect that validation of more complex, and flow-based methods such as the validation of methods tracing information flows. Therefore, the aim of this chapter is to propose a novel validation scheme that could aid computational communications scientists in testing the validity of advanced computational methods used to trace information flows.

Tracing information flows across various entities has been studied by several different disciplines using similar methods. From computer science, communication science to political science, these studies broadly focused on how (textual) content spreads between different (types of) news sources, agendas, platforms, or actors across time and various issues (see e.g. Trilling & van Hoof, 2020; Welbers et al., 2021; Wilkerson et al., 2015; Zannettou et al., 2017). The notion of information flows is at the core of a wealth of studies that deal with such topics and theories as agenda-setting (Barberá et al., 2019), information reliance (Welbers et al., 2021; Wilkerson et al., 2015), content-overlap (Chapter 2), discursive power (Jungherr, Schroeder, & Stier, 2019), or information cascades (Nicholls & Bright, 2019). While these studies may be grounded in diverse theoretical frameworks and conducted on various levels of analysis, they share common methodological approaches that are largely based on document comparison. Specifically, they employ various document similarity metrics to evaluate to what extent the type of information flows they are investigating are discernible among the actors, platforms, or agendas under study.

As demonstrated by Chapter 2 and Chapter 3, existing research offers fruitful avenues for detecting information flows across a variety of different documents and entities. However, it has become clear in these chapters as well as previous research that the validation of such methods remains practically challenging, time consuming and expensive. Generating a "gold standard dataset" – high quality manually-labeled data that is assumed to be correct (Welbers et al., 2018) poses non-trivial challenges not only for more complex and dynamic tasks such as detecting whether subtle information flows occurred between random document pairs, but also arguably less complicated, and more static natural language processing (NLP) tasks such as the identification of political and news-related information within a single domain, in one document at a time (Menchen-Trevino et al., 2023; van Hoof et al., 2023).

What makes the validation of information flows complex? The detection of information flows can be divided into two sub-tasks. The first one is the identification of a particular representation of interest such as "newsflows" (Trilling & van Hoof, 2020; Welbers et al., 2021), the flow of policy ideas (Wilkerson et al., 2015), or issue attention and agenda-setting (Barberá et al., 2019). These information representations can take various forms when measured in text, including bag-of-words, named entities, word embeddings, or topics. The validation of these computational approaches is arguably well-understood by the literature and can usually be performed without too many difficulties (see e.g., Nicholls & Bright, 2019; Trilling & van Hoof, 2020).

However, the second sub-task of linking the previously identified representations is much less understood, making the validation of information flows a complex endeavour. To illustrate, validating information flows entails the need to identify representation (i.e., sub-task 1) and then perform pairwise document comparison by taking the passing of time into consideration (i.e., sub-task 2). Even if we disregard the time issue for now, document comparison alone poses a significant challenge. When comparing documents for information flows, annotators need to read at least two (or more), often random documents side by side, and it is often not immediately clear where the similarity between the documents lies. For example, take the counting of overlapping named entities (e.g., people, locations). To accomplish this task, the annotator must identify and then retain all the entities from one text in their memory while scrutinizing the other text. Doing this systematically is difficult and tiresome. The more aspects the annotator needs to consider, the more cognitively taxing and time-consuming the task becomes.

Moreover, the biggest challenge is the exponential nature of annotating for information flows. To illustrate, annotating for information flows across a corpus of 10 news articles already requires 45 annotations per annotator when annotating for information flows across all possible combinations of text pairs. For 50 articles the number of required annotations is 1,225. In practice, computational social scientists work with much larger corpora, hundreds, thousands, or even millions of documents. For 1000 articles, the number of required annotations skyrockets to 499,500. Performing such high numbers of annotations tasks is not realistic. Therefore, when generating gold standard data for more dynamic tasks, that are exponential in nature and that require the careful consideration of multiple features side by side (i.e., document comparison), it becomes extremely cognitively taxing to obtain high quality annotations.

Against this backdrop, when thinking about validating automated text analyses approaches we can distinguish between more *static* validation tasks that can be used to benchmark the identification of concepts, and more *dynamic*, document-comparison-based validation tasks that study *flows*. This study focuses on innovating the dynamic aspect of validating information flows (i.e., sub-task 2). We propose and test a proof of concept for a novel, *LLM-assisted human validation scheme* that harnesses the power of light-weight and self-hosted Large Language Models (LLMs) as well as an expert annotator to optimize the generation of a gold standard data for the detection of information flows.

LLM-assistance in this study can either be *comprehensive* or *partial*. Comprehensive LLM-assistance entails that human annotators receive highlighted keywords, as well as evaluations (reasoning) regarding match levels across document pairs. Partial assistance on the other hand only includes highlighted keywords. With the testing of this novel validation scheme we hope to aid computational social scientists in making well-informed decisions about their complex validation tasks in order to ultimately achieve higher quality as well as more cost- and time-efficient gold standard data generation for their complex NLP approaches. We provide detailed guidance to researchers by scrutinizing if and under what conditions can the proposed LLM-assisted human validation be of use to optimize the validation process of complex annotation tasks that involve dynamic components. Moreover, we make available a high-quality gold standard dataset for information flows ($N = 990$) on the main topic level as well as on the news event-level that can be used by researchers for the evaluation and training of state-of-the-art tools[55].

Background and related research

Detecting information flows

Before diving into the subject of generating gold standard data for information flows, we must first outline what information flows are and why is it so complex to devise approaches to not only detect these phenomena, but also to validate the outcome. The origin of the information flows concept can be traced back to the domain of information theory (Shannon & Weaver, 1964). Information flows are defined as the transfer of information from point **A** to point **B** within a system. The simplicity of this definition has enabled scholars of several disciplines to study information flows within their own field of research. For instance, Trilling and van Hoof (2020), Welbers et al. (2018), and Wong and Trilling (2023) delved into the notion of "news flows", and Nicholls and Bright (2019) into the concept of "news story chains" by uncovering patterns of content overlap and the transfer of information across various media outlets over time. Similarly, Wilkerson et al. (2015) mapped the flow of policy ideas in legislatures by measuring text reuse.

If we break down the task of identifying information flows we arrive to two fundamental sub-tasks (see Figure 15). The first sub-task is the identification of a representation of *information* in documents. The literature to date highlights a diverse set of such representations such as bag-of-words, named entities, proper nouns, topics, or embeddings (Nicholls & Bright, 2019; Trilling & van Hoof, 2020; Wong & Trilling, 2023). Although researchers may opt for different approaches to identify different representations of information depending on their scope, the validation of this task is generally well-understood amongst scholars. A comprehensive codebook and some annotator training should in principle yield a high-quality gold standard with some caveats related

[55] https://github.com/psyronika/LLM_validate

to annotator fatigue (Grimmer & Stewart, 2013) and biases (Hube et al., 2019; Sap et al., 2022) which will be discussed in another section (i.e., Towards a potential solution).

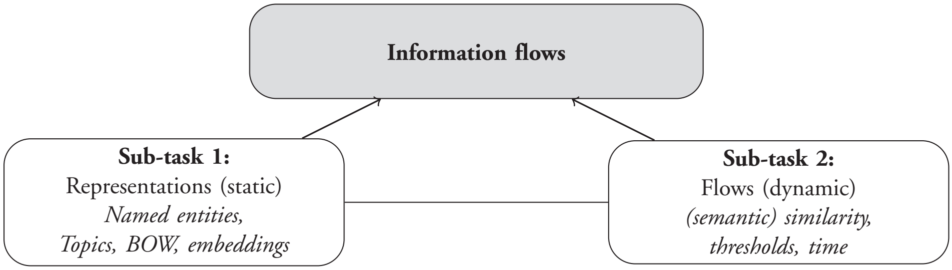

Figure 15: Breakdown of components of tracing information flows

The second, arguably more complex sub-task deals with the *flow* aspect of information flows by linking the aforementioned representations across documents and or domains over time. Hence, while sub-task 1 is more static, sub-task 2 is dynamic in nature. To accomplish sub-task 2, researchers must establish a connection across the identified representations while taking the passage of time into account. To this end, researchers have devised more or less strict measurements and assessment criteria using various thresholds and time windows. To illustrate, while some scholars conceptualized the flows in terms of direct overlap of key features across documents (see Chapter 2), others opted for a threshold-based approach via (semantic) similarity calculated across document pairs over a time period of interest (Novo & Gedikli, 2023b; Trilling & van Hoof, 2020; Welbers et al., 2018, 2021; Wilkerson et al., 2015). Regardless of these differences, scholars appear to agree that pairwise document comparison is crucial to capture flows.

However, the conceptualization and execution of sub-task 2 of tracing information flows can differ a lot across studies. To illustrate, while one study may use cosine similarity to capture flows across news articles and they find that a threshold of 0.6 is the ideal level of similarity that indicates that there is an information flow across two documents within a time-window of 3 days (Trilling & van Hoof, 2020), others may opt to count the number of overlapping entities as the main indicator of flows (see Chapter 2). While both approaches may reveal some level of information flows, it is impossible to say which approach is better or more suitable as they are conceptually very different. Hence, the validation of sub-task 2 is much less clear-cut and well-understood than the validation of sub-task 1. In fact, the big unresolved issue is the absense of a comprehensive validation scheme for information flows that is able to take us one step closer to a unified solution to generate a gold standard that can take both sub-task 1 and sub-task 2 into account with a particularly dire need for solutions for sub-task 2. Henceforth, this study deals with the issue of validating the *flow* aspect of information flows.

Different levels of information flows

The literature to date established various levels in which two documents can share a connection through the information they have in common. For the sake of clarity we will discuss these levels of information flows within the context news, and socially and politically relevant information. According to the literature, information flows can occur on a broad topic or issue level or on a more finer-grained news event (Trilling & van Hoof, 2020) or news story level (Nicholls & Bright, 2019).

First, according to Trilling and van Hoof (2020) “news events are specific events that lead to news coverage, such as a specific debate on a specific day in a specific parliament, a specific accident, or a specific football match. They can be covered by one or more articles in one or more outlets, but relate to one specific and identifiable event and are thus much more fine-grained than news topics, issues, or news categories” (p. 1321). News story chains are similarly conceptualized to news events. The key difference between the two concepts is the expectation of (chronologically) “repeated coverage”. More specifically, Nicholls and Bright (2019) define news story chains “as events or single issues which receive repeated coverage in the news media through a series of initial articles and followup pieces” (p. 44). As such, while the news events definition leaves the possibility open for multiple news outlets simultaneously reporting on the same event, the definition of news story chains is more constraining as it expects a chronological repetition of reporting on the same event with clearly distinguishable initial and follow-up articles. Since the detection of finer-grained information flows is already difficult as is, we stick to the definition of news events that loosens the requirement of a repeated coverage of initial and follow-up articles and also allows for the detection of single story news events.

Finally, news topics are “ thematic news areas [...] which receive repeated coverage but which naturally encompass multiple events, and whose time span is much longer” (p. 44, Nicholls & Bright, 2019). Several news topics represent recurring items of news coverage. To illustrate, let us consider these distinctions in the context of election coverage in news media. The topic “elections”, is a broad and recurring subject in the news. It encompasses various elections happening at different times, locations and levels of government (e.g., presidential or local). The topic sets the stage for coverage and discussions surrounding elections. A news event within the context of elections is more specific than a topic. It focuses on a particular election or related developments with a series of articles that include updates from news outlets, all contributing to the coverage of that specific election or its surrounding events. As such, a news event about elections might revolve around the presidential election of a specific year, detailing campaign events, candidate profiles, polling data, and key issues. An event is a singular occurrence or happening within the broader context of an election news event such as a presidential debate, election day itself, the release of election results, or a major campaign rally.

The problem with validating methods for tracing information flows

Automated methods for tracing information flow have unique challenges in terms of validation. To generate a gold standard dataset, researchers typically hire crowd-workers or undergraduate students to annotate a substantial subset of a large dataset. This practice is not only expensive and time-consuming, but also prone to error. When labelling extensive amounts of data, human annotators will inevitably suffer from fatigue, and a limited attention span (Grimmer & Stewart, 2013). Moreover, differences in the ways in which annotators perceive the task at hand as well as the inevitable influence of knowledge gaps, attitudes and opinions may further bias the quality of the labelling process (Hube et al., 2019; Sap et al., 2022).

Grimmer and Stewart (2013) argued that as novel and sophisticated automated methodologies become the prevailing way to perform text analysis, there is a crucial need for innovation with respect to validation approaches. Validating automated text analysis methods entails the need for a benchmark, in the form of a "ground truth" dataset which is considered the true, absolute, and most trusted known information that typically stems from direct observation (Cardoso et al., 2014). As ground truth is not always available, researchers must invest resources into building a so called "gold standard" dataset which is a manually labeled, high-quality subset of their corpus that is assumed to be correct (Welbers et al., 2018). It must be noted that the terms ground truth and gold standard are often used interchangeably. For the sake of clarity, we use the aforementioned definitions.

Although a wealth of automated methods exist to perform automated text analysis (Boumans & Trilling, 2016), innovation with respect to the validation of these approaches is arguably lagging behind. The gold standard dataset serves as a benchmark that can be used to assess the performance of automated text analysis methods. Generating a gold standard is neither trivial nor cheap. The conventional approach is to first obtain a random sample of the corpus to be annotated as a validation set.

However, if one tries to validate approaches that detect information flows across a large corpus of documents, following the traditional approach would be unfeasible as information flows between random pairs of documents within a large corpus are often extremely rare. Estimating recall requires a complete and accurate gold standard dataset that contains all the relevant instances of information flows. Since information flows are exceptionally rare in a large corpus, any sample makes it impossible to create a comprehensive gold standard dataset, as one may never identify all the possible rare instances of information flows. This is precisely the reason why studies to date have focused validation efforts on document pairs that were more likely to contain information flows within a given "meaningful" time period (see e.g. Nicholls & Bright, 2019; Trilling & van Hoof, 2020; Wilkerson et al., 2015). This represents the first major challenge of validating approaches that detect information flows.

The second major challenge is related to the annotation task itself. While we agree with Nicholls and Bright (2019) that the basic task of manually identifying whether there is some type of relationship between two documents may seem simple at a glance, the practical challenges that arise from the sheer amount of document pairs annotators would need to go through while considering multiple indicators of connectivity paint a different picture. More specifically, the ideal validation scheme for information flows should be able to incorporate previously studied indicators of connectivity such as publication date, named entities (Novo & Gedikli, 2023a), topics (see Chapter 2), events, as well as (semantic) similarity measures (see e.g., Trilling & van Hoof, 2020; Welbers et al., 2018). These indicators are necessary to capture such as latent construct as information flows that is made up of representations (sub-task 1) as well as flows (sub-task 2).

Even if one had access to a small army of highly skilled annotators and a large budget to pay them, the expectation of consistently producing high-quality validation sets while considering all of the aforementioned indicators (and more) meanwhile performing hundreds of document comparisons borders on impracticality. Besides suffering from fatigue and limited attention span, the quality of the gold standard would inevitably be affected by cognitive biases, opinions, knowledge gaps, and the inherent subjectivity that comes with interpretation (Hube et al., 2019). For instance, if annotators' task is to decide whether two news articles refer to the same news event, their annotations may differ simply based on the extent of their knowledge of the particular event or actors being covered.

To illustrate, consider two annotators (A and B) who are tasked with assessing whether two news articles pertain to the same news event based on a document pair of news articles published by different news outlets. News article 1 details a televised debate of a high-profile election with various quotes from candidates, analysis and experts. Article 2 just mentions the debate fleetingly while discussing the elections and politics in general without mentioning particular candidates, or in-depth analysis by experts. Sub-task 1 in this example will be identifying all relevant document representations such as entities, while sub-task 2 will be assessing whether a *flow* exists across the two documents through the identified representations on the news events level. While annotator A who has in-depth political knowledge may instantly connect the two articles even if the match between the identified representations appears to be minimal, annotator B who does not have a great deal of political knowledge and does not follow the news so closely may be hesitant to classify the two documents to be linked on the news events level as they may not recognize the connections due to a lack of knowledge on the subject matter. Therefore, sub-task's 2's outcome depends a great deal on the identified representations (i.e., sub-task 1) but it is also very much influenced by subject-matter knowledge of annotators as well as their ability to retain representations from document 1 in memory while simultaneously comparing these to the representations of document 2.

Hence, the nature of sub-task 2 in particular, that involves complex and multi-step decision-making, demands a very high cognitive load. Consequently, relying solely on human annotations to produce a gold standard for the detection of information flows, may prove highly expensive, time consuming and likely not very successful. We argue that validating methodologies that capture more complex concepts and intricate relationships such as information flows may not be inherently suited to human capabilities as these concepts often involve subtle relationships and multifaceted dynamics that can challenge even experts.

Towards a potential solution

In the past decade scholars have increasingly adopted computer-assisted methods to perform various NLP tasks that traditionally have been done by human annotators (Boumans & Trilling, 2016). Using holistic or indicator-based approaches to teaching computers how to identify latent concepts have become the standard practice in the field of computational social science (Burscher et al., 2015). As NLP techniques become more sophisticated the validation of these approaches is considered even more important (Birkenmaier et al., 2023). Yet, little work addressed the need for clear, and unified guidelines to validating computational text-based approaches in general (Birkenmaier et al., 2023). Moreover, to the best of our knowledge, no study to date addresses the issue of validating flow-based annotation tasks.

Circling back to Grimmer and Stewart (2013), there is a clear need for advances in validation approaches for flow-based annotation tasks. Our work takes a step towards filling this gap by addressing the second challenge, namely the optimization of the manual annotation task pertaining to the dynamic aspect of identifying information flows. We do so by exploring:

To what extent can Large Language Models (LLMs)be used to advance the validation of methods for tracing information flows?

More specifically we propose and test a proof of concept for a novel *LLM-assisted human validation scheme* that harnesses the vast knowledge basis of LLMs to optimize the generation of gold standard data for the detection of information flows. Below we outline the potential benefits and caveats of using such an LLM-assisted human validation scheme for information flows based on what we know so far about using LLMs for research purposes that would replace human annotations. The recent rise of LLMs such as OpenAI's ChatGPT or GPT-4 have taken the world by storm due to their impressive capabilities of performing human tasks such as writing text, code, summarizing, reasoning, translating, etc. Although this technology has only become mainstream recently, it has already had major implications on how information is created and how research is conducted (van Dis et al., 2023). Several scholars have compared these novel LLMs' zero-shot and few-shot classification abilities to more traditional approaches (van Hoof et al., 2023), as well as crowd workers and expert human coders (see e.g. Gilardi et al., 2023; Törnberg, 2023a; Wu et al., 2023).

Early findings suggested that LLMs perform well on various NLP tasks. Some indicated that LLMs can outperform crowd workers as well as trained human coders on a variety of annotation tasks including the detection of stance, topics, frames (Gilardi et al., 2023), ideology (Wu et al., 2023) and political affiliation from natural language (Törnberg, 2023a). Overall, these studies show that LLMs can perform well on conventional text classification tasks (Gilardi et al., 2023; Törnberg, 2023a; Wu et al., 2023) and that the cost of using these models represents only a fraction of standard practices such as hiring crowd workers from MTurk, or training and employing expert coders to perform annotations (Gilardi et al., 2023).

While these studies address important tasks and show promising results, they rarely detail the potential downsides of using LLMs for conventional NLP tasks. This is a concern as there are caveats and shortcomings that are associated with using LLMs for research purposes. If we wish to consider the usage of LLMs in the optimization of how we create validate sets for information flows, we must acknowledge and discuss the biases and issues that may arise from using these models for this particular task and then argue how we can try to mitigate these shortcomings.

The first and perhaps most important caveat of LLMs are the biases that are encoded in the models due to the data they were trained on. Most of the LLMs available today were trained on vast amounts of texts from the internet, and as such they encapsulate the biases and prejudices that were represented by these texts. For instance, multilingual BERT has been found to embody gender bias (Bartl et al., 2020) and ChatGPT has been shown to embody biases against conservatives (McGee, 2023). Therefore, when using LLMs for text generation we must treat the responses generated by LLMs with caution. Instead of taking the responses of LLMs at face value when it comes to performing NLP tasks, we must use some form of human verification to mitigate the biases that may be embedded in the responses of LLMs.

Second, scholars have raised concerns about issues related to reproducibility, transparency and data security when using commercial models such as GPT-4 (Spirling, 2023). To illustrate, all data sent to OpenAI's API is automatically given access to the company that runs the models, which in this case is OpenAI (Törnberg, 2023b). To date, OpenAI has no clear policies in place with respect to this particular aspect of data security and transparency. We neither know whether this data will later be used to train their models, nor whether the data is securely stored on OpenAI's servers. As such, sending potentially sensitive data to commercial API's may jeopardize research integrity and may cause unexpected harm. While scholars caution against using these commercial API's for sensitive or confidential data (Törnberg, 2023b), we argue that sending any kind of research data to commercial API's is problematic. There are safer and more transparent alternatives for researchers than OpenAI's LLMs.

Open-source LLMs are viewed as the way for forward for scientific research (Spirling, 2023). To date only a single truly open-source LLM exists. BLOOMZ, was a result of a large international effort of researchers and practitioners under the BigScience Research Workshop to build the World's first open multilingual Large Language Model [56]. While BLOOMZ represents a truly exceptional model, its overall performance on popular benchmarks is significantly behind a number of self-hosted LLMs [57] not to mention that the computational resources needed to run this model are also rather demanding. Despite this, countless models that are labelled as open-source, fall short on openness in numerous ways, including a lack of access to code, LLM weights, or details about the training data (Liesenfeld et al., 2023). We refer to these somewhat open but partially black boxed models as more open and self-hosted LLMs.

As opposed to closed-source, and commercial LLMs such as the GPT models that can be accessed via OpenAI's API, self-hosted LLMs are built in such a way that does not provide data access to any third party entities. Therefore, in absense of well performing truly open-source LLMs, choosing self-hosted models appears to be the second best ethical solution.

The third caveat is the influence of researcher choices when formulating prompts that instruct the LLMs to perform the desired tasks. Albeit, this is not very different from composing a comprehensive codebook in order to avoid the issue of misunderstanding the annotation task. As designing comprehensive prompts is the most vital conceptual part of text analysis methods assisted by LLMs (Törnberg, 2023b), it is crucial to make this an iterative process. A simple

[56] https://bigscience.huggingface.co/
[57] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard$

change of wording may alter the annotations provided by LLMs to a great extent. Therefore, we must set aside a small sample to perform various prompt tests on until we achieve a desired outcome. We will delve into the exact steps of this process in the following section.

Despite these limitations we argue that the vastness of LLMs "knowledge basis", offer a fruitful avenue to researchers with respect to conducting interpretive text analysis on phenomena that are not directly observable or such that can only be inferred by considering hundreds of contextual parameters across thousands or even millions of words and documents. The detection of information flows across thousands of documents and platforms, arguably represents such a complex, multi-step and not directly observable tasks. As such, for the first time, we may be able to take a larger step towards optimizing the validation of information flows by leveraging the capabilities of LLMs. Although the current state of our knowledge concerning the validity and reliability of LLMs in performing various NLP tasks is limited to a handful of early findings, and LLMs suffer from a series biases and shortcomings, we argue that it is reasonable to assume that using LLMs may help us optimize the validation procedure of automated methods tracing information flows. More specifically, using the vast parametric knowledge (i.e., encoded in the model) of LLMs combined with detailed source knowledge (i.e., encoded in instructions/prompts) enables us to leverage LLMs to optimize the validation process for more complex tasks such as the detection of information flows across thousands of documents.

LLM-assisted human validation

The aim of our approach is by no means to replace humans in the generation of gold standard data. Instead, we test whether LLM-assisted human validation can optimize the validation procedure of flow-based methods. Specifically, to optimize the generation of gold standard data for information flows, we test to what extent can assistance provided by LLMs optimize the validation process of information flows. Specifically, after instructing our LLMs to assess whether an information flow exists across text pairs using a comprehensive sequential prompting and a series of input variables such as publication date, keywords, and proper nouns, (indicator variables), we receive evaluations from the LLMs, which explain whether two texts match on a topic-level or on the news events level. We provide annotators with these evaluations to test to what extent this type of assisted validation procedure can work in practice. Besides the evaluations we also provide annotators with highlighted terms inside the texts that the LLMs used to evaluate whether two texts shared a connection through the information they had in common. To assess whether these evaluations contribute to optimizing the validation process. The metrics we use to evaluate the effectiveness of the LLM-assisted human validation scheme are average time of completion of a task and annotation quality.

We chain these instructions carefully in order to keep the process as simple and intuitive as possible, just like we would do for human annotators. The output generated by the LLMs will not only be a set a labels highlighting whether two texts are related or not related from the perspective of information flows, but rather the language models will assess whether two texts match on a specific level of information flows and provide a short reasoning behind their decision. We will employ human annotators to *verify* the decision of the LLMs by exposing them to both the labels as well as the explanations generated by LLMs. The premise is that humans will have a much easier time to annotate when exposed to such a reasoning. Taken together we study, to what extent can LLMs support human validation of the different levels of information flows. Agsinst this backdrop we ask:

RQ1: *How does an LLM-assisted human validation scheme influence the annotation speed compared to no assistance?*
RQ2: *How does an LLM-assisted human validation scheme influence the annotator confidence compared to no assistance?*
RQ3: *How does an LLM-assisted human validation scheme influence the annotation quality compared to no assistance?*

How to set up LLM-assisted human validation

Pangakis et al. (2023) took a first step at providing a useful five-step workflow to validate LLM performance and to refine the prompting for LLMs. They argue that LLMs may be useful in validation tasks, especially used in a human-in-the-loop manner, but they caution researchers to make this an iterative process where deploying LLMs without proper assessment is not advised. The workflow of Pangakis et al. (2023) uses human annotations as a baseline to inform the LLMs. We modify this process and first let LLMs perform the annotation tasks which included the generation of topic-level and news-event-level evaluations. Then we ask human annotators to perform the same task by providing some of them with the evaluations generated by LLMs in order to test the assistance capacity of LLMs. We compare the performance of human annotations in terms of speed, quality, and confidence within annotators. Our expectation is that: *LLMs can serve as annotation assistants that can extract and keep key information for human annotators who can then make better-informed decisions.* The Figure 16 illustrates the phases of the suggested LLM-assisted human validation scheme.

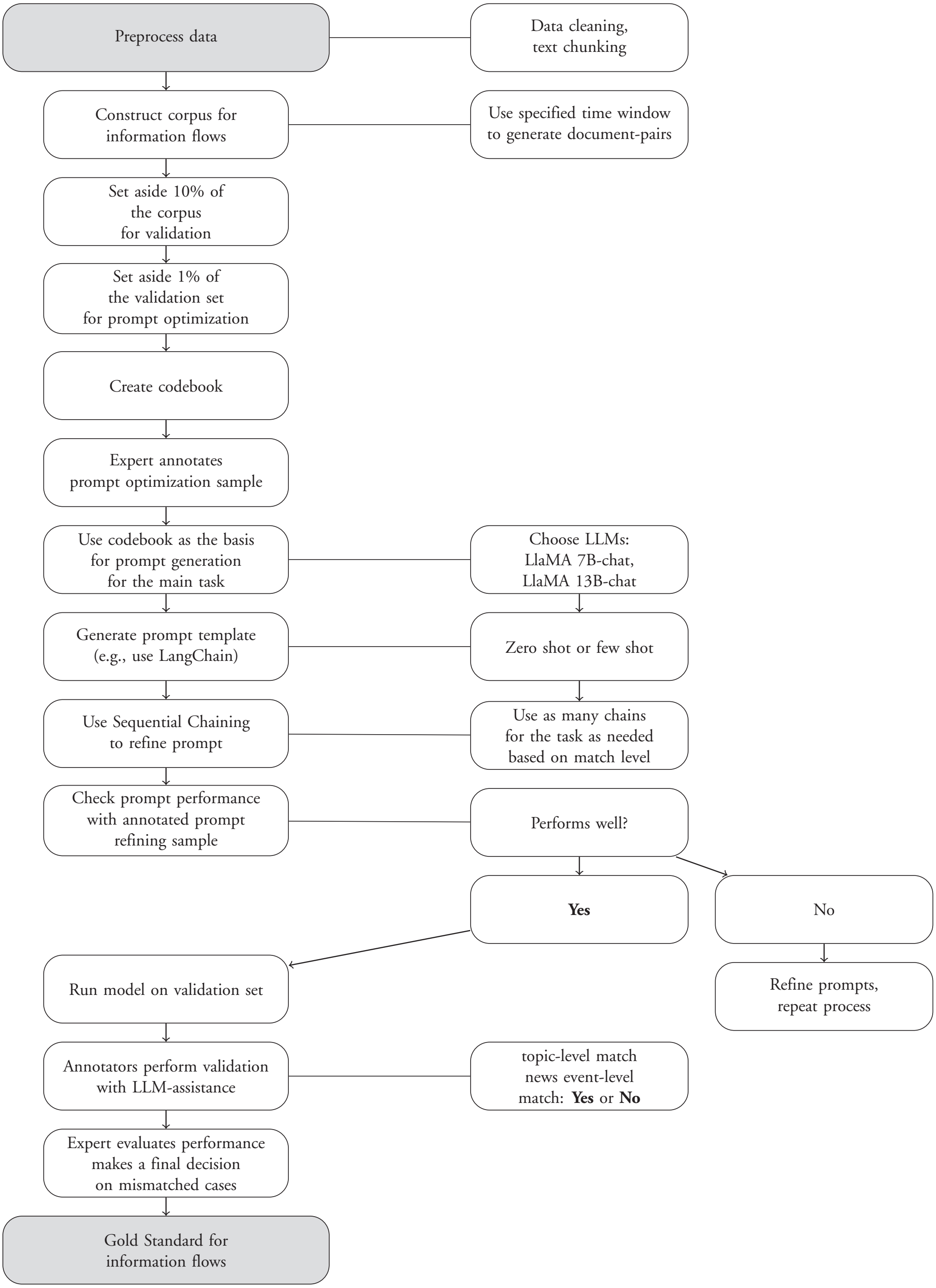

Figure 16: Phases of LLM-assisted human validation for information flows

Data and Methods

Data, sampling, and preprocessing

We test our proposed LLM-assisted human validation scheme on a dataset of Dutch news articles ($N = 6198$) collected by Aaldering et al. (2021) and published across 10 different online and print news outlets during the Dutch general elections of 2021. The original dataset covers a time frame of approximately three months (2021-01-02 to 2021-03-18) leading up to and including the Dutch general elections (15th and 17th of March, 2021) [58]. The dataset includes the five largest print newspapers of The Netherlands (De Telegraaf, NRC Handelsblad, Trouw, Algemeen Dagblad, de Volkskrant), and the largest economic newspaper, Het Finaniceele Dagblad. Besides print newspapers, the dataset also includes the two largest online new websites in The Netherlands (NOS.nl, and nu.nl) including two news sites affiliated with NOS (nieuwsuur and NOS liveblog). Overall, the data consists of a news articles that mention at least one of the 37 participating political parties or their leaders. The dataset was Dutch and besides the full text body, title, and publication date it included information about the publishers, publication date, author, byline, and sections of news articles.

Preprocessing

To clean the dataset, we first removed missing values on articles, then following manual inspection we removed those articles that merely referred to recurring, non-news-related aspects such as TV guides, film tips, or a list of "most read" articles [59]. Then we performed minimal preprocessing steps that included the removal of special characters, line breaks and double spacing. Since most LLMs to date have a limited "context window" (OpenAI, 2023; Touvron et al., 2023), we had to limit the length of the text we could feed into the LLMs. Following the inverted pyramid structure of news articles[60] (Knobloch et al., 2004) we opted for the convention of using the title and lead of each news article for further analyses (see e.g., Welbers et al., 2018). More specifically we took the first 200 tokens of each news article [61] and merged them together with their title for further analyses. Henceforth we refer to these title-lead combinations as article chunks. The next step was to generate additional input variables for our LLMs using the article chunks. After extracting proper nouns from the texts using Spacy's

[58] Due to the COVID-19 pandemic, the government decided to open some polling stations two days prior to the election date 17.03.2021 to ensure safe voting conditions for elderly and those with vulnerable immune systems

[59] 'NPO 2', 'NPO 1' 'BBC 1' 'BBC 2','TLC', 'Tips','kop tag with dummy text', 'Film- en televisietips', 'tv-ladder', 'TV-ladder', 'Best gelezen op fd.nl', 'MEER LEZEN'

[60] The main content of a news article can be found in the first paragraph

[61] In some cases the chunks were shorter in length since we removed additional tokens that occurred after the end of the last sentence

nl_core_news_sm model, we used KeyBERT (Grootendorst, 2023) with the Dutch embedding model, "BERTje" (on huggingface GroNLP/bert-base-dutch-cased) (de Vries et al., 2019), to extract keywords from each text chunk.

Sampling

Following best practices by Nicholls and Bright (2019) and Trilling and van Hoof (2020) we obtained a corpus of candidate pairs for information flows using a randomly chosen three-day window. Similar to Nicholls and Bright (2019), we took all article chunks that were published within a randomly chosen period of three days, assuming that newspapers would not cover the same news event for more than 3 days. This procedure resulted in obtaining a corpus of 142 news article chunks published by 10 news outlets between the three-day window of 2021-02-27 and 2021-03-01.

Since we did not possess a corpus of pre-annotated text-pairs for information flows, we constructed a corpus of good *candidate pairs* for information flows across the 142 article chunks. Previous research showed that pairwise text similarity is a good indicator of the presence of some type of information flows across news articles (Nicholls & Bright, 2019; Novo & Gedikli, 2023a; Trilling & van Hoof, 2020). Therefore, we used BERTje to generate embeddings from the article chunks, then following best practices we calculated cosine similarity across all possible chunks pairs in our corpus of 142 article chunks (Trilling & van Hoof, 2020). From the resulting 10,011 text pairs, 15 appeared to have perfect similarity. After manual checks, it appeared that this was the result of a data collection issue where the same text was collected twice with different ID numbers. Deleting these duplicates resulted in a final sample size of 9,996 article chunks pairs. No article chunk pairs exhibited a cosine similarity lower than 0.5 in our analysis. We attribute this high similarity among text chunks to two primary factors. Firstly, the texts under scrutiny were all published in close proximity to a significant election debate, resulting in a substantial number of articles addressing the event using similar language and vocabulary. Secondly, our calculation of cosine similarity was performed on short text chunks, specifically lead paragraphs, which inherently reduced the linguistic diversity compared to analyzing the entire body of an article. We believe that this low linguistic diversity combined with the proximity to the election debate, explains the high similarity across the text chunks in our sample. All in all, we found some partial evidence for the assumption about news articles published within a three-day time-frame being good candidates for information flows may indeed be valid (Nicholls & Bright, 2019; Trilling & van Hoof, 2020).

Then, using random sampling we set aside 10% of our candidate pair corpus for the purpose of performing the LLM-assisted human validation ($N = 990$). This sample will be annotated by both LLMs as well as three human annotators. Finally, we took 1% of the corpus for prompt optimization purposes. Prompt optimization, also referred to as prompt engineering, represents a crucial step in working with LLMs (Törnberg, 2023b). Since we are giving instructions to LLMs in natural language it is very important to use a small subset of the data to fine-tune these instructions for better performance. The prompt optimization sample was also annotated by an expert annotator (i.e., lead author) so that we could efficiently test the performance of prompts in completing the task at hand.

Methods

LLMs under study

Since truly open-source LLMs are virtually non-existent to date, we decided to opt for the second best alternative, self-hosted LLMs. Therefore, we employ two self-hosted and light-weight LLMs, namely LlaMA-13b-chat and and LlaMA-7B-chat. The choice about these particular LLMs was motivated by two reasons. The first one was their competitive performance with closed, commercial models such as ChatGPT and GPT4. The second reasons was their relatively small size, which makes them practical to use even with more modest computational resources.

LlaMA 2 represent a family of more open, self-hosted, and relatively light-weight LLMs pretrained and fine-tuned by Meta in mid 2023. The models' scale ranges between 7 billion to 70 billion parameters including two LLMs fine-tuned for chat functionality (Touvron et al., 2023). For the purpose of this study and since our system capacity, especially our GPU memory (i.e., VRAM), was limited (see Appendix Q) we opted to use Meta's second largest and potent model from the LlaMA 2 model collection, LlaMA 13B-chat (trained on 13 billion parameters), and the most lightweight model LlaMA-7B-chat (trained on 7 billion parameters). Both of these models were fine-tuned for instructions and can be accessed after a request is granted by Meta, which at the time of running these tests took less than 1 hour.

The greatest advantage of these models is that they can be implemented on relatively small systems such as two A10 GPUs with limited VRAM. We implemented these models via the HuggingFace pipeline with 4-bit quantization to condense the models' 64-bit representation so that we could run them on our modest dual A10 GPU machine that had limited VRAM [62] (Dettmers et al., 2023).

The context length of LlaMA-13B-chat model is 4096 tokens which is quite limited compared to the 8,192 token context length of GPT-3.5 and the whopping 32,768 token length of GPT-4[63]. As such, it becomes challenging to add a lot of instructions and or expect a lot of output

[62]Ideally we would need A100 GPU access with high VRAM to run LlaMA 2 models without any quantization.

[63]https://platform.openai.com/docs/models/overview

from a LlaMA 2 model as the context length is the summary of both the input tokens as well as the output tokens. Nevertheless with some clever prompt engineering, it possible to achieve promising results.

Prompt engineering with LangChain

Prompts represent the instructions formulated in natural language that users give to LLMs to perform a certain task. Prompt formulation is perhaps one of the most critical aspects of using LLMs for any kind of annotation task. However poor prompt formulation can result in unexpected as well as sub-par results. To mitigate the biases that may arise due to poor prompt formulations and hallucinations, we make use of LangChain (Chase, 2022), an open-source framework designed to simplify as well as optimize the usage of LLMs. We use LangChain to devise a comprehensive sequential prompt template for each LLM that can take instructions sequentially using so called *chains*. The idea behind this approach is essentially to 'chain' our instructions to the LLM so that it can break down a complex task into multiple smaller sub-tasks, and to source knowledge from additional indicators tailored to the task at hand to make more precise choices. Following this approach we expect to reduce the bias that may arise from poor prompt formulation, or asking too much at once.

As mentioned before, designing a good prompt represents an iterative process. One must test various formulations to make sure the LLMs understand the task well. To quickly assess the performance of our prompts, we manually annotated the previously obtained prompt optimization sample ($N = 99$) and used this as a benchmark to test to what extent LLMs followed our instructions by comparing the annotations of LLMs to annotations of an expert human annotator. Accordingly, we used parts of this small sample to tune our prompts.

Zero shot or few shot learning. While zero-shot learning implies that the LLM performs a task that it has not been explicitly trained for without any illustrations of how the task should be performed, few-shot learning means that the model is given a very limited amount of examples to learn from (i.e., a few shots) to perform a particular task. After several initial trial rounds, we opted to use few-shot learning for the purpose of this study.

Temperature. In addition to well-formatted natural-language prompts, researchers must also consider the models' *temperature*. This model parameter essentially determines how random and creative the output of the LLMs will be at each request. Model temperature can vary between 0 and 1, where values closer to 1 result in more diverse and unpredictable behaviour while values closer to 0 mean a more focused and deterministic output (Gilardi et al., 2023). The literature thus far has advocated for the usage of a temperature of 0.2 as the ideal level of randomness for various annotations tasks performed by ChatGPT (see e.g., Gilardi et al., 2023; van Hoof et al., 2023). However, the jury is still out with respect to how exactly model settings affect results of more open LLMs (Alizadeh et al., 2023). There are some initial indications that the rule of thumb regarding less randomness leading to higher quality annotations also apply to some of of these more open LLMs (Alizadeh et al., 2023). However, since evidence is scarce regarding the ideal temperature level, we test our models at four different temperature settings: 0.0., 0.2., 0.6, and 1.0.

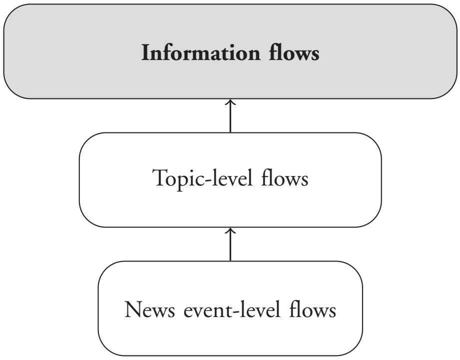

Figure 17: Studied levels of information flows

Two levels of flows

We study information flows on two levels as illustrated by Figure 17. We break up the task of identifying whether text pairs contain information flows on a topic-level or news event-level into multiple sub-tasks by chaining instructions to our LLMs incrementally. These sub-tasks can be conceptualized similarly to the indicator-based approach to measure latent constructs (Burscher et al., 2015). Information flows are arguably a latent construct therefore teaching these LLMs how to discern them across document pairs represents a challenging task. To help the models understand the task, we used a custom system prompt, example prompt, and a main prompt. A system prompt defines the overarching purpose of the model. An example prompt serves as an illustration, guiding the model to follow a specific format and emphasizing important details about the expected input and output. Meanwhile, the main prompt provides clear instructions to the model. Table 6 illustrates the prompt structure of Chain 2 from our study. The prompt also include token markers (e.g., SYS) that are used by LlaMA models specifically.

Topic-level flows. We begin with topic level information flows. To assess whether there is a topic flow between a document pair we need to provide several sequential instructions (i.e., chains) to our model that discern and utilize relevant indicators to make a final assessment regarding topic-level flows across document pairs. We refer to these sequential instructions as chains. A detailed

Table 6: Example of prompts used in this study

Prompt type
System prompt: <s>[INST] «SYS» You are a helpful, respectful, and honest assistant for comparing the main topics of two texts. In this comparison, a match is solely based on the main topic and nothing else. «/SYS»
Example prompt The main topic of each text is described by the following labels: Main topic 1: Politics; Main topic 2: Politics; Based on the information about the main topics above, please write a short evaluation about whether the two texts match on a main topic level. Make sure to only return the evaluation and nothing more in the following format: [/INST] Topic Evaluation: Yes, the two texts match on a main topic level because both texts touch upon the broader context of Politics as seen by their main Topic.
Main prompt [INST] The main topic of each text is the following: {main_topic} Based on the information about the topics above, please write a short evaluation about whether the two texts match on a main topic level. Make sure to only return the evaluation and nothing more in the following format: Topic Evaluation: [/INST]

Note: The tags [INST], [/INST] are used to indicate the beginning and end of example and main prompts for LlaMA models. The tags <s>[INST] «SYS», and «/SYS» mark the beginning and end of a system prompt.

overview of these sequential prompts and corresponding input variables used at each step can be found in Appendix R. In a nutshell, we use one sequential chain with 4 main chains[64] to instruct those model to determine to what extent a there is an information flow on the topic-level across two documents. At chain 1 the model identifies the main topics and most important subtopics in our documents. To do this, we write a prompt that instructs the model to extract the main topics and subtopics of the documents based on a few existing input variables (document pair, proper nouns, and keywords) with the additional instruction to base these main topics on the categories established by Vermeer (2018) [65]. We modified this framework slightly. Rather than forcing the models to only classify the texts into strict categories, we fed some examples of main topics and subtopics into the LLMs as prompts to merely determine the level of abstraction the models had to focus on when generating the main topics and subtopics. At chain 2, the model evaluates whether the texts match on the main topic level following instructions provided in the

[64] we also use some additional sub-chains

[65] e.g., Politics, Business, Entertainment and Other. These categories merely served to indicate the level of abstraction required from topic extraction.

corresponding prompt and the output provided at Chain 1. At Chain 3, the model extracts words used to make topic-level evaluation decision. This is necessary as these words are later used for LLM-assistance in the annotation/validation procedure. At chain 4 we instruct the model to create a classification label based on the evaluation it generated about the topic-level flows. At this step the model will use the evaluation at chain 2 to classify whether there is (1) or there is not (0) a topic-level flow between a document pair.

News event-level flows. In the same vein, we continued the sequential chain in order to instruct the model to determine whether a news-event-level flow exists across a document pair. At chain 5 the model identifies news events based on the definition of Trilling and van Hoof (2020) who discerned that "news events are specific events that lead to news coverage, such as a specific debate on a specific day in a specific parliament, a specific accident, or a specific football match. They can be covered by one or more articles in one or more outlets, but relate to one specific and identifiable event and are thus much more fine-grained than news topics, issues, or news categories" (p. 1321), and by using a series of input variables generated by previous chains. At chain 6 the model is instructed to evaluate to what extent the texts match on the news event-level, whereas as chain 7 it generates a classification label similar to the instructions given at chain 4 for the topic-level matching task.

Results

This section outlines the results of the proposed approach. We begin by providing an overview of the performance of the prompts of the two selected few-shot LLMs at different temperature settings. We measured the performance of prompts against a manually annotated prompt validation set which was done by an expert annotator (i.e., lead author). More specifically following the same instruction we gave to the LLMs, the lead author annotated 99 text pairs for information flows on the topic level and the news events level. Table 7 illustrates performance metrics of our chosen models at different temperatures, a parameter that determines how creative or deterministic the model will be when generating responses. As illustrated by Table 7 our few-shot models performed the same way at each of the four different temperature settings (0.0, 0.2., 0.6, and 1.0). We ran the models twice at each temperature level with no difference in performance. Therefore, it is possible that providing the models with clear example prompts counteracts the creativity/randomness that would otherwise be unlocked by setting the temperature close to 1. Hence, it appears that if LLMs are supplied with a system prompt, example prompt, and main prompt their output will be virtually identical for the task of capturing topic-level and news events-level matches across document pairs. Since all models performed the same regardless of the temperature parameter we will interpret the scores between LlaMA 7B-chat and LlaMA 13B-chat in general on the task at hand.

Since our data is imbalanced with more cases of topic matches and less cases for non topic matches, and less cases for event matches and more cases for no matches on the events level, we consider F1 as the most informative metric to determine the highest performing models. While LlaMA 7B-chat appears to perform better than LlaMA 13B-chat on detecting topic level information flows, LlaMA 13B-chat outperforms LlaMA 7B-chat on all news event-level matches. Although, the overall performance of capturing news event-level matches is arguably low for both models regardless of temperature settings. It must be noted that theses scores represent the best performance we could achieve with prompt optimization using the conceptualizations from the literature. Naturally, we cannot rule out the possibility that better prompts may exist, which would result in better performance than what our prompt optimization process revealed. Nevertheless, we are confident that using prompts that follow the current definition of news events from literature would yield similar results to our attempts.

Thus, we argue that the current definition of what news events are is neither specific enough for LLMs nor for human annotators. Rather, the existing definition of news events is more suitable for less deterministic approaches such that rely on clustering and network analytic methods (see e.g., Trilling & van Hoof, 2020). Regardless of the poor performance of our prompts and models on the news events level, we argue that reporting these findings and annotating for them remains an important contribution that we can learn from. Specifically, analysing the errors and inconsistencies that are generated by the models can reveal ways in which we can clarify the definition of news events for future research.

Error Analysis of Prompt validation sample

To understand the disagreements between the expert annotator and the LLMs we performed a manual error analysis on the prompt validation sample for both topic-level as well as news-events-level matches. Specifically, we extracted a list of cases where LlaMA 13B-chat and LlaMA 7B-chat disagreed with the expert annotator, and we checked why these disagreement occurred. We report the frequencies of the most common pitfalls for each model per match level annotated in Table 8.

The most common type of disagreement between the studied LLMs and the expert annotator were the result of incorrect topic labels assigned by the LLMs, which lead to incorrect evaluations, and ultimately incorrect classification labels. For instance, consider text 1 that discusses the election campaign and text 2 that details the Dutch government's plans with

Table 7: Prompt Performance Metrics Comparison at different model temperatures

Match level	Model	Precision	Recall	F1 Score	Accuracy
Topic Match					
	Llama 7B-chat 0.0	0.798	0.905	0.848	0.758
	Llama 7B-chat 0.2	0.798	0.905	0.848	0.758
	Llama 7B-chat 0.6	0.798	0.905	0.848	0.758
	Llama 7B-chat 1.0	0.798	0.905	0.848	0.758
	Llama 13B-chat 0.0	0.830	0.527	0.645	0.566
	Llama 13B-chat 0.2	0.830	0.527	0.645	0.566
	Llama 13B-chat 0.6	0.830	0.527	0.645	0.566
	Llama 13B-chat 1.0	0.830	0.527	0.645	0.566
News Event Match					
	Llama 7B-chat 0.0	0.200	0.500	0.286	0.596
	Llama 7B-chat 0.2	0.200	0.500	0.286	0.596
	Llama 7B-chat 0.6	0.200	0.500	0.286	0.596
	Llama 7B-chat 1.0	0.200	0.500	0.286	0.596
	Llama 13B-chat 0.0	0.345	0.625	0.444	0.747
	Llama 13B-chat 0.2	0.345	0.625	0.444	0.747
	Llama 13B-chat 0.6	0.345	0.625	0.444	0.747
	Llama 13B-chat 1.0	0.345	0.625	0.444	0.747

Note: Models were tested twice.

respect to economic growth by quoting prime minister Mark Rutte. The LLMs labels the main topic of text 1 as Politics and the main topic of text 2 as Economy. The expert on the other hand, marked both texts' main topic as Politics since the codebook explicitly stated that if words related to politics or politicians are mentioned then the main topic is always considered Politics. So even though text 2 discussed the economy and the topic discerned by the LLMs was not completely incorrect, since it violated the codebook it was considered a mismatch. This type of disagreement also relates to the inconsistencies with respect to the definition of what a topic is and how exactly one should operationalize the definition of a topic.

Table 8: Error Analysis

Mismatch type	Llama 7B-chat	Llama 13B-chat
Main topic mismatch		
Incorrect main topic, incorrect evaluation, incorrect label	19	39
Correct main topic, incorrect evaluation, incorrect label	4	-
Total	23	39
News event mismatch		
Incorrect news event, incorrect evaluation, incorrect label	25	24
Correct news event, incorrect evaluation, incorrect label	12	-
Correct news event, correct evaluation, incorrect label	1	-
Total	38	24

LlaMA 7B-chat appeared to suffer from another matching issue, namely, discerning the correct main topics of a text pair but getting confused by the instructions with respect to the evaluation. The model correctly recognized that both texts' main topic was Politics, but it made an incorrect evaluation despite the correct input variable, which resulted in an incorrect classification label of no topic match.

Turning to the news events-level mismatches, the most common pitfall was the incorrect identification of the news events of the texts which resulted in incorrect evaluations and incorrect labels. The most frequent reason behind this type of error was that the LLMs considered news events to be broader than the definition used in the codebook allowed it to be. For instance, in some cases the LLMs considered the Dutch election of 2021 a news event. However, according to the definition by Trilling and van Hoof (2020) this match would be considered too broad, as "news events in principle take place within a window of three days " (p. 1325). Nevertheless, if we follow the logic of Trilling and van Hoof (2020) but disregard this three-day threshold the response of the LLMs is arguably not entirely ludicrous.

Another type of error that occurred at times was that the LLMs were unable to recognize that two texts matched on the news events level as the texts were not very explicit. For instance, in one case, both texts clearly discussed the election debate, their publication dates were extremely close (a matter of a few minutes). While the subtle signs of news-events-level connection were apparent to the expert annotator, the LLMs failed to recognize these. Finally, another common type of error was correctly identifying the news events but coming to erroneous conclusions despite the correctly identified news events. This was only the case for the less sophisticated LlaMA 7B-chat indicating that this LLMs is not as reliable in following a chain-of-thought instructions as LlaMA 13B-chat. This makes intuitive sense as the model has a much more limited parametric knowledge, which limits its capacity to follow instructions.

Testing the LLM-assisted human validation scheme

The testing of the proposed LLM-assisted human validation scheme had several phases. Figure 22 illustrates the phases of our proposed LLM-assisted human validation scheme. First, three human annotators (A, B, C) and the expert annotator (E) coded an overlap sample of document pairs for topic-match and news event-level match. Following two rounds of overlap annotation, good inter-coder reliability was achieved for both the topic-level matches (Krippendorff's $\alpha = 0.78$) and for news event-level-match (Krippendorff's $\alpha = 0.75$) across the annotators and the expert annotator (for details about the initial round of overlap annotation see Appendix S).

Second, after reaching an acceptable inter-coder reliability across the three human coders and the expert annotator, we tested our LLM-assisted human validation scheme. Since the LLMs appeared to perform identical at all temperature levels, we opted to use both studied models at temperature 0 for the sake of clarity. We asked both LlaMA 7B-chat and LlaMA 13B-chat to perform the annotation on half of the previously obtained validation set ($n = 495$). The other half ($n = 495$) we left unassisted in order to obtain within-person measures of annotation speed, confidence, and quality across assisted and unassisted conditions. As such, we were able to compare assisted versus unassisted conditions.

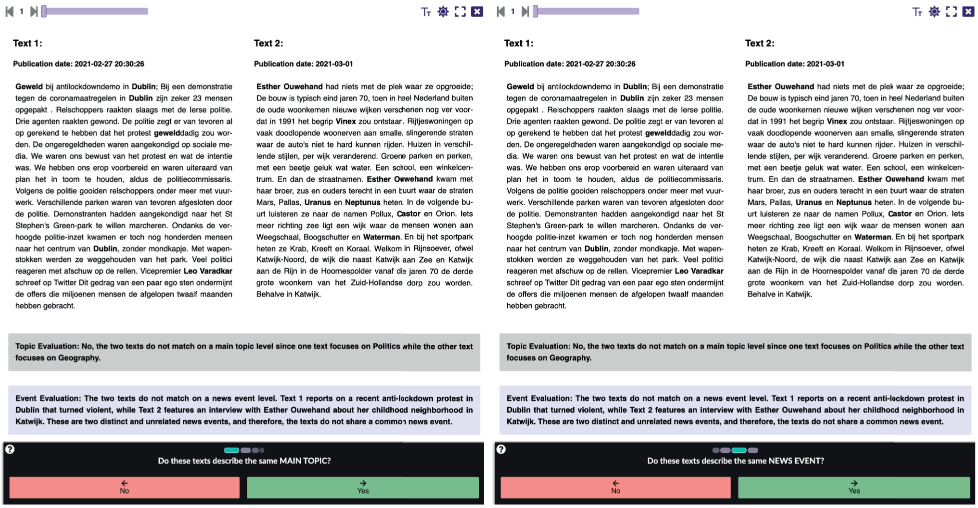

Figure 18: Example of comprehensive assistance in AnnoTinder for topic-level, and news event-level matching.

Third, in the LLM-assisted condition each of the three human annotators was paired with output from a LlaMA model. Annotator A was supported by the output of LlaMA 7B-chat, annotator B by LlaMA 13B-chat, and annotator C received partial assistance from both LlaMA 7B-chat and LlaMA 13B-chat. Annotators A and B benefited from comprehensive LLM assistance, including access to text pairs, publication dates, highlighted keywords, proper nouns, topic match evaluations, and news event match evaluations for each pair (see Figure 18). In contrast, annotator C's support was limited to highlighted keywords and proper nouns besides the text pairs and publication dates (see Figure 19). All three annotators also annotated the same set of tasks without any LLM-assistance (see Figure 20). For both assisted and unassisted conditions, annotators were required to convey their confidence level in the accuracy of their assessments with respect to each of their answers regarding topic-level and news event-level matching (see Figure 21).

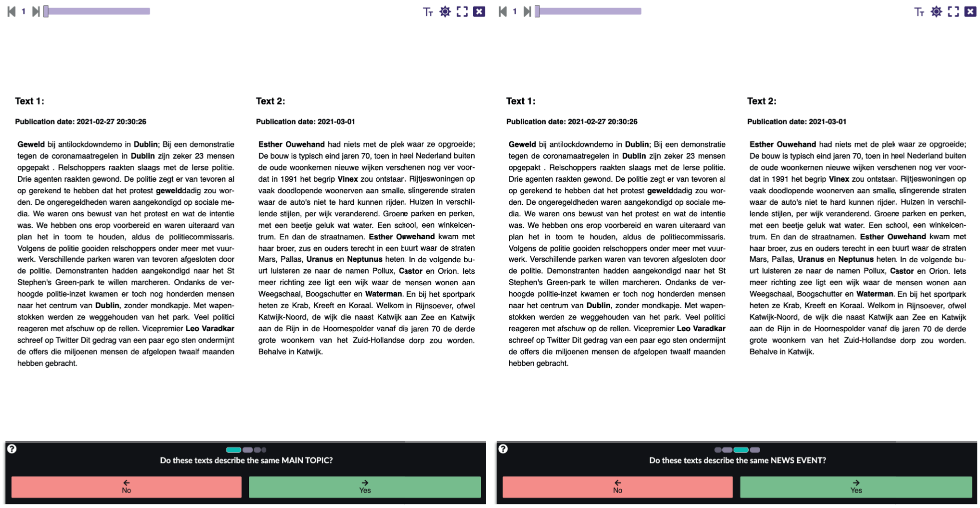

Figure 19: Example of partial assistance in AnnoTinder for topic-level, and news event-level matching.

In the last phase of the validation scheme, the expert annotator, E evaluates performance. The aim of this annotation setup was to help us measure the extent to which LLM-assisted human validation can optimize the validation of information flows. Specially, we asked: *To what extent can LLM-assisted human validation influence the speed of the validation process compared to non assisted validation?* We assessed this by comparing the average time spent per task in an LLM-assisted setup versus no assistance. We also asked: *To what extent can LLM-assistance influence the confidence in annotations compared to non-LLM-assisted validation.* Finally, we checked the quality of annotations by calculating inter-coder reliability across the three annotators and then using an expert coder to verify the output of the assisted, partially assisted as well as unassisted sets.

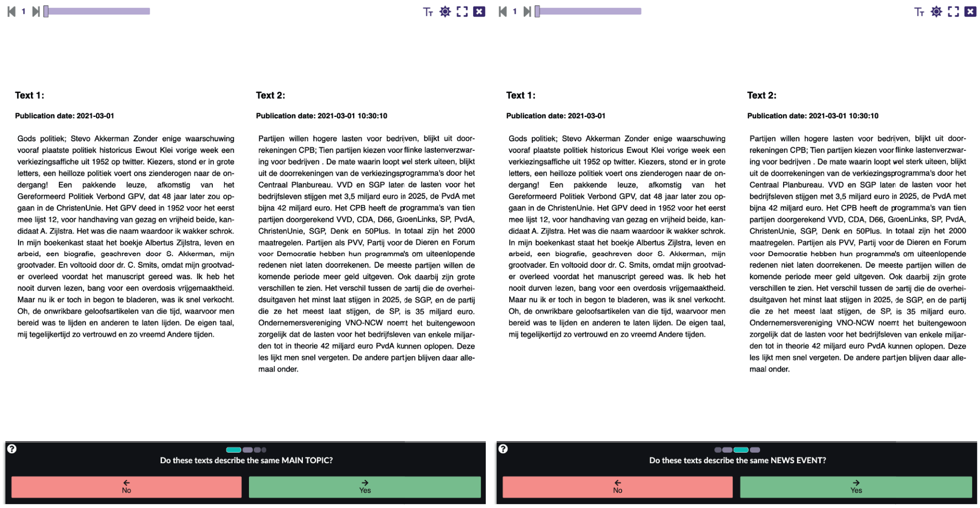

Figure 20: Example of no assistance in AnnoTinder for topic-level, and news event-level matching.

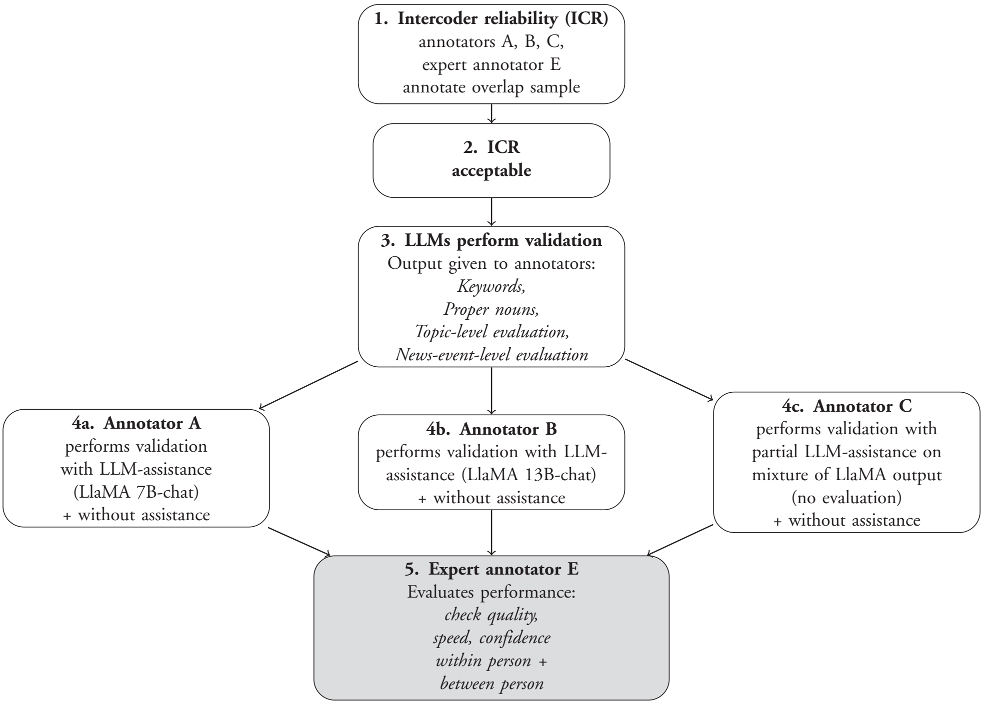

Figure 22: Phases of testing LLM-assisted human validation

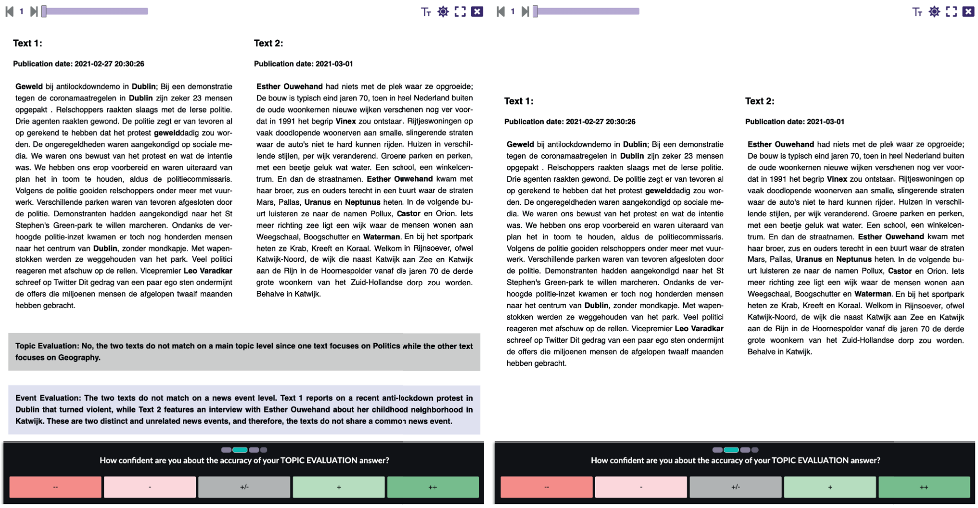

Figure 21: Example of confidence question in AnnoTinder for topic-level, and news event-level matching (comprehensive and partial).

Annotation speed

We begin by highlighting results with respect to annotation speed. Table 9 indicates mean annotation speed levels overall (both for topic-level and news-event-level matching), for topic-level matching, and for event-level matching speed for annotators A, B and C in the assisted as well as unassisted conditions. The pattern for annotator A and B appears to follow the some trend. Namely, comprehensive LLM-assistance seems to lead to longer annotation time compared to the unassisted condition for both annotators. Albeit, independent samples T-tests showed no statistically significant differences between the assisted and unassisted sets for neither topic-level matching, event-level matching, or in general for annotator A and B. Meanwhile, we observed an inverse pattern with respect to the results of the annotation speed of annotator C. Partial assistance (only highlighted keywords, and proper nouns) on a mixture of the output by LlaMA 7B-chat and 13B-chat seems to speed up the annotation procedure for all measured aspects. For annotator C, the difference was only significant for the event-level matching annotation task. Thus, it appears that partial assistance proves more useful for more complex tasks such as determining whether two texts matched on the news event-level.

Annotation confidence

Next, we outline results with respect to annotation confidence for all three annotators. As mentioned before, after completing each annotation task, annotators had to indicate to what

Table 9: Mean annotation speed (in seconds) *SD* in parentheses

Annotator	Type	$\mu_{overall}$	$\mu_{topic-level}$	$\mu_{event-level}$
A	assisted (LlaMA 7B)	12.86*(130.25)*	23.94*(183.58)*	1.77*(3.93)*
A	unassisted	6.45*(19.53)*	11.17*(26.67)*	1.72*(2.84)*
B	assisted (LlaMA 13B)	17.43 *(78.72)*	20.77*(73.00)*	14.08*(83.99)*
B	unassisted	14.81*(85.22)*	20.28*(68.70)*	9.34*(98.80)*
C	partially assisted (LlaMA 7B+ 13B)	5.11*(40.26)*	8.85*(56.69)*	1.37*(1.74)*
C	unassisted	9.00*(68.73)*	14.41*(96.69)*	3.59*(7.00)*

extent they felt confident about their choice with respect to the (1) topic-level match and with respect to (2) news event-level match on a scale from not confident at all (- -) to very confident (++). These values were recoded to form scale between 0 (not confident at all) to 4 (very confident). Results shown on the Table 10 indicated that regardless of LLM-assistance annotators were highly confident about their answers regarding topic matching and event matching tasks. As such, the presence of LLM-assistance neither increases nor decreases the annotators' confidence in performing flow-based annotations. As the scores are almost identical in both conditions for all annotators, we did no perform further statistical tests.

Table 10: Descriptive statistics of annotation confidence (0-4)

Annotator	Type	$\mu_{overall}$	$\mu_{topic-level}$	$\mu_{event-level}$
A	assisted (LlaMA 7B)	4.00	4.00	4.00
A	unassisted	3.97	3.96	4.00
B	assisted (LlaMA 13B)	3.94	3.94	3.95
B	unassisted	3.96	3.96	3.96
C	partially assisted (LlaMA 7B+ 13B)	3.97	3.96	3.99
C	unassisted	3.95	3.95	3.95

Annotation quality

Given that annotation speed and annotator confidence only matters if annotation quality is acceptable, we must also assess annotation quality to obtain a better understanding of the potential utility of an LLM-assisted validation scheme for information flows. Speed and confidence are rather annotator dependent but the annotations themselves should not be so given the previously calculated good ICR score. The most important measure of whether LLMs are able to improve the process of annotations and validations is the annotation quality. We assess this by first calculating inter-coder reliability across the assisted and unassisted sets and then utilizing an expert who verifies the output of the majority vote.

First, the inter-coder reliability across annotators A, B, and C on the unassisted task for topic-level matching was good ($\alpha = 0.83$). The inter-coder reliability for event-level match was acceptable ($\alpha = 0.60$) for the unassisted condition. The ICR across the annotators in the LLM-assisted condition was $\alpha = 0.80$ for topic-level matching, and $\alpha = 0.64$ for news event-level matching. Based on these findings it appears that LLM-assistance can lead to a minor improvement in annotation quality for news-event-level matching but this difference is negligible. Furthermore, LLM-assistance may decrease the ICR slightly. However, this difference is not meaningful enough either.

Alignment between LLM-annotation and human annotation. Digging deeper into the potential influence of LLM-assistance on annotator's performance, we first calculated inter-coder reliability between the annotations of the two LLMs, LlaMA 7B-chat and LlaMA 13B-chat, and Annotator A and B respectively [66]. The ICR between LlaMA 7B-chat, and annotator A was low for the topic-level matching task ($\alpha = 0.29$), and extremely low for event-level matching task ($\alpha = -0.11$).

In the same vein, the ICR between LlaMA 13B-chat, and annotator B was low for the topic-level matching task ($\alpha = 0.22$) and extremely low for event-level matching task ($\alpha = 0.10$). These results suggest a couple of things. First, given the low agreement between the LLMs and the corresponding human annotators but the high agreement between the human annotators, we can conclude that human annotators rightly disregarded the evaluations provided by LLMs during the annotation procedure and instead followed the instructions they were given. Second, despite receiving the same instructions (i.e., codebook) the annotations provided by LLMs do not align well with the annotations provided by human annotators. Hence, it appears that the task was much more understandable to humans than to the LLMs used for the purpose of this study. This is unsurprising as our initial error analysis already revealed some significant shortcoming for both LlaMA 7B-chat and LlaMA 13B-chat in understanding the particular instructions given to them. Nevertheless, what we should consider a positive outcome is that human annotators in this study did not appear to get swayed by the sometimes erroneous evaluations provided by LLMs.

In addition we also calculated agreement between the two LLMs. We found that ICR for topic-level matching task was very low ($\alpha = 0.1$), and the ICR for the event-level matching task was low ($\alpha = 0.35$). This indicates that not only did the LLMs judgement differ to a great extent from human judgments, the two models also very much disagreed with one another despite receiving identical instructions. It must be noted that crafting and testing the instructions (prompt engineering) was a lengthy process (approximately 3 weeks). Despite this long iterative procedure the best performance we achieved with these LLMs was still rather weak. Although we cannot be certain whether the LLMs tested here could have performed better if we had crafted our

[66] we did not compare the LLMs to the performance of annotator C as the output annotator C received was only in the form of highlighted keywords generated by both LLMs (half by LlaMA 7B-chat and half by LlaMA 13B-chat)

instructions differently, we argue that investing any more time into prompt engineering would have defeated the purpose of optimizing the validation procedure of flow-based methods.

All in all, our findings suggest that while the proposed LLM-assisted human validation scheme may show some promise for the validation of complex flow-based methods on the news events level, the current performance of the self-hosted, few-shot LLMs, LlaMA 7B-chat and LlaMA 13B-chat, is simply not good enough to be useful for such a difficult task. Neither, LlaMA 7B-chat nor LlaMA 13B-chat appear to be good at following strict annotation instructions, which have historically been a paramount to generating high-quality gold standard data for automated text analysis methods. Nevertheless, we believe the insights provided in this study may prove useful to scholars who wish to turn to LLMs for solving complex annotation and validation problems similar to the detection and validation of information flows.

Expert finishes gold standard dataset for information flows

As indicated by Figure 16, the final step of the proposed validation procedure is the expert annotator's task to make a final decision on the mismatched cases across the annotators for both the topic-level matching task and the news event-level matching tasks for the assisted and unassisted conditions. We argue that this last step is crucial to ensure high quality gold standard data. 101 out of 495 cases appeared to show some disagreement between annotators in the unassisted condition. In 48 cases the annotators did not agree on the topic-level matching task, and in 58 cases annotators did not agree on the event-level matching task. In the same vein, in 111 out of 495 cases annotators disagreed in the assisted condition. 46 of these disagreements pertained to the topic-level matching task whereas 65 of them pertained to the news event-level matching task. The expert annotator went through all of the mismatched cases and annotated them once more to correspond to the annotation instructions. After this final expert verification step, we were confident that we have achieved a high quality gold standard dataset. We make both the dataset as well as all associated Python and R scripts public on GitHub [67].

Conclusion and Discussion

The validation of complex and flow-based methods poses a challenge to several scholars who apply automated methods to detect information flows across large, multi-platform corpora. The validation of document-comparison and flow-based methods has historically been expensive, time-consuming and prone to errors due to the exponential nature of the task. Despite

[67] https://github.com/psyronika/LLM_validate

significant advancements in the realm of methods available to trace information flows, innovation with respect to the validation of these novel powerful methods has so far been lagging behind. Chapter 2 and Chapter 3 have successfully employed some of these state-of-the-art methods to trace information flows both in a static and a dynamic way over time. Yet, assessing the extent to which the findings of in these chapters were truly valid, remained unclear due to the complex and dynamic nature of the task at hand. The current chapter aimed at advancing the validation of such dynamic methods by leveraging the power of LLMs as well as an expert annotator in optimizing this procedure. Specifically, we devised and tested an LLM-assisted human validation scheme. We tested our proposed scheme on a corpus of 990 candidate text pairs for information flows based on a Dutch-language collection of news articles published by 10 news outlets surrounding the 2021 general elections (Aaldering et al., 2021). Our proposed method involved the testing of two light-weight, self-hosted LLMs, LlaMA 7B-chat, and LlaMA 13B-chat, three human annotators, as well as an expert annotator.

Although our findings are discouraging with respect to the usefulness of the tested more open, and lightweight LLMs for optimizing the validation tasks for information flow-based methods, our study makes important contributions to the field of computational communication science and to the ongoing debate about the usage of LLMs in research for tasks that have traditionally been performed by humans. We provide important guidelines to scholars who are applying methods tracing information flows and those who are considering the usage of LLMs to optimize annotation tasks. We discuss the usage of LLMs in academic research and clarify the openness, or lack thereof of “open-source” LLMs (Liesenfeld et al., 2023). Stressing the importance of testing these models thoroughly, we advocate for comprehensive error analyses as well as verification of the annotations (both LLM- and human-annotated) by expert annotators in order to ensure the generation of high quality gold standard datasets.

Our error analysis offers valuable insights to scholars who wish to use more open and light-weight LLMs for complex annotation or validation tasks that are inherently dynamic in nature. We identify the most common pitfalls regarding the performance of two LLMs on flow-based tasks, and recommend solutions to circumvent these shortcomings. Additionally, we also release a validated dataset for information flows ($N = 990$) both on the main topic-level as well as on the news event-level. We argue that this dataset can serve as a valuable resource to researchers who can make use of this gold standard dataset to train new models or validate their existing models that aim to capture similar flow-based phenomena.

Although the models tested in this study were neither the best performing LLMs available nor the easiest to instruct, we opted for these models for two main reasons. Accessibility was the main reason why we opted for the usage of LlaMA 7B-chat and LlaMA 13B-chat respectively. These models are highly accessible to many computational communication scientists both financially as well as practically. Their usage is completely free, and accessing them is a few clicks away after being granted permission by Meta. Moreover, these models can be set up and used on relatively modest systems such that was utilized for the purpose of this study. The second main reason behind the choice for LlaMA 7B-chat and LlaMA 13B-chat was the ambition for more openness and transparency. Despite the fact that the self-hosted and light-weight models used here arguably fall short on openness (Liesenfeld et al., 2023), they still represent a safer choice compared to models such as GPT4 that are only accessible through commercial API's, thereby jeopardizing research integrity and transparency especially for potentially sensitive or proprietary data (Törnberg, 2023b).

Our findings showed that the tested models are not yet powerful nor reliable enough to be used off the shelf without the need for thorough human verification. Nevertheless, there is some merit in the process of setting up such LLM-assisted validation schemes. Although the LLMs to date may not perform well enough on the task of validating flow-based methods, our findings showed that providing partial assistance to annotators in the form of highlighted keywords, may already improve the validation procedure of flow-based methods in terms of annotation speed. Naturally, obtaining such hints as highlighted keywords need not be performed by LLMs. There are other, simpler, less computationally expensive ways to go about such subtle, yet useful forms of assistance. For instance the transformer-based keyword extraction tool KeyBERT may be a useful tool for this (Grootendorst, 2023).

The usefulness of highlighted keywords during the annotation procedure is in line with research that emphasized the importance of incorporating visual cues into the annotation procedure and software to increase annotation productivity for NLP-based annotation tasks (Stenetorp et al., 2012). Thanks to AnnoTinder (Welbers, 2022), the annotation software used for the purpose of this study, we had the advantage of implementing the validation task as well as the LLM-assistance in a relatively simple, yet highly intuitive and visually appealing manner that enabled us to facilitate a smooth annotation procedure for all human annotators. The relatively good performance of human annotators in terms of annotation speed and quality demonstrated that rather than having to scroll through endless Excel sheets to perform annotation tasks, using a dedicated annotation software can go a long way, especially for more complex, document comparison-based tasks.

Despite the current study's discouraging results, the idea of AI-assisted annotation is still a promising one. Scholars recently tested and a novel AI-assisted annotation tool for a highly challenging image-based semantic segmentation annotation task (Pavoni et al., 2022). Findings showed significant improvements in annotation speed while preserving labelling accuracy. Therefore, we argue that continuing to test the extent to which we can use AI, and novel technologies such as LLMs to resolve complex problems such as the validation of flow-based tasks remains crucial. Furthermore, we argue, that certain annotation tasks may be too complex to perform for the human brain in a reliable and consistent fashion. This is where providing some form of assistance or guidance to humans while performing the task could make a significant difference. Ultimately, all we may need is an assistant that consistently and reliably follows instructions and keeps key information in memory in order to help the process of generating gold standard data. Thus, the idea of LLM-assisted annotation must be further tested until we achieve a safe and reliable model that can perform this task.

Although our study was largely exploratory in nature, we provide some valuable insights and resources to scholars. Tackling the validation of complex and flow-based methods remains a challenging tasks that necessitates the exploration of novel technologies that can have the potential to optimize this crucial but labour intensive, expensive, and overall inefficient task. Overall, we contribute to the debate surrounding the importance of validation for computational methods and the need for innovation with respect to the ways in which validation of such methods has historically been performed (Birkenmaier et al., 2023). Methodologically, we demonstrated how to set up and test an LLM-assisted human validation scheme for flow-based methods by testing the capacity of two self-hosted and light-weight LLMs. Further, we highlight the crucial role of human annotations and the potential for the introduction of expert annotators in improving the quality of gold standard data.

Limitations and future research

The current study had several limitations. First, given the low performance of the tested more light-weight LLMs in following strict instructions, the real value of the explanations provided by these models could not be assessed in this study. Besides their technical limitations, it is possible that that these light-weight and self-hosted models were not suitable to capture information flows based on the theoretical definition of what information flows are. Future research could test these model's ability to assess whether and information flows exists between text pairs using a zero-shot approach. Perhaps part of the reason why it has proven difficult to capture information flows especially on the news events level is related to the conceptual fuzziness of the definitions used to define these flows. An alternative direction for future research could be to let LLMs come up with their own understanding of how information flows across texts would manifest and then verify the results by human coders to assess whether the choices made by the LLMs make logical sense. Such an experiment could contribute to refining the definition of information flows on finer-grained level such as on the news-event-level.