

RESEARCH

Free and Open Access

# The potential of large language models for the assessment of cognitive engagement in informal online communities

Simon Krukowski <sup>1\*</sup>, H. Ulrich Hoppe <sup>2</sup> and Daniel Bodemer <sup>1</sup>

\*Correspondence:  
[simon.krukowski@uni-due.de](mailto:simon.krukowski@uni-due.de)  
Research Methods in  
Psychology—Technology,  
Learning, Collaboration,  
University of Duisburg-Essen,  
Lotharstr. 65,  
47057 Duisburg,  
Germany  
Full list of author information is  
available at the end of the article

## Abstract

Studying the cognitive engagement shown in forum messages of informal online communities can help to associate discussion behaviour with underlying learning processes. The ICAP framework can serve as a theoretical lens for labelling forum messages by distinguishing different levels of cognitive engagement, yet requires time-consuming, high-inference human coding, which is not feasible at scale. Although this can be mitigated by supervised classification models, these rely on large amounts of labelled training data, and their decision-making often remains opaque. In contrast to this, Large Language Models (LLMs) work “out-of-the-box” for qualitative coding tasks via natural language instruction and can produce reasoning traces and explanations for their judgments. In this study, we explore this potential by integrating the ICAP framework into natural-language prompts for few-shot in-context classification, comparing LLM judgments and explanations with those of human raters. Results indicate that LLM judgments generally align with the human ratings, while their reasoning traces and variations offer additional insight into where the coding scheme invites ambiguous interpretations. This suggests that LLMs could serve as cognitive proxies that help diagnose and improve high-inference coding processes. Finally, we apply our approach to large-scale forum data, revealing cross-platform differences in expressed cognitive engagement.

**Keywords:** informal learning, cognitive engagement, large language models (LLMs), qualitative coding, forum discussions

## Introduction

Discussing with peers is widely acknowledged as a powerful catalyst for learning, encouraging us to articulate ideas, consider alternative perspectives, and collaboratively build new understanding. In digital settings, discussion forums serve as social spaces that enable such peer exchanges. Their role is well-researched in contexts like Massive Open



© The Author(s). 2026 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Online Courses (MOOCs), in which active forum participation has been positively linked to improved learning outcomes and course performance (e.g., Houston et al., 2017; Tseng et al., 2016). However, discussion forums can facilitate learning beyond formal educational settings. Informal contexts such as online Citizen Science (CS) projects, where volunteers support researchers by engaging in scientific activities, also offer integrated discussion forums where volunteers can discuss their findings. These discussions can eventually foster the acquisition of scientific skills and domain knowledge (Aristeidou & Herodotou, 2020). Similarly, open online forums like Reddit where users learn about specific topics by asking and discussing them with others (e.g., *r/askscience*) can also serve as networked learning communities (Del Valle et al., 2020). In both cases, learning is primarily spontaneous, emergent, and detached from predefined goals or assessments (Prestridge et al., 2024), and occurs in large communities where it is typically driven by task needs and intrinsic motivation. Consequently, the scale of these communities and the absence of tangible learning outcomes like assessments or grades make identifying learning and providing support particularly challenging.

However, forum discussions of both formal and informal settings leave collectible traces of user interactions. The interactions themselves can be analysed through techniques such as social network analysis (e.g., Hoppe et al., 2016), their content can be qualitatively coded (e.g., Weinberger & Fischer, 2006) or methods can be combined (e.g., Wise & Cui, 2018). Typically, these are related to the final grade, a post-test or course persistence to assess their impact on learning (Joksimović et al., 2018). However, without such explicit learning outcomes or standardised assessments, it is relevant to explore ways of identifying learning through observable changes in behaviour. To this end, the cognitive engagement of individuals is of particular interest, as it can reveal insights into their cognitive processes and offers a structured way to associate knowledge-change processes with overt behaviours such as written forum messages (Joksimović et al., 2018; Wang et al., 2016). The ICAP framework proposed by Chi & Wylie (2014) distinguishes four hierarchical levels of cognitive engagement, namely Passive (P), Active (A), Constructive (C) and Interactive (I), and allows us to classify learning activities accordingly. The framework is particularly well suited to study learning processes in discussion forums and was frequently used in formal contexts to examine learner behaviour and outcomes, e.g., through human coding (Raković et al., 2020; Sharma et al., 2024; Wang et al., 2016).

Applications of the ICAP framework to informal contexts are scarce, despite similar conditions and its potential for measuring distinct levels of cognitive engagement. For example, aggregating data over time (on a forum- or individual-level) can reveal shifts that potentially reflect learning-related processes as predicted by the ICAP hypothesis. However, coding schemes for ICAP, such as the one developed by Wang et al. (2016), involve high-inference decisions between categorical levels that may sometimes be

inherently ambiguous, especially without standardised learning tasks and instructions. Unlike in formal settings, the boundaries of “on-task” versus “off-task” discourse are less defined, and the criteria for the distinct ICAP levels require careful considerations. Therefore, the first aim of this study is to apply and systematically evaluate the ICAP coding scheme by Wang et al. (2016) for identifying cognitive engagement in these less-structured, informal environments.

Another challenge is the size of these environments, as informal communities often comprise thousands of users, making manual coding hardly feasible. While automated classification using natural language processing (NLP) has already been employed to address this, supervised methods rely on large, human-coded datasets for training. At the same time, these approaches remain “black boxes” that usually lack rationales or explanations for their predictions. Recent advances in Large Language Models (LLMs) offer promising new possibilities (Bosch et al., 2024; Pavlovic & Poesio, 2024): Trained on extensive text, they can generalise to new tasks without specific training and generate surprisingly adequate responses and explanations through natural language prompts (i.e., zero-shot prompting). Yet, their utility compared to established supervised methods in this context remains unclear. Unlike supervised methods, LLMs can also generate explanations or articulate their reasoning, and multiple non-deterministic model runs can be used to assess uncertainty. These outputs provide a unique window into judgments that require high inference decisions and could essentially serve as a scalable analogue to think-aloud protocols, as their introspective capability could help determine whether a coding scheme invites divergent interpretations and identify where these occur, having the potential to improve the coding process itself.

The second aim of this study is therefore to explore how LLMs can enrich this coding process by making the reasoning underlying such high-inference decisions visible and diagnosable at scale. To this end, we conceptualise LLMs as *cognitive proxies*, whose outputs could serve a functional, diagnostic role that is analogous to think-aloud protocols. Specifically, we assess how well LLM judgments align with human raters through established agreement metrics, examine how variation across multiple model runs can signal ambiguous cases, and analyse the accompanying explanations and reasoning traces to understand how the models navigate the coding scheme. Finally, we apply our LLM-based approach of identifying cognitive engagement in informal settings to large-scale data of forum discussions to assess what types of cognitive engagement emerge and how they vary across platforms. We thus target the following research questions:

**RQ1:** Can we apply the ICAP coding scheme by Wang et al. (2016) to identify different levels of cognitive engagement within open discussion forums of informal learning communities?

**RQ2:** To what extent do LLM judgments, guided by prompts derived from the coding scheme, align with those of human raters, and what do the accompanying reasoning traces and variations reveal about the nature of high-inference coding decisions?

**RQ3:** In how far does large-scale analysis of cognitive engagement in forum discussions allow us to characterise differences in expressed engagement behaviour across platforms?

## **Background and related work**

### **Learning through discussions in formal and informal settings**

Social constructivism emphasises that learning is fundamentally a social process, where individuals learn through interactions with others (Vygotsky, 1978). In online discussion forums, this means that the asynchronous interactions of users serve as a “vehicle for co-creation of knowledge” (Gunawardena et al., 1997, p. 412). This makes them particularly suitable for the analysis of learning processes. Studies in the field of Computer-Supported Collaborative Learning (CSCL) have typically studied these interactions in formal educational contexts such as blended learning courses or MOOCs, in which participation can be explicitly linked to predefined learning outcomes like course completion, grades, or assessments (Joksimović et al., 2018). For example, Tseng et al. (2016) found forum participants in MOOCs significantly more likely to complete courses successfully, while Houston et al. (2017) reported a direct association between forum interaction frequency and final course grades.

In contrast to that, informal learning settings such as CS projects and online knowledge-exchange communities like Reddit lack pre-defined learning goals and options for explicit assessments, which makes it challenging to apply traditional outcome-based evaluation methods. Still, discussions in these contexts also appear to facilitate learning, although related research remains sparse (Aristeidou & Herodotou, 2020; Hudgins et al., 2020). For example, Kloetzer et al. (2013) found that volunteers in online CS projects gained a range of learning outcomes (e.g., on-topic knowledge, scientific literacy), specifically through incidental learning processes triggered by peer interactions in the forum. Similarly, Rohden et al. (2019) highlighted the integral role of forum discussions for collaborative knowledge construction in CS, although without explicitly measuring associated learning outcomes. Online communities such as Reddit have also been recognised as environments fostering informal learning, as highly engaged users participate actively through questioning, content creation, and community-building activities (Hudgins et al., 2020).

However, quantitative approaches that study forum interactions alone often fail to capture the nuanced processes involved in such knowledge co-construction, rendering its relationship with learning somewhat opaque. Consequently, a qualitative analysis of forum

messages is needed as it provides essential insights into the collaborative nature of learning, particularly for informal settings where explicit learning outcomes are absent. Early models by Gunawardena et al. (1997) provide foundational tools for manually identifying phases of knowledge co-creation in forums, from initial idea-sharing to the synthesis of new meaning, and allow for tracing the evolution of a discussion and determining if higher-order cognitive processes have occurred (see Schellens & Valcke, 2005). Weinberger and Fischer (2006) introduced a multidimensional model, considering the participation, epistemic, argumentative, and social dimensions, effectively capturing different facets of online discourse. While originally developed for formal educational contexts, such frameworks remain valuable when adapted to informal settings. Kumar et al. (2018), for example, successfully used qualitative content analysis of discussions in Reddit communities (e.g., *r/askscience*), categorising exploratory learning dialogue into explanation, socialising, information exchange, and forum-specific interactions.

### **Cognitive engagement and ICAP**

In this work, we focus on the cognitive engagement displayed in forum discussions, as it allows us to associate overt behaviours such as written forum messages with underlying knowledge-change processes and learning (Galikyan et al., 2021; Wang et al., 2016). The ICAP framework by Chi and Wylie (2014) provides a structured way to analyse varying depths of cognitive engagement. In its original conceptualisation, ICAP assumes that passive learning activities such as watching a lecture lead to less learning than, for example, actively taking notes. The framework distinguishes four levels of learning, passive (P), active (A), constructive (C) and interactive (I). The ensuing hierarchy ( $I > C > A > P$ ) further postulates that as learners engage more deeply with the content to be learned (i.e., engage in higher levels), their learning increases. The greatest level of learning is then achieved as they dialogue and interact with other students to collaboratively construct knowledge (i.e., interactive). Following this notion, cognitive engagement can also be assessed by examining the discussion behaviour of forum users itself, i.e., their written messages. Galikyan et al. (2021) examined this for discussion forums in MOOCs, showing that students who demonstrated higher levels of cognitive engagement achieve better course performance, while those with lower levels of engagement tended to perform worse. Coding schemes for ICAP were also used in other studies that examined student behaviour and roles in discussions (Liu et al., 2022; Raković et al., 2020; Sharma et al., 2024; Wang et al., 2016).

This study builds on the work by Wang et al. (2016), who used the ICAP framework to develop a coding manual to identify cognitive engagement in forum discussions of MOOCs. The authors used the distinct levels of cognitive engagement according to ICAP and omitted the *passive* category, as it is not applicable in an already active forum context, and

introduced *off-topic* as a category to discern course-related from unrelated discourse, which was shown to be a relevant distinction (cf. Weinberger & Fischer, 2006). They also further refined the *active* and *constructive* category for a more granular view on the level of cognitive engagement, demonstrating a high reliability for their coding manual. In their analysis, they found that students who displayed higher-order thinking behaviour (i.e., having contributed *interactive/constructive* posts) learned more by engaging more deeply with course materials than those who only showed general attention and engagement (i.e., having contributed *active* posts), which in turn learned more than those who only contributed *off-topic* posts, supporting the proposed rank order by ICAP (Chi & Wylie, 2014). In their study, learning was operationalised through a formal post-test score, which is not applicable in informal contexts. Still, their validation of the ICAP hypothesis further supports the use of cognitive engagement as a relevant indicator for learning processes in forum discussions, particularly for contexts where direct assessment of learning is not possible. However, its application to these contexts warrants a careful consideration of what constitutes task-related discourse and how to operationalise the respective ICAP levels.

### **Automated techniques for coding discussion data**

Qualitative analysis of discussions in large online communities is often paired with automated analysis techniques such as NLP, and many coding schemes themselves are usually crafted for subsequent automatic application (see, e.g., Kumar et al., 2018; Wang et al., 2016). A recent meta review by Ferreira Mello et al. (2024) showed that discussion data is among the most commonly used textual resources in the field of learning analytics, with many studies employing machine learning approaches like Random Forest classifiers or more advanced language models like BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019). Typically, human annotations are used to create a training data set that contains the human-coded labels as the target variable with input features extracted from specific dictionaries or tools such as CoH-Metrix or Linguistic Inquiry and Word Count (LIWC) (see Ferreira Mello et al., 2024). In a process called *supervised learning*, the classifier is then trained on a subset of this human-labelled data, and its performance is assessed against a held-out test subset, typically by comparing the predictions with the human labels through cross-validated metrics such as accuracy, precision/recall, F1-score or quadratic weighted Cohen's Kappa (QWK) to ensure the model's reliability. Typically, the exact rationale behind these predictions remains unclear, as it results from learned weights optimised through training algorithms that are aimed at minimising prediction error. In the case of ICAP, multiple automatic coding approaches exist in the literature, e.g., Liu et al. (2022) who automatically classified cognitive engagement based on the ICAP framework among emotional engagement using the BERT

model. To achieve a reasonable accuracy however, researchers need to acquire sufficient training data, which often means extensive human-coding efforts that need to be cross-validated and revised over several iterations, costing time and resources. The time and resources that are needed often scale with the complexity of the coding task, meaning that the application of high-inference coding schemes like the ICAP-based scheme by Wang et al. (2016) would require significant human coding effort. As a result, many resort to the adoption of microwork platforms that address these constraints by offering human labour as a scalable computational resource (Irani, 2015).

In contrast to this, LLMs have already been instruction-tuned with human feedback and do not necessarily need task-specific training, as they can generalise to unseen tasks by simply being instructed through detailed prompts and examples (*few-shot prompting*). Due to their generative nature, the models can also be instructed to provide explanations associated with generated solutions that can mitigate the “black-box” issue of conventional machine learning models. Recent models also externalise their reasoning process, which additionally allows for studying these answers from a cognitive psychology perspective (Besta et al., 2025; Binz & Schulz, 2023). Because of their ease-of-use, LLMs have been broadly adopted as coders across many disciplines, e.g., for the identification of political ideology in tweets or for the development of inductive qualitative codebooks (Barany et al., 2024; Heseltine & Clemm von Hohenberg, 2024). In the field of CSCL, Bosch et al. (2024) demonstrated that LLMs can code help-seeking behaviour in forum posts, with judgments that mirror those of human annotators.

Despite their appeal, recent work urges caution and emphasises the need for systematic evaluations of LLM applicability (Misiejuk et al., 2025). Critical questions remain regarding the practical trade-offs between large, general-purpose LLMs, which can be computationally expensive, and smaller, fine-tuned encoder-based language models or conventional machine learning approaches. As such, work in this area still needs to explore how reliable LLM judgments for a qualitative coding task are, and how their unique capabilities like generating explanations and multiple judgments can help mitigate prevalent issues and offer new possibilities to optimise the classification workflow.

## Method

To tackle our research questions, we apply and systematically evaluate a coding scheme for cognitive engagement based on ICAP by applying it to forum discussions of three informal learning communities, assessing the agreement between two independent human raters (**RQ1**). We then develop prompts for LLMs with detailed instructions that incorporate this coding scheme, assessing the agreement between three state-of-the-art LLMs and the two human raters. Specifically, we study both human-human and human-model agreement, making use of the advantages of LLMs like uncertainty estimation,

generation of reasoning traces, and explanations (**RQ2**). Finally, we apply our approach to large-scale discussion data of the three communities, analysing how the ICAP categories based on our coding scheme are reflected in real-world data and whether differences exist (**RQ3**).

### **Coding scheme and prompt**

We adapted the coding scheme by Wang et al. (2016) that was originally developed for MOOCs to informal learning settings. Our scheme retains the original six ICAP categories, including the *Off-topic* category, and we define on-task discourse as posts that are related to the content of the specific citizen science project or subreddit. Distinguishing on- from off-topic posts is particularly challenging in informal settings, whose scope is broader than that of most MOOCs. However, we generally found that in our examined forums the distinction works quite well, as most off-task discourse relates to technical difficulties or is otherwise easily identifiable (e.g., personal questions). Accordingly, we also consider posts to be *Constructive (Idea)* if they introduce new content or ideas that go beyond the current discussion rather than beyond the scope of the project/subreddit, which is often difficult to define in informal settings. The coding scheme can be seen in Table 1.

**Table 1**

ICAP coding scheme adapted from Wang et al. (2016) and based on Chi &amp; Wylie (2014)

| ICAP label                              | Explanation                                                                                                                                                                                  | Example                                                                                                                                                                                            |
|-----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Off-topic                               | Comment lacks thematic relevance and is irrelevant to the scientific content or discussion of the project/subreddit                                                                          | <i>Hello from Arkansas!</i>                                                                                                                                                                        |
| Active<br>(Attention/Engagement)        | General engagement with the content of the project/subreddit, without specific paraphrasing or repetition                                                                                    | <i>I am starting a new search for planets now, wish me luck...</i>                                                                                                                                 |
| Active<br>(Paraphrase/Repeat)           | Engagement with project/subreddit content through paraphrasing, repeating or indicating active interaction with the content                                                                  | <i>Looks like a great great candidate, needs a closer look. Beautiful!</i>                                                                                                                         |
| Constructive (Idea)                     | Introducing a new idea or content that goes beyond the current discussion (e.g., referencing external resources) or using technical terms or hashtags when the comment is the thread starter | <i>Likely RR Lyrae variable star, and the inconsistencies could be due to the Blazhko effect.</i>                                                                                                  |
| Constructive<br>(Elaboration/Reasoning) | Elaborating on a point, explaining a phenomenon, or establishing a cause/effect relationship by comparing two conditions                                                                     | <i>#goodtransit. The #transiting is visible in all data segments, the dip is U-shaped with a flat bottom, and the star's metrics confirm it's not an eclipsing binary.</i>                         |
| Interactive                             | The comment is embedded in an interaction thread, directed at a conversational partner and is constructive, involving elaboration, building on ideas, or challenging the partners ideas      | <i>While you're right that many artifacts are not periodic, some, like momentum dumps, can be. For this star, the signal's period of 2.7 days is very close to a known instrumental frequency.</i> |

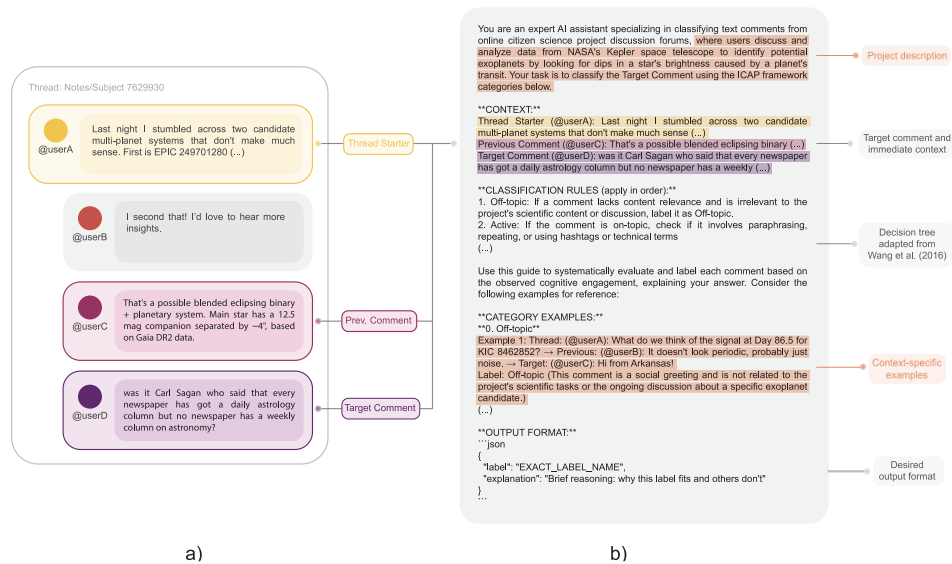
Based on the coding scheme, we developed a prompt template through iterative refinement (see Supplementary Material 2 for the full template, development process, and prompt sensitivity analysis). The final prompt uses few-shot in-context prompting, providing the model with category definitions, two labelled examples per ICAP level and a sequential decision tree. Key design decisions across iterations included the introduction of the decision tree as well as the provision of context-specific examples. In the prompt, the model is instructed to qualitatively code a comment regarding its cognitive engagement and provide an explanation for its choice. In addition to the target comment, we also included its immediate thread context, similar to Wang et al. (2016), which is the thread starter and previous comment. This allows for a more appropriate assessment of the discussion and

can be seen in Fig. 1. Each prompt further includes the specific project/subreddit context with a short description, as well as context-specific examples. This is followed by a decision tree for the coding scheme adapted from Wang et al. (2016), which we ask the model to systematically follow. The post-hoc sensitivity analysis reported in Supplementary Material 2 confirmed that the classification performance is robust to variations in element ordering and that both rules and examples are necessary for achieving the reported agreement levels.

On average, the final prompts (including the target comment with context) contained 1453.28 ( $SD = 198.21$ ) words or 2207.20 ( $SD = 323.32$ ) tokens, respectively. The full prompts can be found in Supplementary Material 2 and on GitHub<sup>1</sup>.

**Figure 1**

a) discussion thread with target comment and context that goes into the prompt. b) the final prompt including the ICAP coding scheme and project/subreddit-specific information and examples.



## Sample and agreement metrics

We used forum posts from the "Talk" page of the *Exoplanet Explorers* and *Galaxy Zoo* citizen science projects hosted on Zooniverse<sup>2</sup> and from *r/askastronomy* on Reddit. In both cases, users discuss and solve astronomy-related questions and problems. The comments were obtained through an API. Deleted or removed posts were excluded. No further filtering for bots, spam, or low-content posts was applied, as such posts are captured by the *Off-topic* category of our coding scheme. For each post, we also have information on its

<sup>1</sup> [github.com/krukowski-lab/RPTel-2026-CogEng-LLMs](https://github.com/krukowski-lab/RPTel-2026-CogEng-LLMs)

<sup>2</sup> [zooniverse.org](https://zooniverse.org)

author, the original thread it was posted to, and whether it is a reply. As context, we consider the previous/parent comment and the thread starter, if applicable. For Reddit, we considered comments until a depth of two sub-comments to the original submission. We followed the approach by Wang et al. (2016) for the provision of the previous comment and thread starter as context, as we are primarily interested in the cognitive engagement shown in the respective forum posts, and considering the whole discussion thread would be (computationally) infeasible.

For the human coding, we instructed two human raters with a comprehensive coding manual (see Supplementary Material 1) to code the dataset. For each target post, raters were provided with the thread starter and previous/parent comment as context. After an initial small sample where the raters could ask questions in case of ambiguities, they independently coded the dataset. In total, we sampled a random subset of 50,000 comments per forum. To obtain an equal distribution of ICAP categories, we applied purposive quota sampling, where Rater 1 was provided with a random subset of the 50,000 comments per forum and selected a total sample of 346 comments, yielding a roughly balanced distribution across categories and forums, which was then given to Rater 2 for coding. This balanced sampling was a deliberate design choice to ensure adequate representation of all ICAP categories for instrument evaluation, as low-prevalence categories would otherwise preclude meaningful per-category assessment. The final sample consists of 112 comments for *Exoplanet Explorers*, 114 for *Galaxy Zoo* and 112 for *r/askastronomy*. After removing 8 comments due to missing labels, the final sample consists of 338 comments, with 28.86 ( $SD = 65.60$ ) words on average. The human ratings were made before any automatic labelling and are not part of the experiment.

To assess agreement among the independent human ratings, we used quadratic weighted Cohen's Kappa (QWK), see Fleiss & Cohen (1973). The weighted measure is particularly well-suited to assess agreement regarding an ordinal scale, which aligns well with the hierarchical nature of the ICAP categories (Chi & Wylie, 2014; Wang et al., 2016) and emphasises the severity of misclassifications by assigning greater penalties to larger disagreements. For example, a disagreement between *Off-topic* and *Interactive* is weighed more negatively than one between *Off-topic* and *Active (Attention/Engagement)*. For the calculation, we treated all categories of our scale as distinct levels, including the two *Active* and *Constructive* categories, respectively. Additionally, we report ordinal Krippendorff's Alpha as a complementary metric.

We further obtained a sample for the large-scale analysis of forum posts, stemming from a full calendar year of forum discussions. After excluding deleted or removed posts ( $N = 1,156$ ), this sample consists of 36,697 comments, with 13,394 from *Galaxy Zoo* (year = 2021), 11,903 from *Exoplanet Explorers* (year = 2018) and 11,400 from *r/askastronomy* (year = 2021).

## Models and experiment

After initial tests with smaller models (Llama3.1-8B, Gemma3-4B) that led to unsatisfactory performance (e.g., QWK = .49 [.39, .58] for Llama3.1-8B and QWK = .20 [.09, .32] for Gemma3-4B with Rater 1), we selected three larger state-of-the-art models, two of which are capable of externalising their reasoning process (see Besta et al., 2025). We thus chose the *Qwen/QwQ-32B*, *deepseek-ai/DeepSeek-R1-Distill-Llama-70B*, and *meta-llama/Llama-3.3-70B-Instruct* models. For performance reasons, we used an API provider<sup>3</sup> to run the models. All models were accessed via this API between June and August 2025 using Python and the OpenAI library with automatic retry logic (up to 5 attempts with exponential backoff) for rate-limit and connection errors. Full retrieval scripts and configuration details are provided in our GitHub repository. Our model choice is certainly not exhaustive, and this work is not intended as a comprehensive comparison between the models, but rather as an applicability test for the use of LLMs for our task, for which we also want to provide a comparison between different models.

To provide a proof-of-concept supervised baseline, we fine-tuned BERT (*bert-base-uncased*), a widely used transformer-based classification model, on the human-coded dataset, training one model per rater ( $N = 338$  examples each), as labels were not adjudicated between raters. We used 5-fold stratified cross-validation with inverse-frequency class weighting and early stopping. To mirror the context provided to the LLMs, each input comprised the target comment, its thread context (thread starter and previous comment), and a short project description, concatenated with special delimiter tokens. The full training configuration and hyperparameters are available in the linked GitHub repository.

For the experiment, we generated responses with each model using Python and the default parameters proposed by their developers on Huggingface<sup>4</sup> (e.g., *temperature* = 0.6, *top\_p* = 0.95 for QwQ-32B). The *temperature* parameter controls the level of randomness of the output (i.e., lower values produce more deterministic outputs, while higher values promote more creative outputs), *top\_p* limits the model's word choices to the most likely options whose combined probability adds up to  $p$ , helping balance diversity and coherence in the output. We made sure that model answers allow for the extraction of the label and explanation by instructing it to generate both in JSON format and additionally enforcing a pydantic model validator. We generated five judgments per comment and used a majority vote paradigm to decide on the final label, as aggregating multiple stochastic outputs reduces the influence of individual sampling variability on the final classification. Five judgments were chosen as they are an odd number ensuring a clear majority while

---

<sup>3</sup> deepinfra.com

<sup>4</sup> huggingface.co

balancing reliability against computational cost. If no majority was achieved, we generated more answers until a majority was achieved. This occurred for 13 comments, with a maximum of 2 additional tries (1.31 times on average,  $SD = 0.48$ ). For the human-coded dataset ( $N = 338$ ), a total of 13,558,851 tokens were processed via the DeepInfra API, with QwQ-32B consuming the most tokens (5,197,239), followed by DeepSeek-R1-Distill-Llama-70B (4,427,666) and Llama-3.3-70B-Instruct (3,933,946), reflecting the longer reasoning traces produced by the two reasoning-capable models.

The majority-vote approach helps to ensure relevant answers and additionally allows for the calculation of a single comment-based uncertainty measure. Malzahn et al. (2023) compared several categorical disagreement measures and showed that pairwise-disagreement-based approaches are largely equivalent to established measures such as the Dispersion Index and Fleiss' Kappa, differing only in normalisation. Following this pairwise logic, we mapped each ICAP category to numeric values (e.g., 0 for *Off-topic*, 1 for *Active (Attention/Engagement)* etc.) and calculated the average of squared differences between all pairs of the five judgments. Unlike purely nominal measures, this captures the magnitude of disagreement between categories, which is meaningful given the ordinal structure of the ICAP framework. After normalising by the maximum possible squared disagreement, this value ranges from 0 (no uncertainty) to 1 (total uncertainty). We refer to this measure as model uncertainty, as it depicts the extent of variation in the individual model judgments. Finally, we saved the majority label and explanation for each of the five individual judgments.

## Results

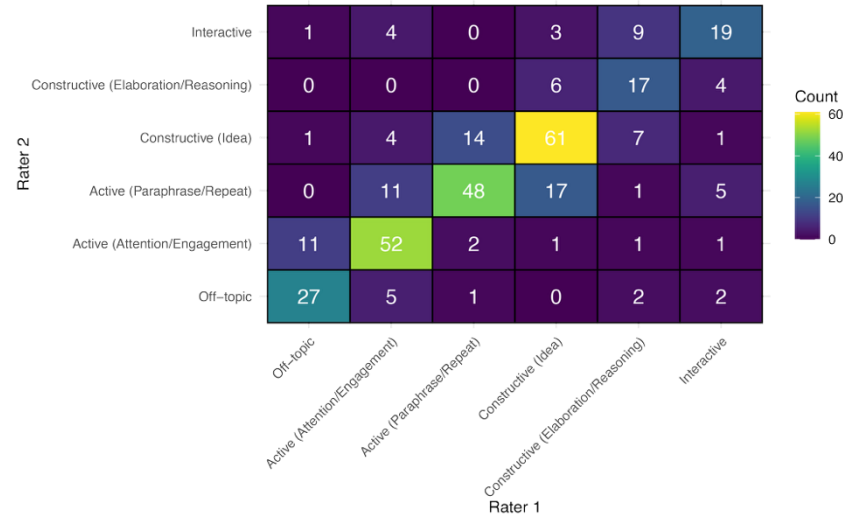
### Applicability of coding scheme

To approach **RQ1**, we assessed how the two human raters agreed with each other on the coding task. For our human-rated sample, QWK between the two human raters was .73, with 95% CI [.65, .82], indicating a “substantial” agreement beyond chance (Landis & Koch, 1977). All reported Kappa values were statistically significant with  $p < .001$ . Krippendorff's  $\alpha$  of .75 [.67, .83] further supports this finding. In Fig. 2, we can see the confusion matrix between the two human raters. Although we can observe some disagreements, they primarily affect adjacent categories, constituting 75% of all disagreements, which explains the high QWK despite some variation in labelling. In particular, the most prevalent type of disagreement affected the *Active (Paraphrase/Repeat)* and the *Constructive (Idea)* categories, which can likely be attributed to the high-inference nature of the coding scheme, e.g., difficulties regarding what constitutes discourse that goes beyond the current discussion, as the raters were only provided with the immediate context of the target comment, similar to Wang et al. (2016).

Apart from these disagreements, we can generally observe that the raters agree on the given categories, and that deviations regarding categories further apart on the ICAP scale rarely occurred, as indicated by the high QWK value. This is particularly relevant considering the ordinal nature of the coding scheme, where we want to avoid larger deviations (e.g., rating *Interactive* discourse as *Off-topic*). Considering this, the scheme appears to be applicable in our context (**RQ1**).

**Figure 2**

Confusion matrix between our human raters on the adapted ICAP coding scheme



### LLMs as raters

To address **RQ2**, we systematically evaluated the three selected LLMs against the human-coded data to assess human-model agreement and investigate the specific strengths and weaknesses of using LLMs for this high-inference coding task, comparing them with the BERT baseline.

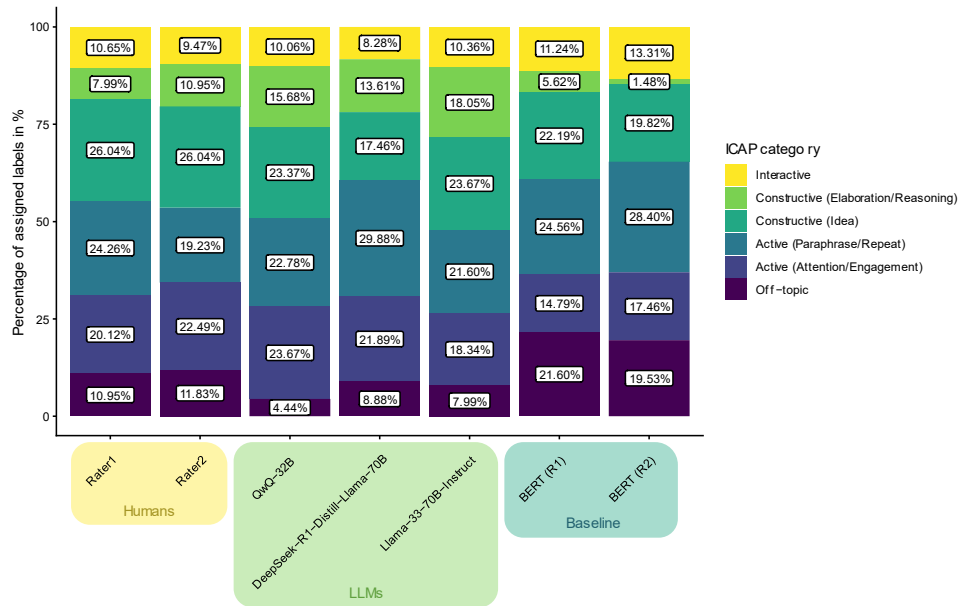
### Quantifying agreement

We started by descriptively comparing the distribution of ICAP categories across our raters (see Fig. 3). Visually, it appears to be comparable across all raters, particularly between the human raters, with the biggest deviation regarding the *Active (Paraphrase/Repeat)* category. We also observed larger differences between the human and LLM label distributions, with a general tendency towards more comments being rated as *Constructive (Elaboration/Reasoning)* and less being rated as *Off-topic*. The largest differences were observed between the BERT baseline and the other raters, particularly for the *Off-topic* and *Interactive* categories. A Chi-square test of independence showed that the distribution of labels significantly differed between all raters, with  $\chi^2(30) = 166.08, p < .001$ , Cramér's  $V$

= .11, although practically negligible. Next, we measured the actual agreement between raters using the described agreement measures QWK and Krippendorff's  $\alpha$ . This can be seen in Table 2.

**Figure 3**

Relative distribution of labels among human and model raters



As expected, the human-model agreement was generally lower than human-human agreement, with most values  $> .60$ , which, however, can still be classified as “substantial”. The baseline comparison with BERT showed less agreement, ranging from .28 to .36. This is likely attributable in part to the limited size of the training set which constrains the capacity of supervised models to learn reliable decision boundaries, particularly for a six-category classification task. As such, the BERT baseline achieved only a fair level of agreement under these data-constrained conditions. Further, the confidence intervals of the agreement values for the LLMs overlapped with those of the human-human agreement, suggesting that while the value itself is lower, we cannot conclude a significant difference between them, meaning that human-LLM agreement approaches human-human agreement. Due to insufficient performance, we did not use the BERT classifier for further analyses.

**Table 2**Inter-rater agreement (QWK and Krippendorff's  $\alpha$ ) between the human raters and models

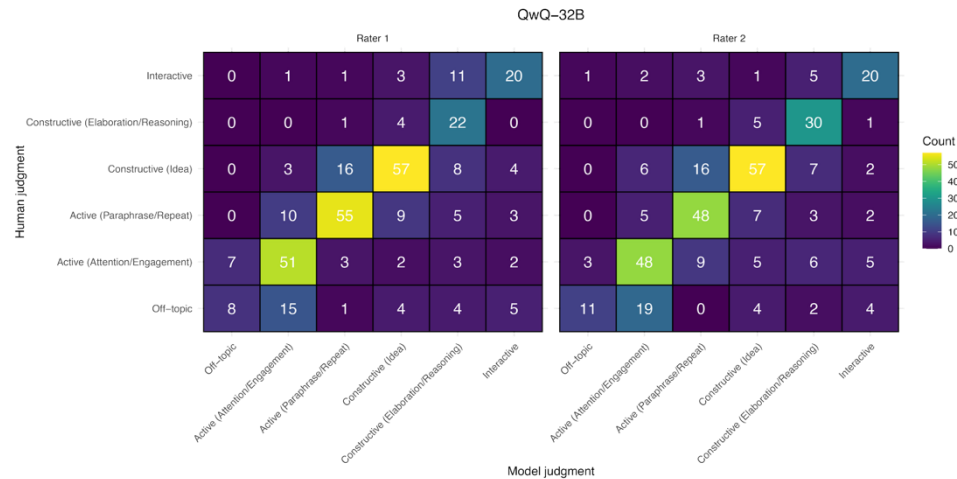
| Rater   |          | QwQ-32B       | DeepSeek-R1-Distill | Llama3.3      | BERT (baseline) |
|---------|----------|---------------|---------------------|---------------|-----------------|
| Measure |          |               |                     |               |                 |
| Rater 1 | QWK      | .64 [.54,.73] | .61 [.51,.71]       | .60 [.50,.69] | .33 [.20,.45]   |
|         | $\alpha$ | .66 [.56,.75] | .62 [.52,.72]       | .60 [.50,.70] | .36 [.22,.48]   |
| Rater 2 | QWK      | .60 [.50,.71] | .61 [.51,.70]       | .57 [.46,.68] | .28 [.16,.39]   |
|         | $\alpha$ | .62 [.51,.71] | .61 [.51,.70]       | .57 [.46,.67] | .31 [.18,.44]   |

Note: The BERT values reflect matched comparisons, each model being trained on and evaluated against the respective rater's labels

We further assessed the agreement by examining the confusion matrix for LLM and human ratings, which can be seen in Fig. 4 for the QwQ-32B model. The diagonal values show the counts of agreements for each category, all other values show the counts of disagreements and their respective direction (e.g., in 11 cases, Rater 1 assigned *Interactive*, while the model assigned *Constructive (Elaboration/Reasoning)*).

**Figure 4**

Confusion matrix of ICAP category classifications for QwQ-32B and the human raters



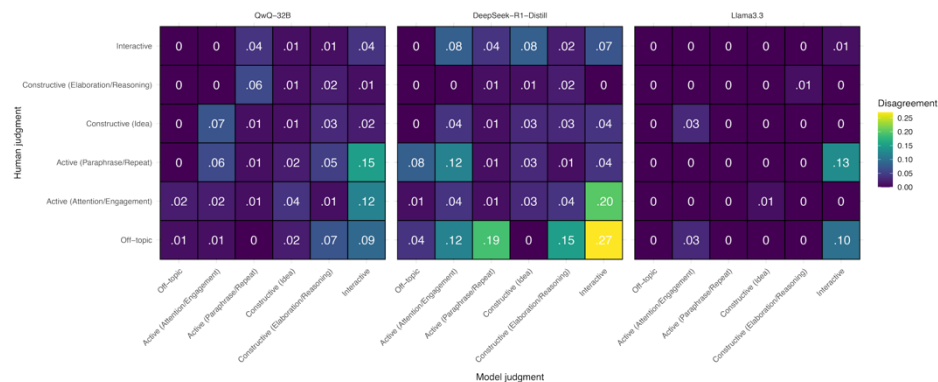
The matrix confirms the earlier finding regarding the inter-human agreement, showing that most human-model disagreements also reflect single-level deviations, explaining the higher QWK value despite certain disagreements. Particularly striking are the disagreements regarding *Off-topic* and *Active (Attention/Engagement)*, where the LLM assigned the latter category, while the human raters assigned *Off-topic*. Similarly, there are many cases where the humans rated *Constructive (Idea)*, while the model assigned *Active (Paraphrase/Repeat)*. This indicates that while the deviations occurred regarding adjacent categories, they were not within the same broader ICAP category in the original conceptualisation (e.g., *Active*), but rather between adjacent but distinct categories, mirroring human-human disagreements. Overall, we see less disagreements towards the

lower end of the ICAP spectrum, e.g., cases when the human raters assigned *Interactive*, while the LLM assigned *Off-topic*, indicating less tendency towards “false positives” of cognitive engagement.

As described above, we used the individual model judgments to calculate a per comment measure for model uncertainty of the LLMs. Across all models, this uncertainty measure was generally small ( $M = .02$ ,  $SD = .06$ ), meaning that the models showed little variation in their individual judgments. It slightly differed between the models, as determined by a Friedman test with  $\chi^2 = 147.67$ ,  $p < .001$ ,  $W = .22$ : For DeepSeek-R1-Distill, it was highest on average with  $.04$  ( $SD = .08$ ), for QwQ-32B it is  $.02$  ( $SD = .05$ ) and for Llama3.3, there was almost no model uncertainty on average, with  $<.01$  ( $SD = .04$ ). Pairwise comparisons using Bonferroni correction showed that the measure significantly differed between each pair of models. To study how the measure relates to each type of disagreement, we assessed its value in a confusion matrix for each model averaged across both human raters, as can be seen in Fig. 5. For each cell (i.e., type of disagreement), we thus calculated model uncertainty and constructed the matrix accordingly. As visible in the

**Figure 5**

Confusion matrices for each model showing the model uncertainty for each type of deviation



diagonal of the matrix, variations in judgments (i.e., high model uncertainty) rarely occurred for cases of human-model agreement and rather affected disagreements that are farther apart, particularly for cases when the humans assigned lower ICAP categories such as *Off-topic* or *Active (Attention/Engagement)*, while the model went for higher ICAP categories such as *Interactive*. Thus, such disagreements likely stem from ambivalent comments that can be interpreted in multiple ways, triggering differences in model judgments that manifest in higher model uncertainty values.

We further assessed this by calculating the external disagreement with the human raters for each comment, based on the same notion like above by using the squared difference between model and human judgments (e.g., if Rater 1 assigned *Off-topic* and the model

assigned *Active (Attention/Engagement)* for a comment, the external disagreement is  $(0 - 1)^2 / 25 = .04$ ). Using the average external disagreement, we can set it in relation to model uncertainty to assess how it relates to disagreements with the humans. For this, we calculated Spearman's correlations, finding a moderate to small positive relationship across all models with  $\rho = .22$  ( $p < .001$ ) for QwQ-32B,  $\rho = .23$  ( $p < .001$ ) for DeepSeek-R1-Distill and  $\rho = .12$  ( $p < .001$ ) for Llama3.3. This finding supports the earlier notion that when the individual judgments of the models vary to a greater extent (i.e., higher model uncertainty), there is also a greater chance of disagreement with the human raters.

To this end, the behaviour of model uncertainty when the humans themselves disagree is of particular interest, as it could indicate an inherent ambiguity regarding the respective ICAP category of a comment. To assess this, we examined the model uncertainty for comments on which the human raters agreed with each other vs. those where they disagreed. The results only showed significant differences for the QwQ-32B model, with a model uncertainty of .017 ( $SD = .046$ ) for cases of human-human agreement and .031 ( $SD = .056$ ) for cases of human-human disagreement, as measured by a Mann-Whitney-U test ( $U = 14678.5, p = .008, r = 14$ ). As such, if the humans themselves disagree regarding a comment, the QwQ-32B model also tended to do so. A similar picture emerges when calculating Spearman's correlation between model uncertainty and human-human disagreement (also measured through a comment-based disagreement like above), which also showed a significant correlation for the QwQ-32B model with  $\rho = .15, p = .005$ .

All of the above analyses offer relevant insights into the different dynamics of human-human and human-model agreement, particularly with respect to how ambiguous comments appear to evoke uncertainty in the LLMs which often results in disagreement with the human raters, a pattern of variability in responses comparable to what we would expect from human raters (**RQ2**). We further studied certain meaningful examples that illustrate this behaviour on a case-based level, focussing on the QwQ-32B model, which we will now describe.

### ***Case-based qualitative analysis***

A simple example for a comment on which both human raters and the LLM agreed on an ICAP category is the comment "*what?*", for which all agreed on *Active (Attention/Engagement)*. The comment did not evoke any model uncertainty, i.e., all five judgments were the same. The LLM gives the following explanation: *The comment 'What?' is a brief, direct reaction seeking clarification on the parent comment's claim about the 'blue moon.' It shows attention to and engagement with the discussion's content without paraphrasing, repeating, or introducing new ideas. It does not qualify as Constructive or Interactive because it lacks elaboration, reasoning, or explicit references to others'*

*arguments beyond a simple reply.* In this example, the final label and given explanation align well with the provided coding scheme.

For the comment *“This has a slightly #irregular shape. I think the blue object (at 7 o'clock) is not physically associated.”*, the human coders disagreed between *Constructive (Idea)* and *Constructive (Elaboration/Reasoning)*. Here, the model assigned *Constructive (Elaboration/Reasoning)*, citing: *“The comment is on-topic and provides reasoning to explain why the blue object at 7 o'clock may not be physically associated with the galaxy. The user introduces analysis (e.g., 'slightly #irregular shape') and causal reasoning ('not physically associated') to elaborate on the observation, fulfilling the criteria for Constructive (Elaboration/Reasoning) (...).”* In this example, the explanation by the LLM reflects the coding scheme well. Another case in which the model agreed with only one of the human raters is the comment: *“Agree, #spiral faint outer arms more obvious in DecALS”*. The comment contains implicit references to another user, yet does not explicitly introduce or elaborate on a new idea that goes beyond the current discussion (i.e., thread starter: *“#spiral? Very faint arms”*). In this sense, it is slightly ambiguous. As such, Rater 1 assigned *Active (Paraphrase/Repeat)*, while Rater 2 assigned *Interactive*. The model's final choice was also *Interactive*, based on three judgments for *Interactive* and two for *Active (Paraphrase/Repeat)*, resulting in an uncertainty of .22. In this example, the different judgments of the model reflect the human ratings and inherent ambiguity quite well. Interestingly, the explanation matches with the given label, so for the *Interactive*, the model cites: *“The comment builds on the thread starter's classification by agreeing and elaborating that the faint outer arms are more visible in the DecALS image processing. The use of 'Agree' explicitly references the prior user's idea, while the mention of DecALS provides reasoning, making it both constructive and interactive.”* For the *Active (Paraphrase/Repeat)* judgment, however, the model cites: *“The comment paraphrases and builds on the thread starter's observation by affirming the #spiral classification and adding specificity (e.g., 'faint outer arms more obvious in DecALS'). It uses technical terminology (DecALS) and a hashtag, aligning with the Active (Paraphrase/Repeat) category. While it adds details, it does not introduce new ideas or causal reasoning beyond elaborating on the existing classification.”* The two explanations help to understand the “rationale” behind the respective LLM choice and exemplify well how each explanation itself is consistent with the coding scheme yet weighs the reference to the other user differently.

There are, of course, examples in which the LLM diverged from human consensus: *“Well i have pictures ive taken. If its clear tonight ill shoot it when its above horizon and when sets which has been consistent all summer... Which makes no sense which is why im asking”*, a comment made in *r/askastronomy* for example. Both raters agreed on *Active (Attention/Engagement)*, as it shows general engagement with the content of the subreddit. Although it contains an idea (*“makes no sense”*) it does not elaborate on it and rather shows

mere engagement with the content (“*which is why I’m asking*”). It also contains implicit references to another user but does not build or elaborate on them. The model uncertainty for this comment is .24, and the different judgments contain 3x *Interactive*, 1x *Constructive (Elaboration/Reasoning)* and 1x *Active (Attention/Engagement)*, all with self-consistent explanations. Particularly helpful is the output of the “reasoning process” which the model provides, e.g., for the *Interactive* label: *[...] Looking at Active (Attention/Engagement), this might fit because they're engaging by indicating they'll take action (shooting photos) and seeking to resolve their confusion. But the user is also explaining their reasoning: the consistency doesn't make sense, hence the question. That reasoning part might push it into Constructive (Elaboration/Reasoning). However, the user is replying to a previous comment, so maybe it's Interactive? Wait, the Interactive category requires the comment to be both Constructive and reference another user. The user is replying to another user, so there's a reference. The comment explains their plan to take more photos (action) and the inconsistency they observe, which is elaboration. So they're building on the previous discussion by adding their own observations and intent to get more data. So, since they're referencing another user and adding their own reasoning (the inconsistency makes no sense, hence asking), this could be Interactive. [...]*”

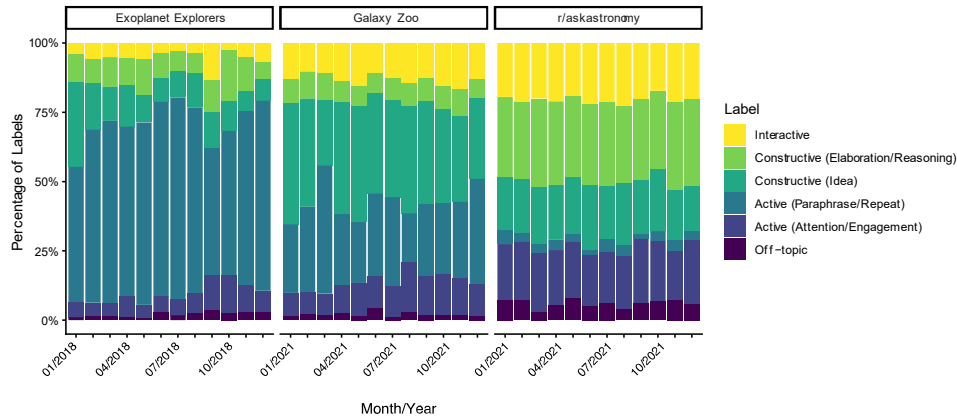
This reasoning process exemplifies how the LLM followed the coding scheme yet falsely interpreted the comment to contain reasoning/elaboration and building on another one’s idea, which led it to judge it as *Interactive*. It also shows that the human consensus label was considered and even became one of the final judgments, thereby underlining the ambiguity of certain comments in the face of a high-inference coding manual. It also shows how the externalised reasoning traces can help identify such issues, fulfilling a diagnostic role analogous to think-aloud protocols by making visible how the coding scheme is navigated and where it invites divergent interpretations.

### ***Application to large-scale data***

For **RQ3**, we demonstrate the scalability of our approach by using the best-performing model from our tests (QwQ-32B) to code all forum messages across one calendar year for all our three selected forums (see Methods section). Such a large-scale study can act as a first indicator for differences in cognitive engagement across forums. Initial tests reveal

**Figure 6**

Large scale assessment of cognitive engagement over a calendar year for each forum based on our coding scheme and using the QwQ-32B classifier



that the forums differ in comment length: *r/askastronomy* comments were longest on average ( $M = 64.45$ ,  $SD = 87.28$  words), followed by *Galaxy Zoo* ( $M = 20.78$ ,  $SD = 34.29$ ) and *Exoplanet Explorers* ( $M = 12.57$ ,  $SD = 37.79$ ). Most comments were top-level posts rather than replies, particularly on Zooniverse (95.80% for Exoplanet Explorers, 81.77% for Galaxy Zoo), while *r/askastronomy* had a higher share of replies (30.71%), though nesting was capped at depth two (see Method section). We found that all ICAP categories from our schema are reflected by the model ratings, with *Active (Paraphrase/Repeat)* being the most prevalent category that represents 31.20% of all comments, while only 3.02% were labelled as *Off-topic*. In Fig. 6, we see the relative share of ICAP categories over time for each of the analysed project forums. As can be seen, the distributions significantly differ between the forums. For *r/askastronomy*, the categories are more evenly distributed, with almost no comments being labelled as *Active (Paraphrase/Repeat)*, while for the two Zooniverse projects, *Active (Paraphrase/Repeat)* and *Constructive (Idea)* are the most prevalent categories. A Chi-square test for independence confirms ( $\chi^2(10) = 11,622$ ,  $p < .001$ ,  $V = .40$ ) significant differences across the forums. These results exemplify different profiles of knowledge exchange, with likely more task-oriented discourse in the CS forums, as indicated by a higher share of comments labelled as *Active (Paraphrase/Repeat)* and more *Interactive* and *Constructive (Elaboration/Reasoning)* discourse indicating a focus on knowledge exchange in *r/askastronomy*. To rule out comment length as a confound, we stratified by median word count and repeated the test with Bonferroni correction. Significant differences persisted for both short ( $\chi^2(10) = 7,699$ ,  $p < .001$ ,  $V = .45$ ) and long comments ( $\chi^2(10) = 1,675$ ,  $p < .001$ ,  $V = .22$ ). As a next step, we plan to aggregate these data on a user level to assess longitudinal, individual learning trajectories by identifying shifts in cognitive engagement over time, manifesting through an increase of comments

with higher levels of cognitive engagement that can be used to approximate learning processes.

## Discussion

In our study, we aimed to approach the challenge of identifying cognitive engagement as a learning-related indicator in forum discussions through qualitative coding. By applying the ICAP-based coding scheme developed by Wang et al. (2016) to informal learning settings such as Zooniverse and Reddit, we explored how we can identify differences in cognitive engagement through the study of written text messages (**RQ1**). We further studied how LLMs can be used to approach this task by comparing the judgments of multiple state-of-the-art open-source LLMs with those of two independent human raters, critically exploring their capabilities and limitations (**RQ2**) and applying them on large-scale data (**RQ3**).

Firstly, our results show that we can indeed identify different levels of cognitive engagement in informal settings using our adapted high-inference coding scheme. Two independent raters showed a “substantial” agreement with each other, and while it was still below  $QWK = .80$ , if disagreements occurred, these were mostly single-level deviations opposed to categories that are farther apart on the ICAP scale and can likely be attributed to the high-inference nature of the coding scheme (**RQ1**).

Secondly (**RQ2**), our results show that elaborated, context-specific prompts incorporating the coding scheme can be used to instruct LLMs to produce judgments that align with those of human raters on a level that can be classified as “substantial” (Landis & Koch, 1977). The reasoning model QwQ-32B showed the highest agreement with both human raters, and overlapping confidence intervals suggest that human-model agreement approaches human-human agreement levels. Notably, the fine-tuned BERT baseline achieved only fair agreement, despite being trained directly on each rater’s labels. However, this comparison should be interpreted cautiously, as the small sample likely disadvantages supervised learning and the performance does not necessarily reflect what such models could achieve with larger datasets.

More importantly, however, these results demonstrate the way LLMs can potentially transform the coding process itself. Opposed to conventional classifiers who often lack explainability, the non-determinism of LLMs can be leveraged to measure variation in judgments, which signals uncertainty and was particularly evident for cases in which the model finally disagreed with human consensus, e.g., regarding *Interactive* vs. lower ICAP categories. For the QwQ-32B model, we found a small yet significant correlation between model uncertainty and disagreement between the two human raters, indicating that ambiguity in the comments regarding the ICAP category is reflected by both the human raters and the model.

Beyond measuring variation, the reasoning traces and explanations that LLMs generate allow us to study the inference chains evoked by our coding scheme (Hagendorff et al., 2024), which are similar to think-aloud protocols that we would expect to be given by humans during the coding process. While it is subject of recent discussions around LLMs whether these externalised inference chains indeed reflect any internal reasoning (Barez et al., 2025), in our case, they provide the ability to match outcomes (i.e., ICAP labels) with underlying rationales based on our provided coding scheme. Especially for high-inference coding schemes like ours, this capability proves helpful, as it allows us to quickly identify ambiguous cases, track faulty assumptions and identify misguided reasonings.

This is particularly exemplified by the qualitative analyses. While there are obvious cases of divergent classifications where both human raters agreed but the LLM disagreed, we also observed ambiguous cases where the human raters disagreed between two categories, and the LLM assigned a category that could either resolve this conflict (typically when the model is highly certain) or simply weighs certain aspects of the comment differently. The latter particularly manifests in multiple different model judgments that reflect this ambiguity and eventually produce a high model uncertainty. However, we also observed instances where the LLMs misapplied the coding scheme and where their judgment is not attributable to ambiguous comments that evoke a high uncertainty but rather to simple misinterpretations of the comment where the model stuck too literally to the coding scheme and did not extrapolate beyond that, indicating a lack for a “broader understanding” of the coding task and underlining the required caution for their unsupervised usage (Misiejuk et al., 2025). Importantly, for all these cases, the reasoning traces and explanations make the rationale behind these faulty judgments diagnosable.

In this sense, LLMs can be understood as *cognitive proxies*, whose externalised decision processes make visible which aspects of a comment and which decision points in the coding scheme led to a given judgment, as well as where and why a coding scheme invites divergent interpretations. While they do not replicate human cognition itself, LLMs can serve the diagnostic role that think-aloud protocols would fulfil during the evaluation of coding schemes. Specifically, this functional parallel is primarily informative when 1) the coding scheme is sufficiently well-defined and embedded in the prompt, 2) a human-coded reference sample exists for validation, and 3) the task involves high-inference decisions where multiple interpretations are genuinely plausible. Importantly however, unlike human think-alouds, which can be understood as imperfect but genuine windows into cognitive processes, LLM reasoning traces are generated outputs whose relationship to internal model computations remains debated, as mentioned above (Barez et al., 2025). As such, conceptualising LLMs as simulation models of human cognition should be treated with caution. Rather, they can act as tools whose outputs serve an equivalent diagnostic function,

surfacing ambiguities that would otherwise require extensive human protocol analysis to detect at a comparable scale.

Finally, our results also show that we can leverage the qualitative coding capabilities of LLMs by scaling to large data sets, which allows us to aggregate measures for cognitive engagement on a project or platform level (**RQ3**). Our analyses showed that all our categories are reflected in real-world data, and that we can observe significant differences between platforms and projects that reflect different types of knowledge exchange. While the present analysis is descriptive in nature and focused on demonstrating the viability of our approach at scale, a natural next step is to aggregate and analyse individual user trajectories over time, identifying shifts in cognitive engagement that can serve as learning-related indicators, consistent with the ICAP hypothesis which assumes that learning increases as learners engage more deeply with the content and move towards higher levels on the ICAP scale (Chi & Wylie, 2014). Such longitudinal modelling of individual learning trajectories in informal settings is the subject of ongoing work.

From a practical standpoint, our LLM-based coding approach offers notable advantages in terms of feasibility compared to conventional supervised pipelines such as BERT. These typically require substantial upfront investment in human coding to produce sufficient training data, which itself needs to be cross-validated and iteratively refined across multiple rounds before a classifier can be trained and deployed. In contrast, our approach requires only a small set of representative examples for prompt construction and can be applied without task-specific training, substantially reducing the time and effort needed to move from a coding scheme to large-scale application. While human coding remains necessary to validate the quality of LLM judgments (Misiejuk et al., 2025), this evaluation sample is considerably smaller than what supervised training demands. At the same time, LLM inference produces per-token costs that scale with corpus size and prompt complexity (our evaluation sample alone required, on average, four million tokens per model), which may become a relevant factor for very large datasets. These trade-offs need to be weighed on a case-by-case basis, additionally considering the diagnostic value that reasoning traces and uncertainty estimates provide, which conventional classifiers do not offer.

Our results come, however, with limiting factors. Although we found that our applied coding scheme proves to be applicable in informal settings with an agreement between the two human raters that can be interpreted as substantial, there is dispute regarding appropriate thresholds for what defines “good” agreement as measured through QWK, particularly depending on factors such as the prevalence of agreements and the number of categories (Doewes et al., 2023; Landis & Koch, 1977). As such, a higher value would have been desirable and could likely be achieved by increasing the selectivity of bordering ICAP categories or selecting a larger sample. Relatedly, as our evaluation sample was balanced by design, reported agreement metrics may differ under natural category

distributions, which future work could examine through prevalence-adjusted sensitivity analyses. In addition, while we carefully constructed the prompts that incorporate this coding scheme, they are the result of several iterations, and while we found that the order of prompt elements does not significantly change performance, the inclusion or exclusion of certain elements does have an influence. As such, it is conceivable that the prompts can be further improved by including additional instructions.

Further, although we compared different state-of-the-art open-source LLMs, this comparison is also by far not exhaustive, as there are many more models that could have been included. At the same time, the development of LLMs is very fast paced, meaning that results obtained with a specific model version may not be directly reproducible as models are updated, though improvements in general capability suggest that performance is more likely to improve than degrade. Our study thus rather serves as a proof-of-concept of an approach that might only be further reinforced by newer models. While we specifically used comments from two distinct platforms (Zooniverse & Reddit) that afford opportunities for informal learning, there are more that could be used for such analyses, including formal learning opportunities such as MOOC forums. However, this was out of the scope for the current analysis.

## Conclusion and outlook

Taken together, our results support the idea that cognitive engagement can be studied in forum discussions of informal settings, and that LLMs can become powerful tools for enriching this coding process. On a methodological level, we can conclude that while the implementation of LLMs for qualitative coding might depend on the specific use case and require careful consideration regarding label accuracy or computational cost, their potential lies beyond a simple replacement of conventional classifiers or human raters, but rather in their capacity to function as cognitive proxies that make high-inference decisions transparent and diagnosable. Building on the notion of “creating new aides for human understanding” (Wise & Schwarz, 2017, p. 441), LLMs could thus augment the research process by serving as additional, explainable judges in the coding process. On the content level, our results also show that we can identify different levels of cognitive engagement in forum discussions of informal settings, which allows for a qualitative, in-depth analysis of discussion behaviour and learning-related processes, especially when assessed at scale. In future work, we aim to further refine and validate our approach, leveraging the unique capabilities of LLMs not merely as substitutes for human raters but as valuable tools to enrich qualitative analyses, optimise large-scale coding workflows, and systematically study longitudinal shifts in cognitive engagement on the user level to uncover individual progression within informal communities consistent with the ICAP hypothesis (Chi & Wylie, 2014).

**Abbreviations**

CS: Citizen Science; QWK: Quadratic Weighted Kappa; LLM(s): Large Language Model(s); NLP: Natural Language Processing; BERT: Bidirectional Encoder Representations from Transformers; API: Application Programming Interface; JSON: Javascript Object Notation; LIWC: Linguistic Inquiry and Word Count; MOOC(s): Massive Open Online Course(s); CACL: Computer-Supported Collaborative Learning

**Endnotes**

<sup>1</sup>[github.com/krukowski-lab/RPTL-2026-CogEng-LLMs](https://github.com/krukowski-lab/RPTL-2026-CogEng-LLMs); <sup>2</sup>[zooniverse.org](https://zooniverse.org); <sup>3</sup>[deepinfra.com](https://deepinfra.com)

**Acknowledgements**

We would like to thank the human raters for their participation in this endeavour and the valuable judgments they provided.

**Author's contributions**

Simon Krukowski designed the research approach and study, developed the prompt design and classification methodology, implemented the classification pipeline, curated and prepared the data, conducted the formal analysis, verified the results, created all figures and data visualisations, drafted the manuscript, and coordinated the overall project. H. Ulrich Hoppe contributed to the scientific framing and conceptualisation of the study, provided guidance on methods and interpretation, and critically reviewed and revised the manuscript. Daniel Bodemer supervised the research process and critically reviewed and revised the manuscript. All authors read and approved the final manuscript.

**Author's information**

**SK** is a research associate at Fraunhofer FKIE, Germany. This work was conducted while he was a PhD candidate at the University of Duisburg-Essen, Germany. His research interests include the data-driven study of user behavior and learning processes in large online communities, with a focus on computational methods and social network analysis. **UH** is an emeritus professor at the University of Duisburg-Essen, Germany, and a fellow of the APSCE and ISLS societies. His research focuses on computer support for learning and knowledge building, with extensive experience in EU projects on Technology-enhanced Learning through his COLLIDE research group. His core interest in Learning Analytics is the integration of computational methods, especially combining social network analysis with other data analysis and mining techniques. **DB** holds the chair of Research Methods in Psychology—Technology, Learning, Collaboration at the University of Duisburg-Essen, Germany. His research focuses on analysing and guiding self-regulated, collaborative, and technology-enhanced learning processes across formal and informal educational contexts. He is an active member of EARLI, ISLS, and APSCE.

**Funding**

The authors received no specific funding for this research from any funding agency in the public, commercial, or not-for-profit sectors.

**Availability of data and materials**

The data can be made available on reasonable request. The additional information regarding the coding process and prompt development can be found in the supplementary materials on the linked GitHub repository. The supplementary files include:

- supplementary\_material1.pdf
- supplementary\_material2.pdf

**Declarations****Competing interests**

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup> Research Methods in Psychology—Technology, Learning, Collaboration, University of Duisburg-Essen, Lotharstr. 65, 47057 Duisburg, Germany

<sup>2</sup> RIAS Institute / University of Duisburg-Essen, Bürgerstr. 15, Duisburg, 47057, North Rhine-Westphalia, Germany

Received: 27 October 2025 Accepted: 27 August 2026

Published online: 1 January 2027 (Online First: 7 October 2026)

## References

- Aristeidou, M., & Herodotou, C. (2020). Online citizen science: A systematic review of effects on learning and scientific literacy. *Citizen Science: Theory and Practice*, 5(1), 11. <https://doi.org/10.5334/cstp.224>
- Barany, A., Nasir, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., & Baker, R. S. (2024). ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (pp. 134–149). Springer Nature. [https://doi.org/10.1007/978-3-031-64299-9\\_10](https://doi.org/10.1007/978-3-031-64299-9_10)
- Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., Bibi, A., Trager, R., Fornasiere, D., Yan, J., Elazar, Y., & Bengio, Y. (2025). Chain-of-thought is not explainability. <https://www.alphaxiv.org/abs/2025.02>
- Besta, M., Barth, J., Schreiber, E., Kubicek, A., Catarino, A., Gerstenberger, R., Nyczyk, P., Iff, P., Li, Y., Houlston, S., Sternal, T., Copik, M., Kwaśniewski, G., Müller, J., Flis, Ł., Eberhard, H., Chen, Z., Niewiadomski, H., & Hoefler, T. (2025). Reasoning language models: A blueprint (arXiv:2501.11223). *arXiv*. <https://doi.org/10.48550/arXiv.2501.11223>
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120. <https://doi.org/10.1073/pnas.2218523120>
- Bosch, N., Williams-Dobosz, D., & Perry, M. (2024). Measuring help-seeking in online course discussion forums with privacy-preserving large language models. In J. Clarke-Midura, I. Kollar, X. Gu, & C. D'Angelo (Eds.), *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning—CSCL 2024*. International Society of the Learning Sciences. <https://repository.isls.org/handle/1/10507>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
- Del Valle, M. E., Grudz, A., Kumar, P., & Gilbert, S. (2020). Learning in the wild: Understanding networked ties in Reddit. In N. B. Dohn, P. Jandrić, T. Ryberg, & M. De Laat (Eds.), *Mobility, data and learner agency in networked learning* (pp. 51–68). Springer. [https://doi.org/10.1007/978-3-030-36911-8\\_4](https://doi.org/10.1007/978-3-030-36911-8_4)
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805). *arXiv*. <https://doi.org/10.48550/arXiv.1810.04805>
- Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. *Proceedings of the 16th International Conference on Educational Data Mining*, 103–113. <https://doi.org/10.5281/zenodo.8115784>
- Ferreira Mello, R., Freitas, E., Cabral, L., Pereira, F. D., Rodrigues, L., Rakovic, M., Raniel, J., & Gašević, D. (2024). Words of wisdom: A journey through the realm of natural language processing for learning analytics—a systematic literature review. *Journal of Learning Analytics*, 11(3), 82–105. <https://doi.org/10.18608/jla.2024.8403>
- Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3), 613–619. <https://doi.org/10.1177/001316447303300309>
- Galikyan, I., Admiraal, W., & Kester, L. (2021). MOOC discussion forums: The interplay of the cognitive and the social. *Computers & Education*, 165, 104133. <https://doi.org/10.1016/j.compedu.2021.104133>
- Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, 17(4), 397–431. <https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG>
- Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., Akata, Z., & Schulz, E. (2024). Machine psychology (arXiv:2303.13988). *arXiv*. <https://doi.org/10.48550/arXiv.2303.13988>
- Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1). <https://doi.org/10.1177/20531680241236239>
- Hoppe, H. U., Harrer, A., Göhnert, T., & Hecking, T. (2016). Applying network models and network analysis techniques to the study of online communities. In U. Cress, J. Moskaliuk, & H. Jeong (Eds.), *Mass collaboration and education* (pp. 347–366). Springer. [https://doi.org/10.1007/978-3-319-13536-6\\_17](https://doi.org/10.1007/978-3-319-13536-6_17)
- Houston, S. L., Brady, K., Narasimham, G., & Fisher, D. (2017). Pass the idea please: The relationship between network position, direct engagement, and course performance in MOOCs. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, L@S '17*, 295–298. <https://doi.org/10.1145/3051457.3054008>
- Hudgins, W., Lynch, M., Schmal, A., Sikka, H., Swenson, M., & Joyner, D. A. (2020). Informal learning communities: The other massive open online 'C'. *Proceedings of the Seventh ACM Conference on Learning @ Scale, L@S '20*, 91–101. <https://doi.org/10.1145/3386527.3405926>
- Irani, L. (2015). The cultural work of microwork. *New Media & Society*, 17(5), 720–739. <https://doi.org/10.1177/1461444813511926>
- Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., Dawson, S., Graesser, A. C., & Brooks, C. (2018). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research*, 88(1), 43–86. <https://doi.org/10.3102/0034654317740335>
- Kloetzer, L., Scheider, D., Jennett, C., Iacovides, I., Eveleigh, A., Cox, A. L., & Gold, M. (2013). Learning by volunteer computing, thinking and gaming: What and how are volunteers learning by participating in virtual citizen science? In *Changing configurations of adult education in transitional times: Conference proceedings* (pp. 73–92). European Society for Research on the Education of Adults (ESREA). <https://eprints.whiterose.ac.uk/id/eprint/132704/>
- Kumar, P., Grudz, A., Haythornthwaite, C., Gilbert, S., Del Valle, M. E., & Paulin, D. (2018). Learning in the wild: Coding Reddit for learning and practice. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 1933–1942. <https://research.rug.nl/en/publications/learning-in-the-wild-coding-reddit-for-learning-and-practice>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>

- Liu, S., Liu, S., Liu, Z., Peng, X., & Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Computers & Education*, 181, 104461. <https://doi.org/10.1016/j.compedu.2022.104461>
- Malzahn, N., Schwarze, V., Eimler, S. C., Aprin, F., Moder, S., & Hoppe, H. U. (2023). How to measure disagreement as a premise for learning from controversy in a social media context. *Research and Practice in Technology Enhanced Learning*, 18, 012. <https://doi.org/10.58459/rptel.2023.18012>
- Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2025). Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. *Journal of Learning Analytics*, 12(1). <https://doi.org/10.18608/jla.2025.8591>
- Pavlovic, M., & Poesio, M. (2024). The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation (arXiv:2405.01299). *arXiv*. <https://doi.org/10.48550/arXiv.2405.01299>
- Prestridge, S., Tan, S. C., Jacobsen, M., Hoppe, H. U., Angeli, C., Milrad, M., Pangen, S. K., Kovatcheva, E., Kafyulilo, A., Flanagan, B., & Khaddage, F. (2024). Disconnected connections of learning beyond formal schooling through human–computer–human interactions. *Technology, Knowledge and Learning*, 29(4), 1791–1807. <https://doi.org/10.1007/s10758-024-09779-6>
- Raković, M., Marzouk, Z., Liaqat, A., Winne, P. H., & Nesbit, J. C. (2020). Fine grained analysis of students' online discussion posts. *Computers & Education*, 157, 103982. <https://doi.org/10.1016/j.compedu.2020.103982>
- Rohden, F., Kullenberg, C., Hagen, N., & Kasperowski, D. (2019). Tagging, pinging and linking – user roles in virtual citizen science forums. *Citizen Science: Theory and Practice*, 4(1), 19. <https://doi.org/10.5334/cstp.181>
- Schellens, T., & Valcke, M. (2005). Collaborative learning in asynchronous discussion groups: What about the impact on cognitive processing? *Computers in Human Behavior*, 21(6), 957–975. <https://doi.org/10.1016/j.chb.2004.02.025>
- Sharma, P., Akgun, M., & Li, Q. (2024). Understanding student interaction and cognitive engagement in online discussions using social network and discourse analyses. *Educational Technology Research and Development*, 72(5), 2631–2654. <https://doi.org/10.1007/s11423-023-10261-w>
- Tseng, S.-F., Tsao, Y.-W., Yu, L.-C., Chan, C.-L., & Lai, K. R. (2016). Who will pass? Analyzing learner behaviors in MOOCs. *Research and Practice in Technology Enhanced Learning*, 11(1), 8. <https://doi.org/10.1186/s41039-016-0033-5>
- Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>
- Wang, X., Wen, M., & Rosé, C. P. (2016). Towards triggering higher-order thinking behaviors in MOOCs. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 398–407. <https://doi.org/10.1145/2883851.2883964>
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1), 71–95. <https://doi.org/10.1016/j.compedu.2005.04.003>
- Wise, A. F., & Cui, Y. (2018). Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. *Computers & Education*, 122, 221–242. <https://doi.org/10.1016/j.compedu.2018.03.021>
- Wise, A. F., & Schwarz, B. B. (2017). Visions of CSCL: Eight provocations for the future of the field. *International Journal of Computer-Supported Collaborative Learning*, 12(4), 423–467. <https://doi.org/10.1007/s11412-017-9267-5>

## Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Research and Practice in Technology Enhanced Learning (RPTeL)**  
is an open-access journal and free of publication fee.