Annotating the Annotators: Analysis, Insights, and Modelling from an Annotation Campaign on Persuasion Techniques Detection

Davide Bassi[1†], **Dimitar Iliyanov Dimitrov**[2‡], **Bernardo D'Auria**[3], **Firoj Alam**[4], **Maram Hasanain**[4], **Christian Moro**[3], **Luisa Orrù**[3], **Gian Piero Turchi**[3], **Preslav Nakov**[5] and **Giovanni Da San Martino**[3††]

[1] CiTIUS, Spain [2] Sofia University "St. Kliment Ohridski", Bulgaria

[3] University of Padova, Italy [4] Qatar Computing Research Institute, HBKU, Qatar

[5] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates

[†]davide.bassi@usc.es;[‡]mitko.bg.ss@gmail.com;[††]giovanni.dasanmartino@unipd.it

Abstract

Persuasion (or propaganda) techniques detection is a relatively novel task in Natural Language Processing (NLP). While there have already been a number of annotation campaigns, they have been based on heuristic guidelines, which have never been thoroughly discussed. Here, we present the first systematic analysis of a complex annotation task –detecting 22 persuasion techniques in memes–, for which we provided continuous expert oversight. The presence of an expert allowed us to critically analyze specific aspects of the annotation process. Among our findings, we show that inter-annotator agreement alone inadequately assessed annotation correctness. We thus define and track different error types, revealing that expert feedback shows varying effectiveness across error categories. This pattern suggests that distinct mechanisms underlie different kinds of misannotations. Based on our findings, we advocate for an expert oversight in annotation tasks and periodic quality audits. As an attempt to reduce the costs for this, we introduce a probabilistic model for optimizing intervention scheduling.

1 Introduction

Annotated corpora are critical in Natural Language Processing (NLP) and supervised machine learning, for training and evaluation (Gururangan et al., 2020; Peters et al., 2019). The quality of these annotations is thus crucial (Sun et al., 2017). However, datasets often exhibit label variance (Cabitza et al., 2023), due to multiple factors: multiple valid categorizations (Bechmann and Bowker, 2019), task complexity (Salminen et al., 2018), the concentration and the proficiency required to detect and to understand the phenomenon (Jagabathula et al., 2017), annotators' susceptibility to cognitive biases, such as overconfidence, confirmation and availability bias, anchoring and halo effect (Eickhoff, 2018).

Poor annotations can substantially impact machine learning algorithms' training and evaluation, especially on benchmark datasets (Northcutt et al., 2021; Klie et al., 2023). While many studies used multi-annotation protocols and inter-annotator agreement for reliability assessment (Abercrombie et al., 2023), Baledent et al. (2022) highlighted that this does not evaluate annotation correctness, failing to determine whether annotators consistently misunderstood label applications. Riezler and Hagmann (2022a) called for sound theoretical definitions and effective methodological strategies to evaluate and intervene in annotation. Riezler and Hagmann (2022b) addressed the issue by introducing a post-hoc measure for assessing the annotation correctness after task completion. Yet, this does not accommodate *in-itinere* assessment, aimed at evaluating and enhancing label correctness *during the annotation process* through "decision hygiene" interventions on annotators (Kahneman et al., 2016).

In this work, we analyze the annotation campaign of Dimitrov et al. (2021a) aiming to get valuable insights and ultimately to improve the annotation protocol. The task is a multimodal and multilabel one that requires annotating 22 persuasion techniques in memes (see Section 3 for further details and Figure 1 for an example of a meme). This presents three main challenges: (*i*) labels require theoretical expertise and extensive training, (*ii*) the task's cognitive demands are high due to its multimodal and multilabel nature with 22 techniques, and (*iii*) accurate span detection requires precise text boundary identification.

Our contributions are as follows: (*i*) we present the first study with continuous expert oversight, for the persuasion techniques detection task, that monitor the process, check the annotations, and provide feedback to the team; (*ii*) we examine the Inter-Annotator Agreement as an indicator of annotation correctness, demonstrating its limitations; (*iii*) we analyze the impact of the feedback sessions on an-

Findings of the Association for Computational Linguistics: ACL 2025, pages 17918–17929

Figure 1: Example of a meme (Dimitrov et al., 2021b); the technique *Thought-terminating cliché* is present in the text, while *Causal oversimplification* can be detected by analyzing both the textual and the visual content.

notator performance and (*iv*) discuss annotator error patterns; (*v*) based on our finding that complex annotation tasks require expert supervision, we propose a probabilistic model of annotators' evolving error rates for adaptive expert intervention scheduling aimed at optimizing costs while maintaining annotation quality; (*vi*) we release a comprehensive dataset that enables researchers to investigate the relationship between annotation variance incorporating the labels from the annotators, the consolidations, and the experts.[1]

The rest of the paper is organized as follows: Section 3 recaps the annotation task protocol and describes the expert's role. Section 2 discusses similar analysis in literature. Section 4 presents our research questions and methodology for analyzing annotator performance, and Section 5 discusses the results. Section 6 introduces a probabilistic model for expert intervention scheduling based on annotators' performance. Section 7 concludes with future research directions.

2 Related Work

Ground truth in Machine Learning (ML) and NLP has been extensively debated. Plank (2022) argued that while some tasks can have objective ground truth (e.g, identifying verb types), other are inherently subjective or allow multiple valid interpretations, configuring disagreement as a potential resource. Building on this perspective, Basile (2020) and Casola et al. (2023) demonstrated how preserving annotation variance can enhance models' flexibility and effectiveness, e.g., for inherently ambiguous phenomena such as irony and hate speech.

However, Cabitza et al. (2023) pointed out the need for a more precise qualification of annotations' variance. Still adopting a perspectivist approach, a dataset characterized by high variance could constitute either a rich or a noisy resource. While inter-annotator agreement can quantify this variability, it cannot qualify it. This distinction is particularly crucial for complex tasks such as persuasion technique detection, where despite some room for subjective interpretation, adherence to fundamental rules remains essential for dataset validity.

Several frameworks have been proposed to address disagreement evaluation. Prabhakaran et al. (2024) (Sandri et al., 2023) and Jiang and de Marneffe (2022) developed taxonomies to analyze disagreement patterns and their underlying causes in subjective tasks. Expert involvement represents another common approach to ensuring annotation quality. Sánchez-Montero et al. (2025) compared expert and non-expert annotations of Spanish metaphors, while Chau et al. (2020) explored expert participation in keyphrase extraction, focusing on cost optimization.

Most closely related to our work, Stefanovitch and Piskorski (2023) analyzed inter-annotator agreement in a large-scale multilingual campaign for annotating persuasion techniques in news articles. Their focus was primarily on measuring the agreement across languages and on developing new metrics for this purpose.

Overall, these approaches, primarily focused on post-hoc analysis, providing useful retrospective insights about misannotation reasons and mechanisms. Our work differs by introducing expert supervision during the annotation process itself, moving beyond measuring annotators' agreement to assess the annotation correctness: by integrating expert oversight while the annotation campaign is ongoing, this enables continuous quality improvement through targeted feedback sessions.

3 Annotation Campaign Setup

The annotation task is about detecting persuasion techniques in memes; see Appendix A for a complete list and definitions, and Dimitrov et al. (2021a) for the detailed guidelines given to the annotators. The resulting dataset was used in the shared task described in Dimitrov et al. (2021a). The annotation campaign followed the pipeline described in Figure 2.

[1] https://joedsm.github.io/pt-corpora/memes/

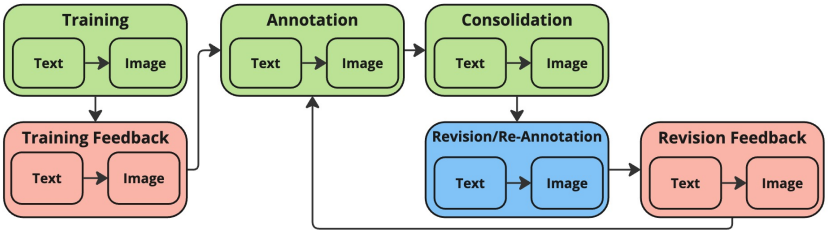


Figure 2: Graph of the phases of the annotation process, shown as colored boxes: **green** phases involve the *annotators*, **blue** phases involve the *expert*, and **red** phases involve both the *annotators* and the *expert*.

We can see in the top row (corresponding to Sections 3.1 and 3.3) replicates the setup in Dimitrov et al. (2021a), while the bottom one details how the expert supervised the campaign, training and giving feedback to the annotators. Each phase is described in the following sub-sections. The annotations were outsourced to two companies (AT1 and AT2) with prior experience in similar tasks. Both the annotators and the consolidator were fairly compensated[2] professionals who possessed C1/C2 levels of English proficiency and held at least a bachelor's degree. The two teams AT1 and AT2 analyzed a total of 9,000 and 1,000 memes, respectively. The expert checked 1,360 memes annotated by AT1 and 301 by AT2. For this study, we focus our analysis exclusively on the subset of memes that were annotated by both a team and the expert.

3.1 Training Phase

The task of detecting persuasion techniques in memes requires an understanding of specific knowledge about argumentation, logical fallacies, and rhetoric. To help the annotators gain this knowledge, they received task guidelines (see Dimitrov et al. (2021a) and a video tutorial featuring expert-annotator discussions of various examples. Then they were asked to annotate a controlled sample that included multiple usages for every technique.

3.2 Training Feedback

During the training phase the expert met with the research team to address any uncertainties about the techniques or the task.

3.3 Annotation Phase

Each meme was annotated by two annotators who performed *independently* two sequential tasks:

Text Annotation (TXT): given the list of the *textual persuasion* techniques[3], the annotators had to identify which techniques were present in the extracted textual content of the meme (without seeing the visual content) and in which text span they occurred. For the meme in Figure 1 the input would simply be "*This is your child* \n *This is your child on vaccines* \n *PERIOD.*", and there would only be one technique: *Thought-terminating cliché*.

Image Annotation (IMG): Given the entire meme (both the image and the text), the annotators have to detect the techniques present in the meme. Note that this time, they could choose from all 22 techniques. Moreover, they were shown the actual meme immediately after having annotated the text only. All techniques identified in the text only would apply to the meme as well, but there could be additional ones (for example in Figure 1 *Thought-terminating cliché* in the text and additionally *Causal oversimplification* in the meme).

3.4 Consolidation Phase

In this phase, the annotators met and discussed with a *consolidator* each annotation, in order to reach a consensus, with the possibility of deleting or adding newly identified technique instances.

This multi-phase design aimed to reduce the annotators' cognitive load through task decomposition, while implementing a weak perspectivist approach (Cabitza et al., 2023) by enabling the annotators to discuss and merge their individual assessments during consolidation discussions.

3.5 Expert Revision/Re-annotation Phase

To analyze the correctness of annotations and address the complexity of the task, we enhanced the protocol from Sections 3.1 and 3.3 by adding an expert in argumentation and persuasion techniques.

[2]Compensation averaged $2.06 per meme for the complete annotation process, including two individual annotations and one consolidation.

[3]We have 20 techniques as *Transfer* and *Appeal to (Strong) Emotions* only apply to visual content. See Appendix A.

This role served two key purposes: to assess the correctness of the annotations and, by operating throughout the annotation process, to enable real-time improvement of the annotators' performance. The expert's contribution in this phase included the following:

- Checking 100 *consolidated memes* weekly (i.e. memes that went through the consolidation phase described above), re-annotating them to create a "*expert gold-label dataset*".
- Creating a weekly report regarding the annotators' performance. By analyzing error frequencies and patterns, the expert generated statistical insights about the distribution of missed, incorrect, and correct annotations across different persuasion techniques.

3.6 Revision Feedback Phase

The expert met weekly with the annotation team: both the annotators and the consolidators. These sessions were occasions, for the expert, to provide feedback on possible errors that emerged. The annotators, on the other side, had the opportunity to ask questions and to discuss hard cases encountered while carrying out the task, thus improving their understanding of the techniques.

4 Research Questions and Methodology

We systematized the objectives of this paper in the following research questions: **RQ1** Is the inter-annotator agreement(IAA) a useful indicator of annotation correctness? **RQ2** Do the revisions of the expert allow us to observe characteristics about the annotation quality that, otherwise, would go unnoticed? **RQ3** Do the follow-up feedback sessions help to improve the annotators' performance?

To investigate RQ1, we analyzed the agreement using AT1's longitudinal dataset, which provided sufficient data for temporal analysis. We calculated four types of Cohen's Kappa coefficients for each session:

Consolidated vs. Expert labels: we compared the consolidated annotations to the expert labels to track the annotation quality over time.

Annotator Average vs. Expert: between each annotator and the expert labels, then averaging them to assess the pre-consolidation quality.

Annotator Average vs. Consolidated: we compared the labels by each annotator and the consolidated labels, then averaged the scores to measure the annotation changes during consolidation.

Inter-Annotator Agreement: for each annotator pair per session based on their overlapping entries, then averaging these pairwise agreements to obtain a single score per session.

To answer RQ2 and RQ3, we defined and tracked the occurrences of errors from a taxonomy of possible errors. First, we identified and removed the correct annotations, defined as cases where the same technique was identified in both consolidated annotations and the expert review, by finding matching techniques between the annotators and the reviewers in both TXT and IMG steps. Second, we analyzed the remaining misannotations as follows:

Substituted annotation (SUB): when both the consolidated annotation and the expert identified a technique in the same text span (75% overlap), but disagreed on which specific technique was present (e.g., *Loaded Language* vs. *Name Calling*), requiring replacement. Substitutions were not considered for the annotations in the image step, since no overlap is possible. After handling text span overlaps through SUB errors, subsequent error analysis focused solely on technique presence/absence, treating the task as a multilabel classification problem.

Incorrect annotations (INC) occurred when a technique present in the consolidated annotation was deemed absent by the expert. These were identified by comparing techniques present in annotators' labels but absent in the expert's.

Missed annotations (MIS) represented cases where the expert identified a technique that was absent in the consolidated annotation. These were detected by finding techniques present in the expert's labels but absent in the annotator's.

5 Results

5.1 Research Question 1

Figure 3 shows the Pearson's correlation analysis of Cohen's Kappa coefficients for RQ1. The low, non-significant correlations between agreement metrics and expert annotations (gold standard) suggest that these metrics alone do not reliably indicate annotation correctness.

This underscores how, while inter-annotator agreements constitute a reliability measure (how consistent they are between themselves), annotators' and expert's Cohen's Kappa is a measure of annotation's correctness. Given this, the former can be understood as a pre-requisite for the latter, but it does not guarantee the correctness of the annotations (Paun et al., 2022).

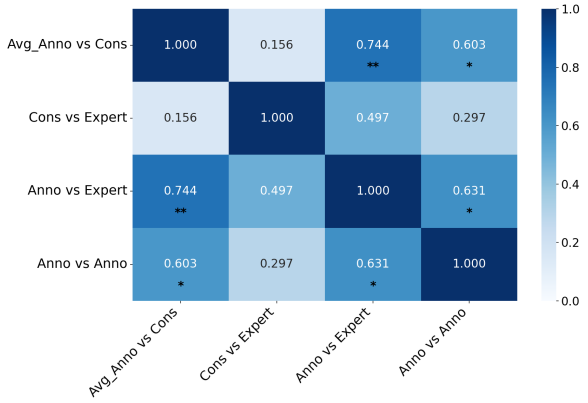


Figure 3: Agreement correlation heat-map.

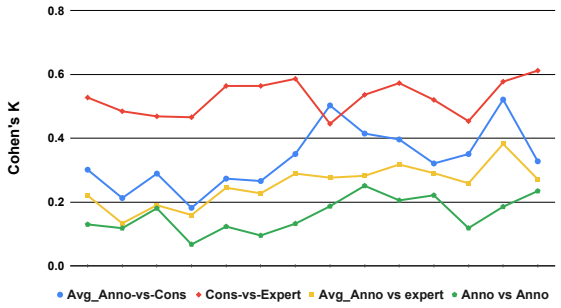


Figure 4: K-Alpha across the annotation period.

Additionally, Figure 4 reveals distinct patterns across different agreement levels. Consolidated vs. Expert annotations (*Cons-vs-Expert*) maintain consistently higher agreement (0.46-0.61) compared to Annotator Average vs. Expert (*Avg_Anno vs expert*, 0.13-0.38), demonstrating the effectiveness of the consolidation phase in improving annotation quality. This improvement is further evidenced by two metrics: Annotator Average vs. Consolidated (*Avg_Anno-vs-Cons*) shows moderate agreement (0.2-0.5), indicating substantial refinements during consolidation that align annotations closer to expert standards. Additionally, the low inter-annotator agreement before consolidation (*Anno vs Anno*) highlights how the discussion process effectively resolves initial disagreements among annotators.

5.2 Research Question 2

RQ2 deepens the analysis of misannotation patterns. After classifying them into the categories in Section 4, we calculated the total number of occurrences of each type, as shown inTable 1. Statistical analysis showed that the missed annotations (MIS) occurred significantly more often than incorrect (INC) and substituted (SUB) ones, as confirmed by two one-sided Z proportion tests.

Error	Freq.	Z Tests
INC	706	MIS vs INC: Z=36.85***
MIS	1102	MIS vs SUB: Z=26.25***
SUB	193	

Table 1: Errors' frequencies and Z tests ($p < .001$).

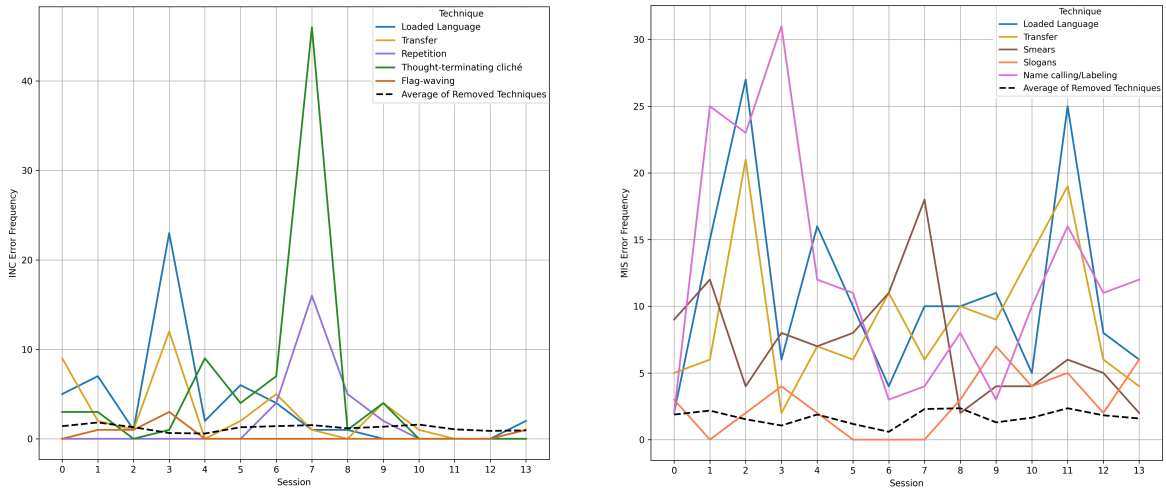
These findings offer an initial understanding concerning RQ2, indicating that the MIS typology represents the most frequent misannotation.

To further investigate the origins of these misannotations, we plotted technique-specific trends for INC and MIS misannotations in each weekly report (which, as described in Section 3.5, preceded each Revision Feedback weekly meeting), as shown in Figure 5a and 5b for AT1[4], and in Figure 6a and 6b for AT2. From these graphs, two additional insights can be obtained: INC misannotations tend to have a more "peaky trend," i.e., a generally low level of incorrect labels interspersed by peaks of errors in specific labels. In our case, this happened especially for *Thought-Terminating Cliché*, *Loaded Language*, and *Repetition* techniques in sessions 3 and 7. MIS misannotations, on the contrary, are characterized by a more spread trend, i.e., lower peaks and a generally smoother trend. Even if it were possible to detect some "critical points" (such as in sessions 2, 3 and 11), Figure 5b shows missing errors for techniques such as *Loaded Language*, *Name-Calling*, and *Smears* throughout the annotation campaign. AT2's data mirrors these patterns: despite the shorter annotation period, Figure 6a shows a notable peak in session 13 followed by a decline, while Figure 6b exhibits multiple peaks paralleling AT1's MIS error distribution.

5.3 Research Question 3

We compared for AT1 the errors in week 1, where the annotators did not get any feedback, to weeks 2,3,4 and 2–13 (all the rest of the annotation campaign) by running a Z-test with $\alpha = 0.01$. The difference is significant for weeks 3,4 and 2–13, suggesting that expert interventions improved the quality of annotations. We argue that we need an effective initial training, but the "acute peaks" in both AT1 (weeks 3 and 7) and AT2 (week 13) INC graphs (Figure 5a and Figure 6a respectively) can be used as indicators of expert interventions' effectiveness during annotation: they show rapid increase in misannotations followed by quick reductions post-feedback.

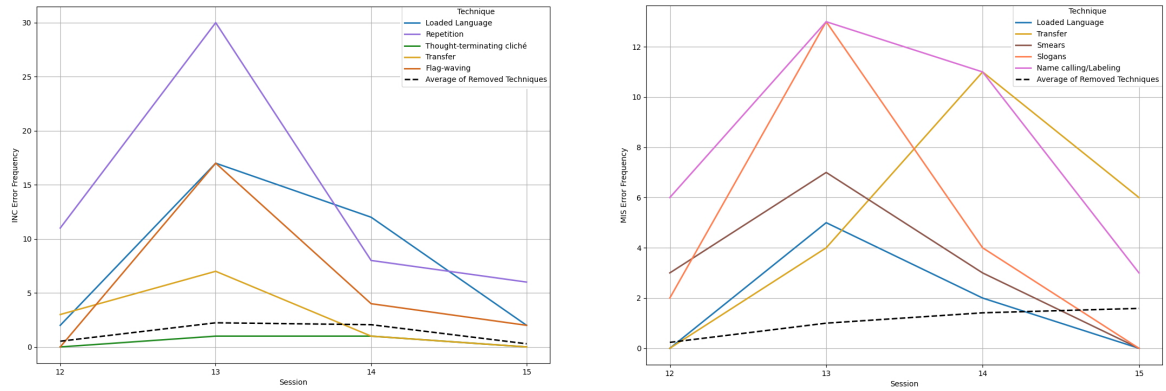
[4]Techniques with low error frequencies are aggregated and shown as the black dotted line for readability.



(a) Incorrect labels per technique

(b) Missing labels per technique

Figure 5: Annotation Team 1 (AT1): misannotation typology per technique.



(a) Incorrect labels per technique

(b) Missing labels per technique

Figure 6: Annotation Team 2 (AT2): misannotation typology per technique.

MIS-type errors showed different patterns. Expert interventions proved less effective in this case, with Figures 5b and 6b showing these errors maintaining elevated frequencies across multiple sessions and techniques, suggesting a lower reduction in this type of misannotation.

These distinct patterns point to different underlying mechanisms. INC errors, occurring when annotators recognize a form of manipulation in the meme, but assign an incorrect technique label to it, appear to stem from a misunderstanding of label definitions and can be effectively addressed through additional training sessions. MIS errors, reflecting a deeper failure in manipulation detection, show a more persistent and erratic trend, suggesting additional underlying factors such as cognitive overload (Chen et al., 2023), biases and attention fluctuations (Gautam and Srinath, 2024), as well as varying subjective interpretations of meme content (Sandri et al., 2023).

6 Toward Efficient Revision Scheduling via Probabilistic Modeling of Error Rates

Our analysis demonstrates that expert supervision can improve annotation quality, particularly for incorrect annotations (Section 5.3). However, expert involvement is resource-intensive, raising the question of how to optimize revision and intervention timing while maintaining good quality. We propose a probabilistic framework for adaptive scheduling based on annotators' evolving error rates. The key insight is that as annotators learn from expert feedback, their error probability should decreases.

Let us start by formalizing the annotation setting (we will inevitably simplify it to keep the model manageable). In the first phase, each annotator analyzes $N \geq 0$ items in autonomy, then curates them with a consolidator, resulting in a final error probability $p \in [0, 1]$.

An erroneously identified item implies a cost $c > 0$. The cost could be interpreted in broader terms than economical: for the model to be sound, it needs a disincentive to make an error; otherwise, there would be no gain in having the expert supervise the annotations. The number of undetected errors, say X, is distributed according to a Binomial distribution with parameters N and p. After the N sentences, additional $M > 0$ sentences are annotated, this time also checked by an expert, who receives a compensation $d > 0$ for their work. The expert identifies and corrects all errors in the M annotations. The number of errors found, say Z, are trivially distributed as an independent Binomial distribution with parameters M and p.

The Z errors are shown and discussed with the annotators and the consolidator, allowing them to improve their ability to perform the task, reducing the error rate p. We model such update with the function

$$f(p, Z) := \frac{p}{1 + bZ} \tag{1}$$

where $b > 0$, p is the error rate before the expert's consultation, and $f(p, Z)$ is the updated rate after receiving the expert's feedback Z.

We assume the interest rate is $a \in (0, 1)$[5], and we call $T(p)$ the average total cost incurred in this improvement process.

The average total cost satisfies the following recurrent equation:

$$T(p) = c\,\mathbb{E}[X] + a^{N(p)}\left(d + a\,\mathbb{E}[T(f(p, Z))]\right), \tag{2}$$

where $N(p)$ is again the number of items annotated by the annotators before the intervention of the expert, but now we define it in terms of the error rate p (our goal is precisely to determine when the expert should do the next intervention based on the current ability to perform the task of the annotators). Eq. (2) states that the cost of the annotation campaign starting with an error rate p depends on the average errors done when the annotators are unsupervised, the discounted price of the expert intervention, and the discounted price of the remaining campaign whose probability p to make errors has been updated according to Eq. (1).

[5]To simplify the calculations, it is standard practice to assume an infinite number of annotations, which in turn require an interest rate to allow for convergence.

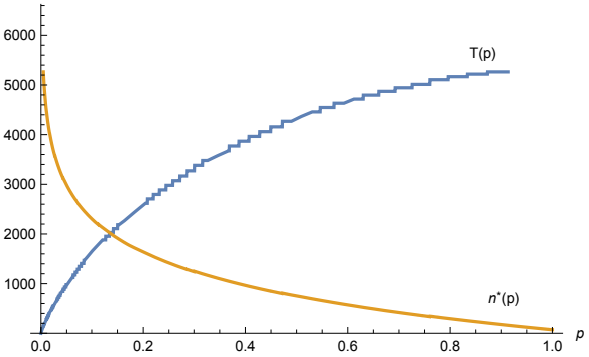


Figure 7: Example of the total amount $T(p)$ as a function of the initial error rate p, and the best choice for $N(p)$, i.e. $n^*(p)$. The parameter values are as follows: $a = 99.9\%$, $b = 4.6 \times 10^{-4}$, $c = 5.0$, $d = 150.0$ and $M = 362$.

Since $Z \sim \mathrm{Bin}(M, p)$, we have that

$$\mathbb{E}[T(f(p, Z))] = q^M T(p) + \bar{T}(p)$$

$$\bar{T}(p) = \sum_{z=1}^{M} \binom{M}{z} p^z q^{M-z}\, T(f(p, z)),$$

$q = 1 - p$, and, substituting it in (2), $T(0) = 0$ and

$$T(p) = \frac{cpN(p) + a^{N(p)}(d + a\bar{T}(p))}{1 - a^{N(p)+1} q^M}. \tag{3}$$

By (2), it is clear that T is an increasing function, and that, by (3) and (1), it depends recursively on lower values of its parameter p. In Appendix B, we show how to derive a closed-form solution for $T(p)$ and the best value for $N(p)$, which we denote as $n^*(p)$.

Figure 7 shows the values of $T(p)$ and $n^*(p)$ as a function of the error rate p. Note that the values on the y-axis are not expressed in terms of examples: in the model, we assumed that the errors have a Binomial distribution, and, since our problem is multiclass multilabel, for each example, we take multiple decisions, and therefore we can make multiple errors as a result. One example therefore corresponds to multiple units on the y-axis. We chose for an example to correspond to the average number of the gold labels plus one (to account for all non-gold-labels wrongly annotated) units, but any other sensible choice would have led to a similarly shaped plot. If the annotators have an error $p = 0.8$ then the expert should intervene after $n^*(0.8) = 294$ decisions (roughly corresponding to 80.08 examples in our corpus) with an estimated cost $T(0.8) = 5167.23$.

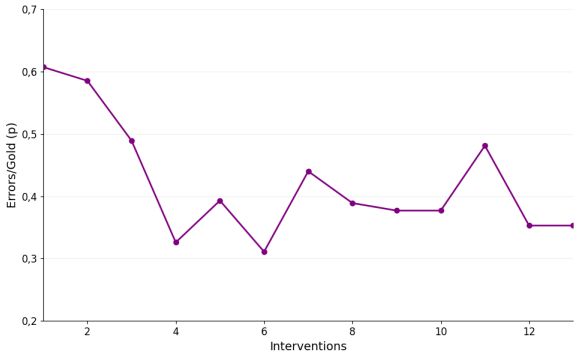


Figure 8: Actual error rate p of the annotators in our annotation campaign.

Here, we consider the values of the parameters a, b, c, d as indicated in the caption of Figure 7, and we assume that p would not change. In practice, p would actually change after the intervention by the expert annotator, therefore resulting in new (lower) values of $n^*(p)$ and $T(p)$. In our experimentation, the expert annotator intervened at fixed regular times for a cost of 5,469. By applying the model and thus intervening at different time intervals according to $n^*(p)$, the estimated cost $T(p)$ would be 4,716. We would like to note that this comparison should by no means be interpreted literally. To have a fair comparison, we should have repeated the annotation campaign in the two different settings, and eq. (1) should be a good model of the learning curve of the annotators. While it is infeasible to redo the annotation campaign, in Figure 8, we plot the actual values of p computed on the data we collected.

The overall decreasing trend in error rates aligns with our proposed model of annotators' learning through expert feedback, as formalized in eq. (1). However, notable spikes in error rates occur at interventions 7 and 11, corresponding to the peaks previously identified in Figure 5a and Figure 5b, respectively. As discussed in Section 5.3, these temporary increases in errors stem from different underlying causes. If the expert intervention strategy could be optimized to promptly identify these errors (the plot is currently based on the actual interventions that were made at regular intervals), the learning curve would likely show a smoother decrease in error rates, more closely matching our theoretical model in eq. (1).

7 Conclusion and Future Work

We examined dataset annotation quality through a case study of persuasion techniques in memes (Dimitrov et al., 2021a). Our work extends the ongoing discussion about annotation error detection in machine learning (Plank, 2022; Klie et al., 2023; Cabitza et al., 2023), emphasizing the need to move beyond inter-annotator agreement to assess annotations correctness. The task of persuasive memes detection underscores the importance of such evaluations, particularly when computational models address misinformation that influences public opinion (Bassi et al., 2024), making dataset integrity both a technical requirement and a matter of political and social accountability (Nannini et al., 2024). A key contribution of our work is the introduction of a weak perspectivist-based annotation protocol (Cabitza et al., 2023) (allowing annotators to merge their individual assessments during consolidation discussion) integrated with expert supervision during the annotation campaign, enabling continuous correctness management through systematic revisions and feedback.

We addressed three research questions (section 4). Our analysis of RQ1 shows that inter-annotator agreement does not correlate with annotation correctness, challenging its use as a sole quality metric. However, discussions among annotators significantly improved annotation correctness, likely due to enhanced awareness of edge cases and collective reasoning strategies. Results from RQ2 revealed distinct error patterns, while through RQ3 we showed how expert supervision effectively addressed incorrect annotations through targeted feedback, missed annotations remained more resistant to improvement across both annotation teams, suggesting systematic cognitive limitations rather than team-specific issues.

Additionally, expert supervision proved effective for improving annotation quality, but remains resource-intensive. To address this constraint, we introduced a probabilistic model that optimizes the frequency of expert interventions while maintaining annotation quality.

As future work we plan to extend the model to differentiate between incorrect and missed annotations and customize it to individual annotator characteristics, such as learning curves. Finally, we plan to integrate our framework into annotation platforms to enable real-time quality monitoring and automated intervention strategies.

Limitations

Our study used a weak-perspectivist approach rather than the strong-perspectivism proposed by (Cabitza et al., 2023), which preserves multiple annotations throughout model training and evaluation. While this approach would be valuable for persuasion technique detection, scope constraints prevented its implementation. To address this limitation we released the complete set of annotations.

Expert annotation validation relied on informal team reviews of a label subsample. As demonstrated by Schmid et al. (2021), experts can disagree and make mistakes, suggesting the need for more systematic validation approaches.

Our study lacks a control group to compare error trends with and without expert feedback, which would have provided stronger empirical evidence for the effectiveness of our intervention.

We partially address these resource-related limitations through our proposed statistical model.

Ethics and Broader Impact

Our research on annotation quality has important ethical implications that we carefully considered. To protect privacy, we fully anonymized the identities of all annotators, consolidators, and experts involved in the data collection process.

While our study focuses on technical aspects of annotation quality, we acknowledge potential societal impact when this work is applied to content moderation systems. The detection of persuasive techniques in memes, while valuable for identifying potential manipulation, could inadvertently reinforce existing biases or create new ones if implemented without proper safeguards. Though we strived for objectivity in our annotation protocol, we recognize that unintended biases may exist in how persuasion techniques are identified and labeled.

We emphasize that automated systems built on this data should be deployed only as support tools for human moderators, who can provide necessary context and nuanced judgment, helping to mitigate the risks of systematic biases against particular groups or communication styles.

Finally, all annotators in our study provided informed consent, were fully aware of the study's objectives, and had the right to withdraw at any time. They were also appropriately compensated as part of their job.

The datasets will be released under the CC BY-NC-ND 4.0 license.

Acknowledgments

This project has received funding from the European Unión´s Horizon Europe research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 101073351.

Declarations

Some parts of this manuscript's English language and writing clarity were enhanced using Claude 3.5, an AI language model developed by Anthropic. However, the authors reviewed and carefully verified every single suggested improvement.

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *ArXiv*, abs/2301.10684.

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin. 2022. Validity, agreement, consensuality and annotated data quality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2940–2948. European Language Resources Association.

Valerio Basile. 2020. It's the end of the gold standard as we know it - leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *AIxIA 2020 - Advances in Artificial Intelligence - XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25-27, 2020, Revised Selected Papers*, volume 12414 of *Lecture Notes in Computer Science*, pages 441–453. Springer.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

Anja Bechmann and Geoffrey C. Bowker. 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data Soc.*, 6(1):205395171881956.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6860–6868. AAAI Press.

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Teresa Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Hung Chau, Saeid Balaneshin, Kai Liu, and Ondrej Linda. 2020. Understanding the tradeoff between cost and quality of expert annotations for keyphrase extraction. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 74–86, Barcelona, Spain. Association for Computational Linguistics.

Ouhao Chen, Fred Paas, and John Sweller. 2023. A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, 35(2):63.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6603–6617. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 162–170, New York, NY, USA. Association for Computing Machinery.

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in NLP. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 82–88, Mexico City, Mexico. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics.

Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 18(1):3233–3299.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Daniel Kahneman, Andrew M Rosenfield, Linnea Gandhi, and Tom Blaser. 2016. Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard business review*, 94(10):38–46.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Luca Nannini, Eleonora Bonel, Davide Bassi, and Michele Joshua Maggini. 2024. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI and Ethics*, pages 1–29.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical methods for annotation analysis*. Springer Nature.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*, pages 7–14. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Riezler and Michael Hagmann. 2022a. Validity. *Synthesis Lectures on Human Language Technologies*, page 9 – 53. Cited by: 0.

Stefan Riezler and Michael Hagmann. 2022b. Validity, reliability, and significance: Empirical methods for nlp and data science. *Synthesis Lectures on Human Language Technologies*, 14(6):1 – 147. Cited by: 2; All Open Access, Bronze Open Access, Green Open Access.

Joni O. Salminen, Hind A. Al-Merekhi, Partha Dey, and Bernard J. Jansen. 2018. Inter-rater agreement for social computing studies. In *Fifth International Conference on Social Networks Analysis, Management and Security, SNAMS 2018, Valencia, Spain, October 15-18, 2018*, pages 80–87. IEEE.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Disagreement in metaphor annotation of Mexican Spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Annette M Schmid, David L Raunig, Colin G Miller, Richard C Walovitch, Robert W Ford, Michael O'Connor, Guenther Brueggenwerth, Josy Breuer, Liz Kuney, and Robert R Ford. 2021. Radiologists and clinical trials: part 1 the truth about reader disagreements. *Therapeutic Innovation & Regulatory Science*, 55:1111–1121.

Nicolas Stefanovitch and Jakub Piskorski. 2023. Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 71–86, Singapore. Association for Computational Linguistics.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 843–852. IEEE Computer Society.

A List of Persuasion Techniques and their Definition

Both Textual and Visual Techniques:
Loaded Language: *Using specific words and phrases with strong emotional implications to influence an audience.*
Name calling/Labeling: *Labeling an entity as either something the target audience fears, hates, finds undesirable, or loves, praises.*
Smears: *An effort to damage or to call into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups.*
Doubt: *Questioning the credibility of someone or something.*
Exaggeration/Minimisation: *Representing something in an excessive manner, making it larger, better, worse; or making it seem less important or smaller than it really is.*
Slogans: *A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.*
Appeal to fear/prejudice: *Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases, the support is built based on preconceived judgments.*
Whataboutism: *A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.*
Glittering generalities: *Words or symbols in the value system of the target audience that produce a positive image when attached to a person or issue. Virtue can be also expressed in images, where a person or an object is depicted positively.*
Flag-waving: *Playing on strong national feeling (or to any group such as race, gender, political preference) to justify or promote an action or idea.*
Repetition: *Repeating the same message, so that the audience eventually accepts it.*
Causal Oversimplification: *Assuming a single cause or reason when there are actually multiple causes for an issue. It includes transferring blame to one person or group of people without investigating the complexities of the issue.*
Thought-terminating clichè: *Words or phrases that discourage critical thought and meaningful discussion about a given topic.*
Black-and-white Fallacy/Dictatorship: *Presenting two alternative options as the only possibilities, when in fact more possibilities exist. It includes dictatorship, where one tells the audience exactly what actions to take, eliminating any other choices.*
Straw Man: *An opponent's proposition is substituted with a similar one, which is then refuted in place of the original proposition.*
Appeal to authority: *Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered.*
Reductio ad hitlerum: *Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated or in contempt by the target audience.*
Obfuscation/Int. vagueness/Confusion: *Using words that are deliberately unclear, so that the audience may have their own interpretations.*
Presenting Irrelevant Data: *Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.*
Bandwagon: *Attempting to persuade the target audience to join in and take the course of action because 'everyone else is taking the same action.'*
Visual-Only Techniques:
Transfer: *Projecting positive or negative qualities (praise or blame) of a person, entity, object, or value onto another one to make the second one more acceptable or to discredit it.*
Appeal to (Strong) Emotions: *Using images with strong positive/negative emotional implications to influence an audience.*

B Derivation of the Solution of $T(p)$ and $n^*(p)$

We show how to solve eq. (2) and derive the optimal value for $N(p)$, i.e. $n^*(p)$. We copy here for convenience eq. 3:

$$T(p) = \frac{cpN(p) + a^{N(p)}(d + a\bar{T}(p))}{1 - a^{N(p)+1}q^M} .$$

Assuming for a moment that $N(p) = n$ is constant, we can rewrite the total cost as a function of two arguments,

$$T(p, n) = \frac{cpn + a^n(d + a\bar{T}(p))}{1 - a^{n+1}q^M} , \tag{4}$$

that allows to optimize for the value of n, in search for its minimum value, as a function of p.

Taking derivatives, we have

$$\begin{aligned}\partial_n T(p, n) =& \frac{a^n \log(a) \left(aq^M cpn + d + a\,\bar{T}(p)\right)}{\left(1 - a^{n+1}q^M\right)^2} \\ &+ \frac{pc(1 - a^{n+1}q^M)}{\left(1 - a^{n+1}q^M\right)^2} .\end{aligned} \tag{5}$$

We have that $\lim_{n\to\infty} \partial_n T(p, n) = pc$, and

$$\partial_n T(p, 0) = \frac{\log(a) \left(d + a\,\bar{T}(p)\right)}{\left(1 - aq^M\right)^2} + \frac{pc}{1 - aq^M} .$$

To have $\partial_n T(p, 0) \le 0$ it should be $\bar{T}(p) \ge t(p)$ where

$$t(p) := \log_a \left(\frac{1}{e}\right) \frac{1 - aq^M}{a} pc - \frac{d}{a} .$$

The values for n where the derivative is non-negative satisfy the follow inequality

$$a^n \left(n + \frac{\bar{T}(p) + d/a}{cpq^M} - \frac{1}{\log(a)}\right) \le -\frac{1}{\log(a) aq^M} ,$$

and the minimum is attained at the largest non-negative $n = n^*(p)$ such that

$$a^{n^*+\gamma(p)} \left(n^*(p) + \gamma(p)\right) = \alpha(p) , \tag{6}$$

so that $n^*(p) = \infty$ whenever there are no non-negative solutions to (6).

In (6), we made the following definitions

$$\alpha(p) := \frac{a^{\gamma(p)-1}}{q^M} \log_a(1/e) , \tag{7}$$

$$\gamma(p) := \frac{1}{pq^M} \frac{\bar{T}^*(p) + d/a}{c} + \log_a(1/e) . \tag{8}$$

For $a \in (0, 1)$, let $W_a(z)$ be the maximum real solution, w, of the equation $w\, a^w = z$, for $z \le (1/e) \log_a(1/e)$. It follows that $W_a(z) = W_{-1}(z \log(a))/\log(a)$ with $W_{-1}(z)$ being the secondary branch of the Lambert W function, defined as the minimal real solution, w, of the equation $w\, e^w = z$, for $z > -1/e$.

By (6), we have

$$n^*(p) = \left(W_a\left(\alpha(p)\right) - \gamma(p)\right)^+ , \tag{9}$$

with $(x)^+ = \max\{0, x\}$.

To make sense, the argument of the W_a in (9) should be less than $(1/e) \log_a(1/e)$, therefore

$$\begin{aligned}\gamma(p) &\le \log_a(q^M/e) + 1 \\ \bar{T}^*(p) &\le cpq^M \log_a\left(q^M\right) - d/a .\end{aligned}$$

In Eq. (9), if $\alpha(p) > \gamma(p) a^{\gamma(p)}$, then $n(p) > 0$ otherwise $n(p) = 0$.