Quality Assurance in Spam, Fraud, and Abuse Fighting

Google

Abstract

Quality assurance as a discipline goes back hundreds of years, but its adoption within spam and fraud detection, content moderation, and abuse fighting i.e., the Trust & Safety industry, is nascent. While many of the techniques and standards deployed in manufacturing, health care, aviation, etc. are useful here, the Trust & Safety space comes with its own sets of challenges which require rethinking how quality measurement and improvement is approached. The present whitepaper focuses on the **measurement** aspect, i.e., how to most effectively and efficiently evaluate the quality of manual reviews performed by spam, fraud, and trust & safety agents. We discuss which quality metrics to rely on, how to effectively sample for quality measurement, how to obtain reliable quality metrics (including how to address the lack of definitive 'ground truth'), and how to ensure that the quality metrics are representative of the work performed by the agents (i.e., unbiased).

Certain sections go into more detail than what is needed to obtain a high-level understanding of the topic, and these can be skipped. These sections have been highlighted in blue.

Introduction

Manual reviews play a critical role in enforcing the violations of rules and standards intended to ensure the safety and trust of users in various online products and platforms i.e., the Trust & Safety industry. Many companies (tech- and social media companies, banks, and others) rely on manual reviews to protect users and their products. In this context, It is essential to be able to monitor the quality of decisions made by manual reviewers to ensure accurate outcomes and enable trusted experiences.

The present whitepaper provides an overview of *best practices in regards to measuring the quality of manual reviews*. It focuses on the development of robust quality metrics which enable regular monitoring of performance and allows us to take immediate action if a deterioration in quality is detected. Any significant drop in regularly monitored quality metrics immediately initiates a **r**oot **c**ause and **c**orrective **a**ction (RCCA). The RCCA aims to identify 1) the primary root causes of the quality issues and 2) associated corrective actions. Post implementation of the corrective actions, the monitoring process is again key to assessing the effectiveness of each of those and to ensure that improved performance is sustained.

In documenting our sampling and metrics approach, we wanted to ensure that anyone working in the Trust & Safety industry - quality experts as well as quality novices - could understand and when needed apply these approaches.

Measurement concepts and definitions

Before diving into the details, let's introduce some terminology which we will be using throughout this paper.

Defining 'quality'

In this paper, we will be concerned with the quality of the *outcome* of the review process only. That is, we focus only on the manual reviewers making the *correct decisions*, regardless of the process to get to that decision. The quality of the process (i.e. 'process compliance') may be of interest for a variety of reasons, including how it relates to outcome quality, how it ensures consistency regardless of outcome, and the role it places in assurance. On the other hand, focusing on complying with a process as a way to ensure outcome quality may miss defects that are caused by agents coming to the wrong conclusion, despite perfect process compliance. In practice, measuring process compliance in the world of digital content reviews is complex and fraught with potential error. Thus, *we recommend measuring decision quality* and working backwards in a way to unearth any root causes around process non-compliance.

The standard quality metrics and their interpretation

When assessing quality in the Trust & Safety industry, we are concerned with:

1. To what extent manual reviewers are able to catch spam, fraud, abuse, and other policy violations;
2. To what extent manual reviewers label/enforce only on items that are in fact spammy, fraudulent, or abusive.

#1 is typically measured in terms of *recall*, and #2 is typically measured in terms of precision. More details on quality metrics are provided in the following section.

Constructing quality metrics

Assessing quality relies on comparing the ground truth to the applied label for a given set of content.

- The **ground truth label** is the true state associated with a reviewable item. The ground truth state is 'positive' if the item is truly policy violating. A state is 'negative' if the item is *not* policy violating.
- The **applied label** is the label *applied* to a reviewable item by a human agent. The applied label state is 'positive' if an item is *labeled* as policy violating and 'negative' if an item is *labeled* as not policy violating.

Thus for any entity, the ground truth and applied label can be aligned (positive-positive, negative-negative) or misaligned (negative-positive, positive-negative).

We then construct quality metrics using three core comparison metrics:

- **Precision (PPV):** The fraction of all items with a positive applied label whose ground truth label is also positive.
- **True Positive Rate, or Recall (TPR):** The fraction of all items with a positive ground truth label whose applied label is also positive.
- **False Positive Rate (FPR):** The fraction of all items with a negative ground truth label whose applied label is positive.

We can more easily visualize these definitions as they relate to the confusion matrix:

	ground-truth positive	ground-truth negative	
applied positive	TP	FP	PPV = TP / (TP + FP)
applied negative	FN	TN	FOR = FN / (FN + TN)
	TPR = TP / (TP + FN)	FPR = FP / (FP + TN)	

It is worth noting that recall and false positive rate are both defined relative to the ground-truth label, and are therefore generally more robust to the prevalence of a particular abuse type in the sample for human review (i.e. fraction of ground truth positives in the overall set of entities for review). Although precision is useful to know operationally (and is a metric for which targets are commonly set), it is inherently sensitive to prevalence. An increase in prevalence in the review queue may artificially increase precision without an associated improvement of the raters' ability to correctly rate items (and vice versa).

Measurement approaches to manual review processes

Manual review processes

Companies (tech- and social media companies, banks, and others) rely on manual reviews for a range of different activities:

1. *Direct Policy enforcement:* items which algorithms struggle to classify as abuse/non-abuse (a.k.a. 'gray area cases') are submitted for manual review.
2. *Appeals and User Communication:* When users appeal or otherwise seek additional feedback on a policy enforcement action, manual reviewers evaluate whether the account or content in question is indeed policy compliant, and grant appeals accordingly.
3. *Indirect Enforcement via Machine Learning:* Manual reviewers assign labels to specific content and entities which are used in training algorithms to automatically detect spam, fraud, and abuse.

Review flows are typically organized into 'queues' which surface a stream of content that is related by policy violation and/or product in order to facilitate efficient and accurate reviews. Within the queues, there are typically two tiers of manual reviewers charged to overseeing the review of content and policy violating entities:

- *Tier 1 agents* perform the manual reviews described above (policy enforcement, appeals handling, labeling). Tier 1 can consist of multiple agents independently reviewing the same item (a.k.a. a *multi-rated workflow*) and entities can be reviewed for multiple kinds of violations at once.
- *Tier 2 agents (a.k.a. QAs, or QA agents)* evaluate the quality of the work performed by the Tier 1 agents to assess for accuracy of decision and surface any doubts or common issues to policy experts. Tier 2 agents are typically agents who have developed a certain level of experience and subject matter expertise, by working and upholding a high quality standard as Tier 1 agents for an extended period of time.

Assumption

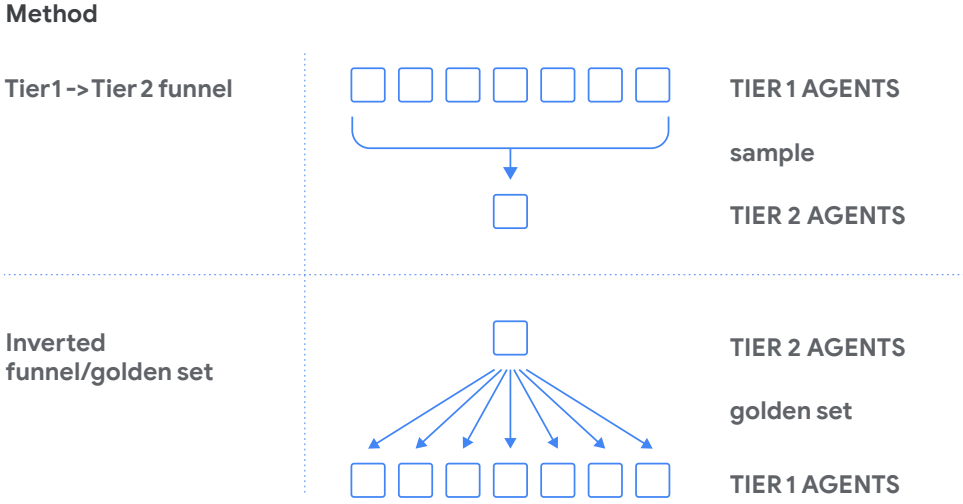
We consider the decision provided by Tier 2 agents to be the 'ground truth', as defined above. We do recognize, however, that Tier 2 agents have imperfect quality, as does any human being. In the sections, 'Tier 1:Tier 2 appeals process' and 'Independent oversight of quality', we discuss ways of compensating for this and correcting the metrics for Tier 2 agent defects.

Quality evaluation process

The Tier 2 agents described above review and assess quality by re-review of previously reviewed content. In general, there are two ways of assessing the quality of manual reviews (illustrated below):

- Standard Tier 1 -> Tier 2 funnel: a sample of items reviewed by Tier 1 agents is submitted for re-review by the Tier 2 agents.
- Inverted funnel (golden set) approach: Tier 2 agents review a representative sample of items from the Tier 1 workflows. We call this a *golden* set. The golden set is submitted for (blind) review by all Tier 1 agents.

Quality is defined as the agreement between Tier 1 and Tier 2 decisions in either the sample or the golden set. More on this below.



Applying quality metrics in single-rater workflows

In single rater review flows, the construction of the core quality metrics is relatively straightforward:

- **PPV:** The probability that a positive label applied by a Tier 1 agent is **correct** given the label of the Tier 2 agent;
- **TPR:** The probability that a Tier 1 agent **correctly** applies a *positive label* to an item whose *ground truth (Tier 2) label* is also positive;
- **FPR:** The probability that a Tier 1 agent **incorrectly** applies a *positive label* to an item whose *ground truth (Tier 2 agent) label* is negative.

Again, it is worth explicitly noting that these metrics rely on the assumption that Tier 2 agent decisions represent the ground truth.

Applying quality metrics in multi-rater workflows

The above definitions apply to single-agent workflows. For multi-rater workflows there are two broad definitions of quality:

- **Agent-level quality:** PPV, TPR, and FPR defined at the agent-level decision.
- **Aggregate-level quality:** PPV, TPR and FPR defined at the aggregate-level decision.

Aggregate-level quality therefore consists of the same metrics, except they're defined in terms of the applied-label after aggregating over all agents' decisions. It therefore depends on the aggregation rules employed in a particular workflow. As an example, in a workflow where the labels are binary (i.e. positive or negative) and a majority vote between three agents is used, the aggregate-level quality metrics would be defined based on the majority vote label obtained from aggregating the three agents' decisions. Consider the following example reviews, with mistakes highlighted in red and positive labels in bold:

Agent 1 (T1)	Agent 2 (T1)	Agent 3 (T1)	Aggregate	Ground-truth (T2)
Positive	**Positive**	Negative	**Positive**	Negative
Positive	**Positive**	**Positive**	**Positive**	**Positive**
Negative	Negative	Negative	Negative	Negative
Negative	**Positive**	Negative	Negative	Negative
Negative	Negative	Negative	Negative	Negative

Quality metrics would be measured as:

Agent 1 (T1)	PPV	FOR	TPR	FPR
Agent 1	0.5	0.0	1.0	0.25
Agent 2	0.5	0.0	1.0	0.50
Agent 3	1.0	0.0	1.0	0.0
Aggregate	0.5	0.0	1.0	0.25

Which population should be considered for quality assessment?

The quality of work performed by Tier 1 agents can be measured relative to different populations. The populations most frequently used for quality measurement are:

1. **Enqueued items only**
   With this approach, quality is measured relative to a sample of the items which are enqueued for review on a day-to-day basis.
2. **Entire corpus**
   With this approach, quality is measured relative to the entire corpus of items *eligible* for review. For example, in their day-to-day work, Tier 1 agents may be exposed to only so-called 'gray area' content, but one could still evaluate the quality of their work relative to all items eligible for review.
3. **Curated set**
   Policy specialists may curate a set of items to be used when assessing the quality of manual reviews. For example, when rolling out a new policy where the performance relative to the new policy is of particular interest, policy specialists may curate the set in such a way that it addresses all the primary applications of the new policy.

Because the quality metrics that one measures are dependent on the population considered for quality assessment, relative comparisons on comparable populations (either between reviewers or over time) are often more informative than the absolute value of the metrics.

The following table describes the pros and cons of each of the approaches above:

Population considered for quality evaluation	Pros	Cons
Enqueued items only	Representativeness: the quality metrics most accurately reflect the quality of the output from the regular review process.	Potential volatility: if the content of the review queue changes significantly over time (e.g., if the prevalence of policy-violating items fluctuates or if the prevalence of gray-area cases fluctuates), these changes in themselves can cause the quality metrics to fluctuate.
Entire corpus	Stability in metrics: the composition of the corpus is expected to be relatively stable over time. Hence, the corpus composition in itself should not cause fluctuations in the quality metrics.	Translation: one may not be able to easily relate the quality metrics to the quality of day-to-day work, especially if the items enqueued for review on a daily basis don't represent the entire corpus well.
Curated set	Focus: ability to focus the quality measurement on the topic of interest (e.g., reviewable items which touch upon all the key aspects of a new policy).	Only suitable for one-off measurement: a set curated for this month's measurement, say, may be very different from a set curated for next month's measurement. This makes it difficult to compare metrics over time and say anything meaningful about improvement or deterioration in quality.

Tier 1:Tier 2 appeals process

As mentioned previously, our assumption when computing quality metrics is that the Tier 2 agent provides the ground truth against which Tier 1 quality is measured. However, there are instances where the Tier 1 agent may have made the right decision, despite disagreement between Tier 1 and Tier 2, some of which are mentioned here:

A. Human error: the Tier 2 agent could be wrong

B. Dynamic content: the content of a website or the signals on an account at the time of the Tier 1 review may differ from that at the time the Tier 2 review is completed

C. Policy gray area: the policy in question may leave room for interpretation, to the extent that the Tier 1 and the Tier 2 decisions could both be viewed as policy-compliant

D. Abuse-type not addressed by policy: the Tier 1 agent may have come across a new type of abuse which is not yet addressed by any policy

Being able to track the occurrence of all of the above is important in and of itself. We therefore recommend establishing an appeals process, by which Tier 1 agents can appeal the decisions of Tier 2 agents. An appeal is granted if it is agreed that the Tier 1 agent made the correct decision at the time of review.

The appeals process allows one to:

1. Adjust the quality metrics for all of the above (A-D): we recommend reporting the raw quality metrics as well as those adjusted for any appeals granted
2. Indirectly track the performance of Tier 2 agents: if appeals are frequently granted due to A), then this alerts us of Tier 2 quality issues
3. Revisit policies to address C) and/or D).

As such, the appeals process provides insights which are critical to the understanding and continuous improvement of quality.

Independent oversight of quality

While the above-mentioned Tier 1:Tier 2 appeals process helps identify instances where the Tier 2 agents make incorrect decisions, it doesn't help identify instances where all agents agree to be wrong. For example, agents (Tier 1 and 2) may over time have agreed to interpret the policy in a certain way, which is inconsistent with the spirit of the policy. Such issues can be surfaced only with some level of *independent oversight* of quality. To this end, any or all of the below can help ensure alignment between agents and policy owners/experts, thereby establishing trust in the quality metrics:

1. Frequent calibration/gray area sessions in which policy owners and Tier 2 agents go through particularly challenging examples and align their decisions
2. Policy owners review a sample of Tier 2 decisions on a regular basis (similar to the 'funnel approach' described below)
3. Policy owners review a representative sample of items which is submitted for blind review by Tier 2 (possibly also) Tier 1 agents on a regular basis (similar to the 'inverted funnel approach' described below)

Types of quality workflows

The process of quality sampling and review can be designed in a number of different ways. We describe the four most commonly used approaches in the following sections and summarize it all with a side-by-side comparison of the four.

Funnel approach (FA)

Under the standard funnel approach, a sample of reviews performed by Tier 1 agents is submitted for review by the Tier 2 agents. Quality metrics are computed based on the agreements and disagreements between Tier 1 and Tier 2. The Tier 2 reviews may be blind or non-blind. Pros and cons of this approach are described below.

Pros	Cons
• Ease of implementation, relatively speaking. • Ease of metrics computation, incl. confidence intervals/sample size estimation.	• Not knowing whether the Tier 2 agent has perfect quality and hence to what extent we can trust the metrics. This can to some extent be dealt with by allowing Tier 1 agents to appeal the decisions of Tier 2 agents and via independent oversight by policy owners (cf. the previous sections). • The Tier 2 resource requirement is higher than for the Inverted Funnel Approach (described below). • Unable to guarantee a certain sample size per Tier 1 agent (which tends to be of interest to the ones who coach and manage individual agents).

Inverted funnel approach (IFA)

With this approach, Tier 2 agents (or other subject-matter experts) review a representative sample of items from the Tier 1 workflows. We call this a *golden set*. The golden set is submitted for (blind) review by all Tier 1 agents, and their quality is evaluated based on the agreement with the Tier 2 decisions.

Pros	Cons
• More economical from a Tier 2 resource perspective than the Funnel Approach described above: one can obtain more accurate[1] metrics with the same amount of Tier 2 resources. • Ability to drill down to and compare agent-level quality (which may be of particular interest to the ones who coach and manage individual agents): all Tier 1 agents review the same golden set of items.	• Not knowing whether the Tier 2 agent has perfect quality and hence to what extent we can trust the metrics. This can to some extent be dealt with by allowing Tier 1 agents to appeal the decisions of Tier 2 agents and via independent oversight by policy owners (cf. the previous sections). • Technical infrastructure requirements: Parallel routing of a single reviewable item to a Tier 2 agent and to all Tier 1 agents Blind review process required Sample size/confidence interval computations are more involved May need distinct golden sets for different workflows (e.g. workflows using different languages or dealing with different content).

[1] 'Accurate' in the sense of low margin of error.

Multi-rated funnel approach (MR-FA)

In the two approaches described above, we mentioned that Tier 2 agents may not have perfect quality. The quality metrics produced by the Funnel Approach or the Inverted Funnel Approach should therefore be interpreted with some level of caution. While it is hard to entirely solve this problem, one can at least increase the reliability of the metrics by replacing each Tier 2 agent decision by the majority of 3 or more Tier 2 agents. That is each item sampled from the Tier 1 corpus of reviews would be reviewed (blindly) by 3 or more Tier 2 agents. The Tier 1 decision would be compared against the majority vote amongst the Tier 2 agents. The pros and cons of this approach are described below:

Pros	Cons
• By introducing multiple Tier 2 reviews per item, we increase the overall quality of the Tier 2 decision-making. • The level of disagreement/consensus between Tier 2 agents can provide useful information about clarity of policies and general difficulty of the review process.	• Most expensive from a Tier 2 resourcing perspective. • One needs to have the technical infrastructure required to allow Tier 2 agents to review the same item independent of one another. • Systemic mis-interpretations of policies and similar by the Tier 2 agents would not be compensated for in this process.

Multi-rated inverted funnel approach (MR-IFA)

Lastly, one can also introduce multiple Tier 2 ratings into the Inverted Funnel process. This has all the benefits of the Inverted Funnel Approach but increases our trust in the metrics, in the sense that each item in the golden set would be reviewed by 3 or more Tier 2 agents (and Tier 1 decisions would be compared against the Tier 2 majority vote). While this set-up puts the most demands on technical infrastructure, the resulting metrics are the most trustworthy compared to any of the above.

Pros	Cons
• By introducing multiple Tier 2 reviews per item, we increase the overall quality of the Tier 2 decision-making • The level of disagreement/consensus between Tier 2 agents can provide useful information about clarity of policies and general difficulty of the review process. • Ability to drill down to and compare agent-level quality (which may be of particular interest to the ones who coach and manage individual agents): all Tier 1 agents review the same golden set of items.	• Technical infrastructure requirements: Parallel routing of a single reviewable item to multiple Tier 2 agents and to all Tier 1 agents Blind review process required • Sample size/confidence interval computations are more involved • May need distinct golden sets for different workflows (e.g. workflows using different languages or dealing with different content).

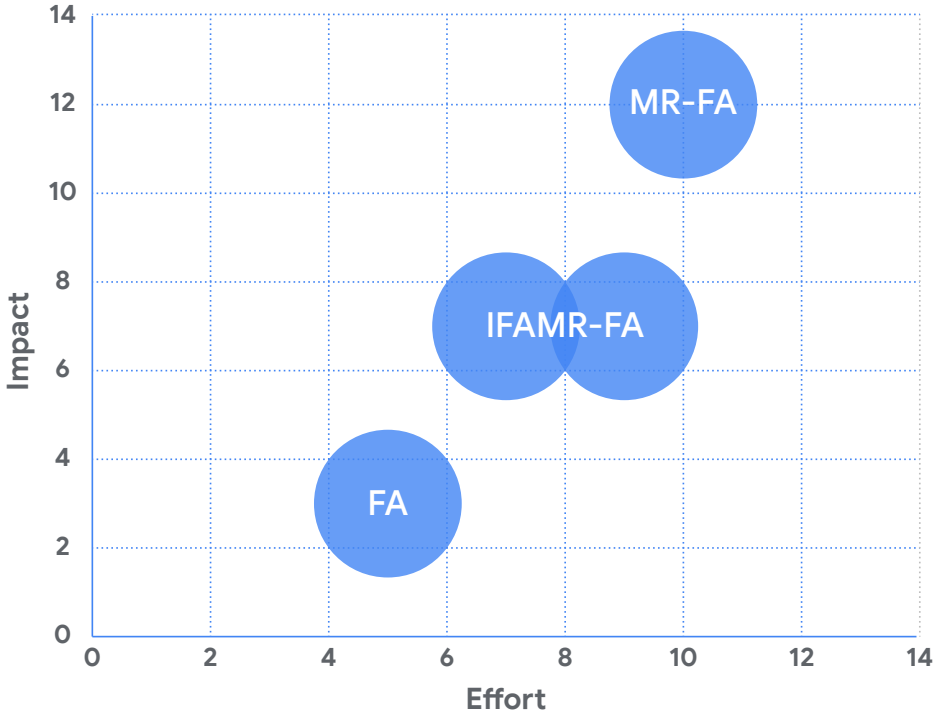
Comparison of quality workflows

The following table compares the above-mentioned types of quality workflows side-by-side, rated on a scale from 1 to 4 (1 = “low”, 4 = “high”):

		Funnel Approach (FA)	Inverted Funnel Approach (IFA)	Multi-rated Funnel Approach (MR-FA)	Multi-rated Inverted Funnel Approach (MR-IFA)
EFFORT	*Infrastructure requirements*	1	3	2	4
	Complexity of metrics computation	1	3	2	4
	Tier 2 resource requirements	3	1	4	2
IMPACT	*Reliability of metrics*	1	2	3	4
	Feasibility of agent-level quality measurement	1	3	1	4
	Ability to derive insights from consensus amongst reviewers	1	2	3	4

In the chart below, we sum up the scores for impact and effort, respectively, for each workflow and plot them against one another[2]. As the chart shows, the Multi-rated Inverted Funnel Approach is the most complex to implement but also provides the most informative and reliable quality metrics (with reasonably low Tier 2 resource requirements). *The Multi-rated Inverted Funnel Approach is our preferred quality measurement framework.*

Impact vs effort for different quality measurement frameworks



[2] We have assigned equal weight to each of the dimensions considered, but one could consider one dimension (e.g., infrastructure requirements) more costly/important than another (e.g., Tier 2 resource requirements) and assign weights accordingly.

Practicalities of measuring quality

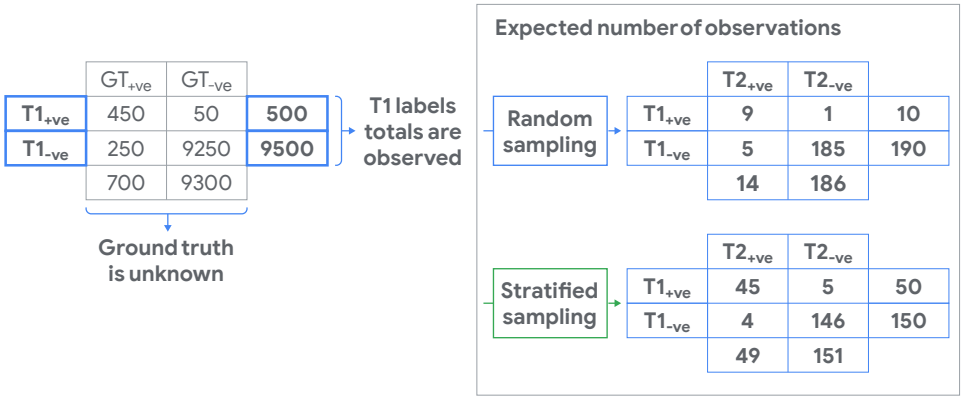
The process of measuring quality depends on which type of quality workflow is employed (cf. the section above), so we discuss the regular funnel and the inverted funnel approaches separately. We will also describe alternative ways of evaluating quality (through multi-rated workflows) in a third subsection.

Measuring quality using the funnel approach

Under the standard funnel approach, a sample of reviews performed by Tier 1 agents is Quality measurement using the FA is done by randomly sampling entities reviewed by Tier 1 agents and having those entities re-reviewed by Tier 2 agents. The process of random sampling has a significant impact on the robustness of the quality metric estimates produced through this workflow. The most straightforward way to obtain a quality sample is to perform **simple random sampling** (i.e. sample Tier 1-reviewed entities uniformly at random). Once this sample is reviewed by Tier 2 agents, the desired metrics can be directly estimated from the resulting confusion matrix. However, abuse fighting/content moderation workflows tend to have **low prevalence**, meaning that the fraction of reviewable entities which are policy-violating is very small. As a result, if one selects a sample of Tier 1-reviewed entities uniformly at random, there will generally be very few positive labels in the sample, making it hard to evaluate the precision (PPV) of these workflows. To combat this, it is common to employ **stratified sampling or weighted (a.k.a. importance) sampling**, in order to oversample entities which were identified as policy-violating by the Tier 1 agents. This introduces bias to the sampling process, which must be corrected for when the estimates of the quality metrics are being calculated. We illustrate these two random sampling methodologies below.

Estimating metrics with simple random sampling

Example



In our example, we have 10k labels generated through Tier 1 reviews (left). Out of those 10k labels, 9500 are negative labels, while only 500 are positive. The GT labels are unknown (otherwise we wouldn't need a quality workflow) but, for the sake of the example, we assume that out of the 500 positive labels, 450 agree with the true label, and 50 do not. Similarly, out of the 9500 negative labels, 9250 agree with the true label, while 250 do not. So, in our example, our true (and unknown) quality metrics read:

- PPV = 450 / 500 = 90%
- TPR = 450 / 700 = ~64%
- FPR = 50 / 9300 = ~0.5%

If we draw a simple random sample (top right) of size 200, we expect to sample 10 positive labels and 190 negative labels, given the proportions in the Tier 1 population. As a result, we have (on average) only 10 labels at our disposal to estimate precision, and only 14 labels available to estimate recall. Even though metric estimates computed from the confusion matrix will be unbiased (i.e. they will average out to the correct value over many measurements), these sample sizes are too small to produce an accurate estimate of either metric.

Stratified sampling, on the other hand, allows us to oversample from the positive labels. In this case, we drew a random sample of 50 reviews from the positives, and an independent sample of 150 reviews from the negatives, so the total sample size is still 200. However, we now have 50 samples at our disposal to estimate precision, resulting in a much more accurate estimate of that metric. However, because we artificially changed the balance of positive vs. negative labels in our sample (relative to the Tier 1 population), we can no longer estimate TPR and FPR from our resulting confusion matrix, without doing some additional calculations (i.e. TPR != 45 / 49, and FPR != 5 / 146). Even though these simple ratios will move in the same direction as the unbiased metrics, failing to correct for the bias introduced by the sampling makes it harder to draw comparisons across different workflows.

Estimating metrics with stratified sampling

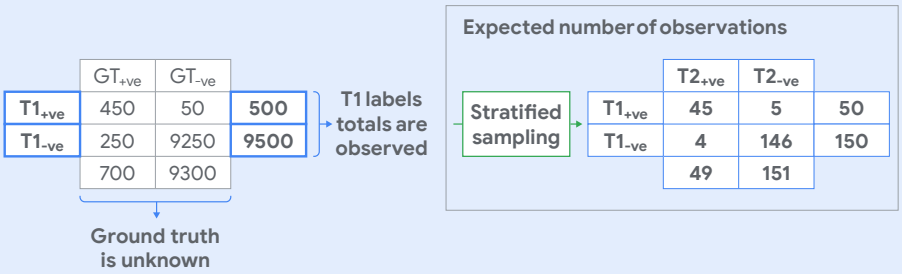
We saw in the example above that we can **stratify** the sample or, alternatively, draw a **weighted** sample, with weights conditional on the Tier 1 label in order to increase the representation of positive labels in the quality sample.
The downside of this, as we mentioned,was that we could no longer directly use the confusion matrix counts to estimate the metrics, due to the bias introduced by the sampling process. This bias can be corrected for using a Horvitz-Thompson or a Hansen-Hurwitz estimator (if sampling with replacement).
The latter results in significantly simpler variance formulas (so we assume it throughout)[3]. Specifically, we need to compute the probability that each unit in the population is selected in the random sample. Unlike for simple random sampling, we have two distinct values for this probability: p_{+ve}, and p_{-ve}, depending on whether the Tier 1 label is positive or negative. With these values at hand, we can estimate the required metrics as follows:

$$\text{TPR} = \frac{TPs/p_{+ve}}{TPs/p_{+ve} + FNs/p_{-ve}}$$

$$\text{FPR} = \frac{FPs/p_{+ve}}{FPs/p_{+ve} + TNs/p_{-ve}}$$

For precision (PPV), we do not need to worry about the weights, since they are the same for every term in the numerator and denominator (namely, p_{+ve}). We revisit our earlier example below.

Example



Based on our sample sizes, we can compute p_{+ve}= 50/500 = 1/10, and p_{-ve}= 150/9500 = 3/190.
Applying our formulas to the expected confusion matrix returns:

- PPV = 45 / 50 = 90%
- TPR = (45 * 500 / 50) / (45 * 500 / 50 + 5 * 9500 / 150) = ~64%
- FPR = (5 * 500 / 50) / (5 * 500 / 50 + 146 * 9500 / 150) = ~0.5%

which coincides with the true values.

[3] When sampling with replacement the sample is typically deduplicated prior to review, in order to prevent wasted review effort. However, the number of instances of each entity in the sample should be correctly accounted for when computing any metrics.

Aggregating multiple stratified samples

Inverse-probability estimators generalize effectively to multiple stratified samples. A common example occurs when measuring the aggregate quality of reviews conducted by multiple agent groups. For example, if we have a group of agents reviewing items in Region A, and another group reviewing items for the same workflow in Region B, we may draw independent stratified samples to measure their respective quality. We may then also want to report their aggregate quality (i.e. the quality of their pooled reviews). We illustrate this with another example.

Example with 2 regions and 2 strata in each region

In our example, we have 10k labels generated through Tier 1 reviews (left).

Region	Stratum	$T1_{population}$	$T2_{population}$	$P_{inclusion}$	w
Region 1	+ve	1000	100	0.1	10
	-ve	10000	100	0.01	100
Region 2	+ve	2000	100	0.05	20
	-ve	50000	100	0.002	500

Where $P_{inclusion}$ denotes the probability that an item from that stratum ended up in the sample (i.e. the sample size divided by the respective stratum size), and w is the inverse of that quantity. Let the true confusion matrices and quality metrics be the following:

Region	TP	FP	TN	FN	PPV	TPR	FPR
Region 1	900	100	9500	500	90%	64.3%	1.04%
Region 2	1500	500	45000	5000	75%	23.1%	1.10%
Total	**2400**	**600**	**54500**	**5500**	**80%**	**30.4%**	**1.09%**

Let the confusion matrices for the quality sample be the following:

Region	TP	FP	TN	FN	w
Region 1	90	10	95	5	10 (TP & FP) 100 (TN & FN)
Region 2	75	25	90	10	20 (TP & FP) 500 (TN & FN)
Total	**165**	**35**	**185**	**15**	

If we were to try to estimate the quality without any probability adjustment we would obtain an incorrect estimate, namely:

Estimates using canonical formulas (wrong)

PPV=TP/(TP+FP)=165/(165+35)=82.5%
TPR=TP/(TP+FN)=165/(165+15)=91.7%
FPR=FP/(FP+TN)=35/(35+185)=15.9%

Using the probability-adjusted formulas yields the correct result

Estimates using inverse-probability estimator (correct)

PPV=(9010+7520)/(9010+1010+7520+2520)=80%
TPR=(9010+7520)/(9010+5100+7520+10500)=30.4%
FPR=(1010+2520)/(1010+95100+2520+90500)=1.09%

Measuring quality using the Inverted funnel approach

The primary difference between measurements obtained with the standard FA and those obtained with the IFA is that the base metrics computed with the latter are **agent-specific metrics**. That is, because in the IFA, each agent reviews the same ground-truth set, we have a direct measurement of their quality metrics on that specific ground-truth set, which can be compared with that of other agents. An important advantage of this approach is that it enables agent-to-agent comparisons (unlike FA, where the total number of QA reviews for each particular agent is typically too small to enable any meaningful comparison between agents) and allows us to assess 'how good it gets' (i.e. it allows us to better evaluate what level of quality the best agents are able to deliver).

Measuring agent-level metrics

Under the IFA, the core principles for selecting the *golden set* are identical to those in the standard FA, and typically a stratified sample is recommended for low prevalence workflows, with the goal of improving the balance between the classes in the golden set. The typical way to achieve this is to draw a golden set using a stratified sample where the stratification criterion is the label applied by the Tier 1 agents.

Once each Tier 1 agent completes their review of the golden set, the same formulas from the previous section can be employed to obtain each agent's quality scores. However, in the FA we are measuring the aggregate quality of the *process* (i.e. the quality of the labels as a whole, aggregated over all reviewers). In the IFA, the direct measurement is on the agent-level quality, so if an aggregate measurement of the process is desired, then there are two options:

Option 1 (Direct Estimate)	Estimate aggregate quality directly from the golden set labels, using the approach illustrated in the FA section. Just like the FA, this suffers from data sparsity, as the size of the golden set tends to be limited.
Option 2 (Agent-level aggregation)	Estimate aggregate quality by aggregating agent-level quality metrics with appropriate weighting. This requires some technical assumptions, but better leverages the large amount of data that the IFA approach produces. Details of this approach are discussed in the next section.

Aggregating agent-level metrics

The discussion of the Funnel Approach (FA) above also provides us with the methodology we need to estimate each agent's quality using a golden set (IFA). Imagine that these are PPV_i, TPR_i, and FPR_i, for a given agent, *i*. We often want to know what the quality of the process is, on aggregate. This kind of metric would be comparable to the one obtained from a standard FA.

However, different agents may contribute to the workflow in different amounts (e.g. part-time work vs. full-time work, experienced agents vs. trainees, etc). Therefore, we often can't simply average the agent-level metrics, and it is generally preferable to weight each agent's metric by the volume of reviews that agent contributed to the workflow. This requires some technical assumptions, namely

Assumption

The distribution of content from which the golden set distribution is drawn is similar to that encountered by each of the agents.

Technical formulation

The probability that a review is routed to agent i is proportional to the rate of reviews of the agent and is independent of the ground truth label of the reviewable.

Example

Note that under this assumption, the prevalence of policy-violating content in each individual agents' queue should be approximately the same[4]. We illustrate the metric estimation process below with a simple example.

	Confusion matrix counts are unknown.				Total +ve and -ve counts are observed for Tier1 agnets		Agent-level quality metrics are estimated from golden set.		
	TP	FP	FN	TN	+ves	+ves	PPV	TPR	FPR
Agent 1	400	100	200	4800	500	5000	80.0%	~66.7%	~2.0
Agent 2	60	50	100	1200	110	1300	~54.5%	37.5%	4.0%
Agent 3	480	20	70	4950	500	5020	96.0%	87.3%	0.4%
Total	**940**	**170**	**370**	**10950**	**1110**	**11320**	**~84.7%**	**~71.8%**	**~1.5%**

In this example, we see that the three Tier 1 agents contribute very different review volumes, with different levels of quality. In particular, we see that Agent 2 contributes a meaningful portion of the errors (FNs, in particular), despite a significantly lower review volume. This kind of behavior needs to be accommodated when aggregating reviewer-level metrics if our goal is to understand the quality of the labels on aggregate. That is, it is not sufficient to simply average the quality metrics of individual agents in order to estimate the aggregate quality. Instead, we may resort to a simple weighted average, as follows:

1. Estimate the probability that a review is routed to agent i as
   a. $P[r=i] = N_i / N$
      e.g. $P[r=1] = (500 + 5000) / (1110 + 11320) = 44.2\%$
2. Estimate aggregate metrics (PPV, TPR, and FPR) as
   a. Metric (aggregate) $= \Sigma_i P[r=i] \times$ Metric (agent i)
      $PPV = 44.2\% \times 80.0\% + 11.2\% \times 54.5\% + 44.4\% \times 96.0\% = 84.1\% \approx 84.7\%$
      $TPR = 44.2\% \times 66.7\% + 11.2\% \times 37.5\% + 44.4\% \times 87.3\% = 72.4\% \approx 71.8\%$
      $FPR = 44.2\% \times 2.0\% + 11.2\% \times 4.0\% + 44.4\% \times 0.4\% = 1.5\%$

There are alternatives to the weighted average approach depending on which technical assumption makes more sense in the context of the workflow. Regardless, the point to note is that aggregate metrics estimated in this way from an IFA approach are only approximate, and typically not unbiased unless the assumption is exactly satisfied.

[4] Such an assumption is typically satisfied (at least approximately) so long as the golden set comes from a single common workflow and the agents all conduct reviews from a single common queue. The IFA (or golden set) approach can't typically be applied outside of this context anyway, since agents reviewing content of a distinct nature cannot generally be evaluated on a common golden set.

Appendices

Label quality vs. decision quality

This white paper has primarily discussed the quality of the final outcome with an implicit assumption that such an outcome is binary in nature (e.g., should an item be removed from the corpus or not? Is the item policy-violating or not?). However, there are many scenarios in which either:

- The final outcome is not a binary decision (e.g. a video could be OK, de-monetizable, or removable);
- We have a strong reason to care about which policies the item violates - this is also not binary as there are typically many different policies a single item could violate.

The canonical quality metrics - as we've defined them - do not work cleanly outside of a situtation where the label of interest is non-binary. However, we often face scenarios such as the one illustrated in the diagram below.

In such situations, we often define two categories of metrics:

- **Decision Quality:** these are quality metrics defined in terms of the final action. I.e., they measure whether actions taken as a result of the review are the correct actions, but do not consider the *reason* for the decision.
- **Label Quality:** these are quality metrics defined in terms of the individual labels, so a decision is considered incorrect if any individual label is incorrectly applied (or not applied).

Label quality metrics themselves can either consist of a set of regular quality metrics (TPR, FPR, and PPV) for each individual policy (Policy A, Policy B, ...), or be **aggregated**. In the latter case, individual policy verdicts are aggregated together at the entity level to define the confusion matrix. In the above example, that means that the illustrated review would contribute 7 TNs, 1 TP, 1 FP, and 1 FN. TPR, FPR, and PPV can then be defined in terms of the aggregate confusion matrix. The latter can be an effective summary of the average label quality, but suffers from some drawbacks which we describe briefly below.

When is it important to measure label quality?

While it is generally interesting to understand and measure quality at the label level, it is more important in some scenarios than in others:

	Scenario	Importance of measuring label quality
1	Different labels lead to significantly different enforcement actions (e.g., different degrees of friction for- or different messaging to the user/publisher/advertiser/etc.)	High - applying the wrong label would lead to the wrong enforcement action
2	Labels are used to train classifiers at the policy-/abuse-type level	High - applying the wrong label impacts the quality of the training data negatively
3	All (positive) labels lead to the same enforcement action	Low - in this case, the specific label has no downstream impact on users or classifiers
4	All (positive) labels lead to the same enforcement action, but the specific labels are used to train classifiers at the policy-/abuse-type level	High - applying the wrong label impacts the quality of the training data negatively
5	Groups of labels lead to the same enforcement action, e.g., labels 1 and 2 lead to action A, and labels 3 and 4 lead to action B	High - although in this case it would make sense to measure quality for the two actions A and B, rather than the labels 1-4

Robust label-level quality measurement is generally harder to achieve, especially for labels which are less prevalent in the population of interest (more on this in the following). Hence an encouragement to first understand and compare the downstream implications of applying different labels before developing sophisticated sampling techniques to support the more granular label-level quality metrics.

A note on aggregate label quality metrics

While aggregate label quality metrics do provide a general sense of label-level correctness, some care is important in using them, since they are sensitive to the composition of the policy set. As a silly, but concrete example, imagine that there are two policies, blue and red, that are violated whenever the entity is blue or red, and which have a recall of 90% and 10%, respectively. If the prevalence of these two policy violations is approximately the same, then the average aggregate recall works out to 50%. Now imagine that we introduce an additional policy, azure, which is violated when the entity is azure (basically, blue). This policy is very similar to our blue policy, with similar label-specific recall, but the inclusion of this policy in our set of policies inflates aggregate recall to ~63% even though the agents are no more (or less) able to make decisions than they were previously.

Similarly, as the number of policies grows, there is a tendency for the proportion of TNs in the sample to grow disproportionally because policies aren't independently distributed and many policies rarely co-occur, resulting in a progressively decreasing FPR. This kind of behavior can lead to the misleading notion that quality is improving (or worsening) over time as new policies are added (or removed), making it hard to track these metrics over time when the set of policies is constantly evolving.