AA GTM Scaled Review Protocol

Updated 17 d View Recent Updates

Overview

This protocol is for the Access & Compromise High Touch Support (HTS) pilot, for Ground Truth Measurement (GTM) labelling. Ground Truth is a label based review process used to validate our topline metrics (‘Recovery Rate’ and ‘Bad Clear Rate’) and to provide insights into our Gaps.

Recent Updates

As of 8/7/25:

- Added new question (Q3) for comp classifier "What was the compromise classifier output?”
- Added short protocol for Q3 under ‘Compromise Classifier Guidance’
- Q4 - Q10 have been updated.

As of 7/10/25:

- Added options to a few questions
- Added multi-select to Q8
- Added options for Q8
- Added ‘NA - Recovery Not Performed’ to Q6

- Added ‘NA - No login / recovery’ to Q7
- For Q1, choose what the issue is, rather than what the customer says (updated question phrasing)

High Level Process

1. **Compromise Review:** Review the user account for account compromise, this will also help determine other answers in the review.
2. **HTS Job Review**: Review the original case to understand the actions taken, the issue at hand, and who contacted support.
3. **Outcome Assessment:** Combining entry point signals with Expert Review we will determine if the correct outcome was reached.
4. **Gap Assessment:** When we do not reach the correct outcome we will perform a gap assessment to identify which area(s) were ineffective, broken, or incorrect. These insights will be used to inform future builds/iterations/trainings of our systems, processes, and reviewers.

Intake Process

Task Folder

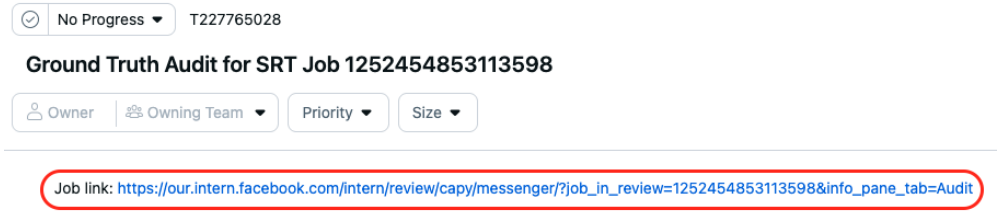
1. Open the task folder and select task.
2. Claim task to assign to your name.

3. Change status to In **Progress**.
4. Open the AGC job from link in task description.
5. Complete the AGC job (this will automatically close the task).

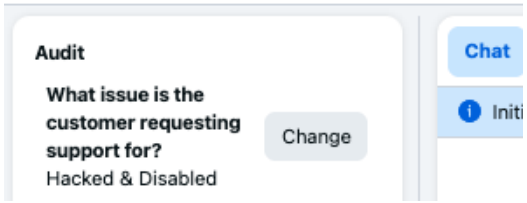
Note: Do not pick up a task that has already been assigned to another person, check if the task has already been reviewed previously or is already closed.

Agent Connect Audit

1. Open the job link from the task.

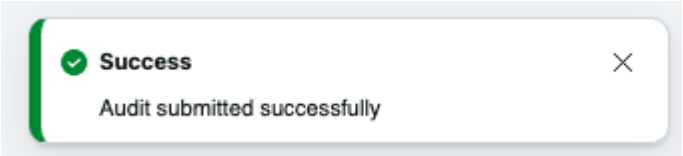


1. Complete the Audit Questions from the **Audit** widget (top left-hand menu).



1. Add comments and any issues you saw in the **Notes** field.
    a. If you do not have any comments, enter “NA - No Comments”.

1. One your notes are complete, press **Submit**.
2. You will see a pop-up that the feedback has been received on the bottom right of your screen (see image below).



1. After the audit is submitted, close out of the AGC job window.
2. The task will close automatically.

Warning: Do not change anything else on the AGC job itself (ex. do not change job status in AGC, do not change or add notes to the activity log or AGC wizard).

Ground Truth Measurement (GTM) Protocol

For the full question list, see the **UDT** section. Below are the questions that require more information.

For other protocols and additional resources, review the **Materials** tab of [AC] GTM Scaled Review Training - Access.

Compromise Investigation

Review the user account that is writing into support for compromise. Make sure to take note of attributes (ex. Device, location, etc.) of the real user vs the hacker for questions that come later in the process.

- Facebook Investigation Playbook: https://www.internalfb.com/wiki/Compromised_Accounts_Investigations_Playbook/FB_Compromised_Investigations_Playbook_v4/
- Instagram Investigation Playbook: https://www.internalfb.com/wiki/Compromised_Accounts_Investigations_Playbook/IG_Compromised_Investigation/

High Touch Support Overview

Review the HTS job and the outcomes using the appropriate knowledge base:

- Account Access: https://fburl.com/gms/2qh6jh78
- UFAC: https://fburl.com/gms/9wrlgbfa

Q3. Compromise Classifier

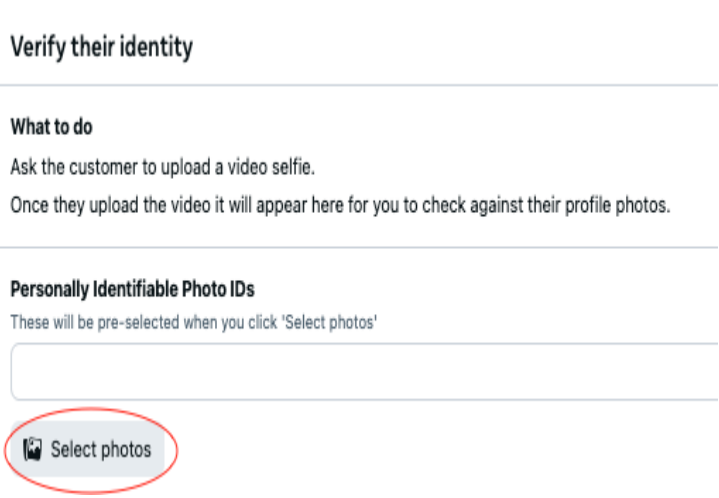
Be sure to review the account for compromise before looking at the classifier output. Do not rely on the compromise classifier to make your compromise decision, we need these to be separate.

- The Compromise Classifier widget is located on the righthand AGC Wizard Panel
- The Classifier results can also be seen in the ‘Completed Summary’ section
- General guidance for compromise review period is 30-90 days from time of case submission

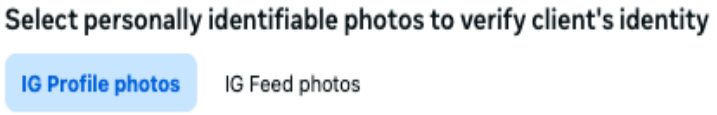
Q4. Submitter Identification Process

Use the following signals to identify if the submitter of the case was the real account owner, or a malicious actor. Leverage a combination of the signals below to make your decision.

Signal	Signal Strength	Steps to Review

Authentication Process	High	• Review the selfie submission by the user for authentication to the profile photos on the user account
		  ◦ Check the selfie submission from the user in the Chat or Email thread OR check the Auth Viz tool linked in the Agent Wizard.
		  ◦ Click through the Agent Wizard on the righthand side to get to the **Verify the Identity** page.
		 

a. Review the Personally Identifiable Photos (PiP) on the two tabs.



- Compare to the Authentication photos for likeness (check that the person in the photo matches the user). Check the hair, facial expression, facial attributes, etc.

Important: Be aware of recent trends where bad actors are.

- Submitting selfies that will populate in the PiP section. Be conscious of the dates the photos were uploaded, and whether those overlap with the compromise timeline.
- Submitting deepfake selfie videos (ex. j616280350799260)

Case Creation Details	High	Review the information from the user that submitted the case using GTM Reviewing Case Creation. • Locate the information from the real user (DATR, IP, Device, Location, Browser).

		◦ Check the activity log for before the compromise OR if not compromised, check for regular user behavior. • Locate the information from the case submitter (DATR, IP, Device, Location, Browser). ◦ Check against the **Requesting User Details** widget in AGC. ◦ Check against the **Support Case Submitted** event in the Activity Log. • Locate the information from the hacker (if compromised) ◦ Leverage the time of compromise from the Compromise Labeller. ◦ Compare the information to see which user submitted the case. **Note:** For case abandoned or limited visibility, use **Case Creation Details** only.
Post Recovery Behavior	Mid	Does the activity look suspicious after recovery was complete (ex. Linking or Unlinking Accounts, Suspicious posts, Contact Point updates, etc.)? • Example: did the user post a suspicious or sketchy looking post?

		• Example: did the user continue similar compromise actions that were seen before recovery?
Conversational Indicators	**Low**	Sometimes bad actors will provide suspicious flags in the conversations themselves, to indicate they may not be the real owner of the account. Check for: • Short and direct messages / responses • Refusal to submit authentication methods • Knowledge of the process (ex. knowing authentication step is coming) • Suspicious wording indicators
User Story + Account Activity	**Low**	Does the user stated story / timeline match up with what we are seeing on the backend for compromise dates? • If the account has been compromised, there is a higher chance the hacker is writing into support to regain access to the account. If there are no compromise indicators, it is more likely that the submitter is the real user. ◦ Check compromise labeller score. ◦ Review the **Ground Truth Measurement Protocol** section above. • Is the user saying they are having trouble logging in, and we can see in the activity log

		that the timeframe and login issues match? ◦ Check the **Activity Log**.

Q7. Successful Login

Review the activity timeline to determine if the real account owner was able to log back in after support was provided.

1. Check the real owner information, use the info from Step #2 to find the real owner in the activity log
2. Check when support was provided, use the activity only after “Support Case Created” or “Agent Connect Job Created” happened (see step #2 for more details)
3. Confirm if the real user (from step #2) successfully logged back into the account (ex. login, session update, post, contact point update, etc.)

You can also get a hint from the chat thread or email (within AGC) to see if the user mentions they logged back in successfully (be sure to confirm this in Activity Log).

GTM Audit Responses (Access)

TITLE	QUESTION PROMPT	RESPONSE OPTIONS	DETAIL	NEXT QUESTION	RESOURCE	MULTIPLE SELECT
1 - issue_diagnosis	*What issue is the customer facing?*	Hacked / Compromise	Hacked or Compromise issue	**account_compromise**	In scope for HTS	No

		(This is not what the customer is stating, but rather what the real issue is.)	Hacked and Disabled	Hacked account that is also disabled		
			Login - Reset Password	Password reset or forgot password		
			Login - Two Factor	Two factor issue		
			Login - Login Challenge	Login Challenge checkpoint		
			Login - Lost Credentials	The user lost their credentials and cannot login		
			Not enough signal to evaluate	The user has not provided enough information to review the job	**notes**	Out of scope for HTS	No

	(ex. Immediate abandon)
Unsupported - User requesting support for another account	User is requesting support for an account that they do not own OR requesting support for multiple accounts at once
Unsupported - User requesting support for out of scope asset	Issue unrelated to a user account (ex. Page, Business Manager, Group, etc.)
Unsupported - Non-Access Issue	User is reaching out about a non-access related issue (ex.

			account strikes, ban, payment)			
		Unsupported - Account Deletion	User is requesting an account or data deletion			
		Unsupported - Language Issue	The user is writing in using an unsupported language (not English)			
		Unsupported - Other	Unsupported use case, or an issue that is Out of Scope for this flow. *Note the new issue type in the ‘Notes’ section below			

2 - account_compromise	*Was the account compromised?* *(At the time of review)*	Yes	Yes the user account was compromised / hacked	**comp_classifier**	Compromise Review Protocols	No
		No	No the user account was not compromised			
3 - comp_classifier	*What was the compromise classifier output?*	Compromised	The classifier states the account was compromised	**support_requestor**	Compromise Classifier Guidance	No
		Not Compromised	The classifier states the account was not compromised			
		Unavailable	The classifier is unavailable for the case			

4 - support_requestor	*Who contacted support?*	Account Owner	The user that wrote into support is the real account owner (see protocol)	**requestor_confidence_signal**	Submitter Identification Process	No
		Malicious Actor	The user that wrote into support is not the real account owner (malicious actor)			
		Other	Relative / Friend of the user (note, check for bad actor). This should not be used regularly			

	5 - requestor_confidence_signal	*What signals did you use to make your determination?*	Authentication Methods	Using authentication methods (ex. Selfie, auth viz, profile photos)	**recovery_decision**	Submitter Identification Process	Yes
			Case Creation Details	Using the case creation details to match the users (see protocol)			
			Post Recovery Behavior	Using the behavior after the user recovered (see protocol)			
			User Story / Account Activity	Using the user story, timeline, and account activity (see protocol)			

		Conversational Indicators	Using the user conversation details			
6 - recovery_decision	*Were the HTS recovery steps performed?*	Yes	The HTS reviewer performed the recovery steps (ex. SUA, PRL, 2FA)	**successful_login**	WS2 [HTS GenPop] - Protocol + Content Design	No
		No	The HTS reviewer did not perform any recovery actions (ex. Access denied)			

7 - successful_login	*Did the requestor successfully log back in after support?*	Yes	The user successfully logged back in after receiving support (confirm with protocol), at the time of review	**post_login_experience**	Successful Login	No
		No	The user has not logged back into the account after receiving support, at the time of review (confirm with protocol)			
		NA - Recovery Not Performed	The recovery steps / support were not performed for this case			

	8 - post_login_experience	*After login, did the requestor experience something else that may have prevented recovery?*	Enrolled in a Checkpoint / Experience	The user is enrolled into a checkpoint after recovery (ex. FB Protect, 2FA, ORCA)	**issue_type**	[AC] HTS GTM Scaled Review Protocol	Yes
			Incomplete Recovery	The user remains blocked due to a related issue (ex. business asset impacted which got the acct taken down)			
			Re-Compromise	The account was re-compromised by the hacker after support was received from HTS			

		Out of Scope for HTS	The user is experiencing an error from a different violation that is unrelated to Access (ex. VAN disable, Content Violation, etc.)				
		No - No issues	The user did not see any issues after login				
		NA - No login / recovery	The user did not login and/or did not receive support				
		Other	The user experienced an issue not listed above, please				

				specify in the notes section			
	9 - issue_type	*Why was the requestor unable to log back in?*	Quality: Incorrect resolution (Agent)	The issue was not solved correctly by the reviewer (ex. the incorrect action was taken)	**notes**	WS2 [HTS GenPop] - Protocol + Content Design	Yes
			Quality: Incorrect resolution (AI)	The issue was not solved correctly by the AI bot			
			Quality: Incorrect protocol applied (Agent)	The HTS reviewer used the incorrect protocol / guideline to address the problem			
			Quality: Incorrect	The AI used the wrong protocol / issue			

			protocol applied (AI)	diagnoses to address the problem			
			User is experiencing a bug	The user is getting a bug/error causing them to not be able to recover their account (ex. PRL was sent but the user cannot open the link)		[AC] HTS - Bug Reporting & Management Process	
			User is experiencing a checkpoint	The user is stuck in a post recovery checkpoint and cannot log back in (ex. FB Protect)			
			Process/documentation is	The protocol does not have an entry for the		WS2 [HTS GenPop] -	

missing or incorrect	issue stated, or, the current information is outdated or is incorrect *Please clarify the issue in the 'Notes' section for what is missing/incorrect
Tooling Issue (reviewer)	There were limitations in the tooling available to the HTS reviewer that resulted in the user not regaining access (ex. SUA action did not go through, no tool available, etc.)

Protocol + Content Design

User abandoned - long wait time (Agent)	The user left or stopped responding due to long wait time from the HTS reviewer
User abandoned - long wait time (AI)	The user left or stopped responding due to long wait time from the AI Bot
User abandoned - authentication error or refusal	The user refused or was unable to submit authentication methods

User abandoned - more time to complete the review	The user specifies that they need more time to complete their task or cannot complete authentication at the time (ex. need help submitting authentication)

				User abandoned - resolved elsewhere	The user leaves the support thread because they were able to self resolve or resolve in another channel (you can see this from the user stating it, or checking activity log)						
				User abandoned - duplicate case	The user abandoned the case because they submitted another case for the same issue						
				User abandoned - unknown	The user leaves the chat for an unknown reason and/or						

			leaves the chat soon after chat creation				
		User was a malicious actor and access was denied	The user that wrote into support was a bad actor, and access was denied correctly (see Question 3)		Submitter Identification Process		
		User did not pass authentication	The user uploaded authentication, but did not pass the authentication step				
		Recovery action not	The user was unable to log				

Choose a ticket