**Vision Agent Quality Assurance Report - Release Process Summary - Week 18**

**Release Testing Completion Date:** May 7th, 2025

**Release Start Date:** Apr 30, 2025

**Environment:** Staging

**Deployed Services:**

* Vision-agent-iu
* Vision-tools-rest-api
* mi-models-server

**Environment details and testing scope:** This was a release whose main scope was to test two new features which are listed below.

**New Features/Improvements:**

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **DRI** | **Tester** |
| **Light Theme** | [Mingrui Zhang](mailto:mingrui@landing.ai) | [Daniel Giraldo](mailto:daniel.giraldo.external@landing.ai) [Luisa Aristizabal](mailto:luisa.aristizabal.external@landing.ai) [Victor Diez](mailto:victor.diez.external@landing.ai) [Kelly Castrillón](mailto:kelly.castrillon.external@landing.ai) |
| **Page Limit Playground and API Request** | [Mingrui Zhang](mailto:mingrui@landing.ai) [Camilo Zapata](mailto:camilo@landing.ai) | [Kelly Castrillón](mailto:kelly.castrillon.external@landing.ai) [Luisa Aristizabal](mailto:luisa.aristizabal.external@landing.ai) [Daniel Giraldo](mailto:daniel.giraldo.external@landing.ai) |

For further details on how we tested them, you may refer to [the following spreadsheet](https://docs.google.com/document/d/1BudchpzWsdsRaPBmAHHUl47hnwEfNe3TfsvHbZwfDy4/edit?tab=t.0).

**Testing Methodology:**

During this release process the testing cycle comprised the following validations:

* Automated test cases to check the stability of the VA tools and run the RL. These tests comprise the following elements:
  + Backend:
    - 1 request to each one of the tool’s APIs
    - RL tests with 320 responses based on [Junjie Guan](mailto:junjie.guan@landing.ai)’s test which can be reviewed [here](https://github.com/landing-ai/vision-agent-ui/tree/main/tests/ratelimit). During this release these tools were covered by these tests.
      * AOD
      * ADE
      * SAM2
    - RL and 500 status requests with parallel requests with a good image which can be reviewed [here](https://drive.google.com/drive/folders/12sr-U6lIXqzt-xfymK4yftROQaZE1inI?usp=drive_link). During this release these tools were covered by these tests.
      * AOD
      * ADE
  + Automated UI validations:
    - Check that the URL in the API request example code of each tool points to the right environment, in this case STG.
    - Check that the requests through the playground with example images in AOD finish successfully.
    - Check that the chats in the VA finish successfully.
* Manual testing. These manual tests aim to reinforce the results of the automated tests and double-check the stability of the VA platform. This smoke test consist of:
  + Smoke test of some of the VA tools.
    - Smoke test of all tools with the playground and an API request.
    - Smoke test of the vision agent in the chat, test the examples and additionally 5 more prompts.
  + Smoke test of ADE tool with the playground and one API request.
  + Smoke test of AOD tool with the playground and one API request.
  + Validation of the page limit according to [this matrix](https://docs.google.com/spreadsheets/d/1aPlZSwMtyJixnl4bc0icjkyLplTe79k3A_nryH5H-c4/edit?gid=702435464" \l "gid=702435464), which is expected in each of the tiers.

Additionally, more focused tests were conducted on the recent **Playground updates for Agentic Document Extraction (ADE)**:

**URL:** [ADE Playground](https://va.stg.landing.ai/demo/doc-extraction)**Playground Tests:**

* Upload an image and verify if markdown and JSON outputs are correctly generated.
* Upload a 50 page PDF and verify markdown and JSON visibility.
* Validate that every example has corresponding markdown and JSON outputs.
* Verify the responsive design across devices (desktop, tablet, mobile).
* Measure latency across different file sizes.
* Confirm that usage is correctly charged when using the playground.

**Chat with Document Tests:**

* Verify the generation of relevant questions.
* Validate that answers in the chat are properly generated and include clickable references.
* Test the chat functionality using an image, a 2-page document, and a 50 page document.
* Measure latency during interactions.

**API & Python Library Tests:**

* Test input methods using direct file upload (image, 2-page PDF, 20-page PDF, 21-page PDF, 50-page PDF, 51-page PDF, 100-page PDF in tier 3), folder input, and remote URLs.
* Verify functional features:
  + Marginal chunk removal (page headers and footers).
  + Visual grounding functionality, ensuring correct operation of crop-based visual chunks.
* **Light Theme Validation:**
  + Ensure that all pages apply correct light theme styles.
  + Confirm that the light theme is the default setting and no dark mode elements appear when in light mode.

**Results:**

1. **Automated tests**
   1. **API’s test cases tools automated:**

|  |  |  |
| --- | --- | --- |
| **Tool Name** | **Status** | **Note** |
| Owlv2 | Pass |  |
| Countgd | Pass |  |
| Florencev2\_Roberta | Pass |  |
| Florencev2\_OCR | Pass |  |
| Florencev2QA | Pass |  |
| Depth\_Anything\_V2 | Pass |  |
| Text\_To\_Od | Pass |  |
| nsfw\_classification | Pass |  |
| Wsi\_Embedding | Pass |  |
| qr\_reader | Pass |  |
| loca | Pass |  |
| Pose\_Detection | Pass |  |
| Barcode\_Reader | Pass |  |
| Agentic OD | Pass |  |
| Agentic DE | Pass |  |

1. **RL tests:**
   1. The RL tests during the current VA release have been completed successfully. After sending 100 parallel requests in each tier with a good image for AOD and ADE, no 500 errors were generated. Additionally, the number of 200 and 429 responses matched the expected amounts according to the RL settings for each tier.
2. We also executed the RL tests from the vision-agent-ui repository for ADE, AOD, and SAM2 tools, and all results were aligned with the RL established for each tier. The actual request rates observed were as follows:

* **ADE Tier 1:** 12.88 requests/minute
* **ADE Tier 2:** 32.03 requests/minute
* **ADE Tier 3:** 76.03 requests/minute
* **AOD:** 12.85 requests/minute
* **SAM2:** 106.72 requests/minute

No unexpected behavior was observed, and the tests confirmed that the system behaves according to the configured RL limits.

1. Automated UI Validations
   1. The automated tests validating the VA chats stability passed successfully in this release. No bugs were detected, and all validations completed as expected without issues.
2. **ADE-Python library:**

* Direct file upload
* Folder input
* Remote URLs

1. **Light Theme:**

The Light Theme was correctly implemented across all pages. It was verified that the light theme is now the default setting, with no dark mode elements appearing when active. Automated UI validations passed successfully, and manual tests confirmed consistent styling and behavior.

1. **Playground and API (Page Limit feature):**

The page limit functionality was tested across different tiers through the Playground and API. Uploads of documents of various sizes were validated, confirming correct markdown and JSON outputs, responsive design, and correct latency handling. [in this sheet are the results we obtained from the latencies in each tier.](https://docs.google.com/spreadsheets/d/1slPpaYyh78OLPIyJ1RzZ6Xv5MjJYifQrMmDv1cAiU64/edit?gid=0" \l "gid=0)

1. **Bugs Identified:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Asana Ticket** | **Priority** | **Team** | **Status** |
| [✓ ADE is priced at 0.3 instead of 0.03](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210138875387863?focus=true) | P0 | DM | Fixed |
| [✓ Error handling for >50 pages on playground](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210136439857659?focus=true) | P0 | DM | Fixed |
| [✓ Chat Visual Reference Scroll](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210150967263641?focus=true) | P0 | DM | Fixed |
| [✓ Visual refrence shifted for this pdf](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210150968499827?focus=true) | P0 | DM | Fixed |
| [✓ [STG][Vision Agent][ADE] The preview of the files is not visible when the file contains more than 50 pages](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210137422878777?focus=true) | P0 | DM | Fixed |
| [✓ [STG][ADE] Intermittently, but still frequently enough to impact customer's experience, ADE is returning requests with "Internal server error" message.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210144871380573?focus=true) | P0 | MI | Fixed |
| [✓ Uploading new files UI experience](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210165073558023?focus=true) | P1 | DM | Fixed |
| [✓ Light theme should be default](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210150967263653?focus=true) | P1 | DM | Fixed |
| [✓ Feedback button text update](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210150967263645?focus=true) | P1 | DM | Fixed |
| [✓ [Duplicate] [STG][Vision Agent][ADE]In the Playground environment, the page limit for Tier 0 and Tier 1 accounts is set at 50. However, when accessing through the API, different limits apply: 2 pages for Tier 0 accounts and 20 pages for Tier 1, which aligns with what is expected for these tiers.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210145109102197?focus=true) | P1 | DM | Fixed |
| [✓ [STG][Vision Agent][ADE]Page limit in T1 is not working as expected for requests sent through the API. The limit is still 20 pages instead of 50.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210177813085094?focus=true) | P1 | MI | Fixed |
| [✓ [STG][Vision Agent][Mobile Responsive][AOD]When using the AOD tool on a mobile device, the user is unable to operate it because the "Run" button is not displayed. Additionally, the section where the results are shown appears broken or visually incorrect, making the tool almost impossible to use from a mobile device.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210150632356504?focus=true) | P1 | DM | Closed |
| [✓ Decimal Points](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210167058809317?focus=true) | P2 | DM | Fixed |
| [Generate example prompts for the chat when extraction generated](https://app.asana.com/1/504311096896991/project/1208384446321539/task/1210165073558027?focus=true) | P2 | DM | PR |
| [[STG][Vision Agent][AOD] The requests made through the Playground in the Agentic Object Detection tool are not generating any cost.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210151327697462?focus=true) | P2 | DM | Open |
| [[STG][Vision Agent][AOD] The requests made in the AOD tool examples are listed in the Detailed Logs section of the Usage page, even though these requests do not incur any cost.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210151327697456?focus=true) | P2 | DM | Open |
| [✓ [Duplicate] [STG][Vision Agent] The cost of processed images is displayed on the usage page using exact decimal values that are not rounded. This causes the value in the “Total cost” section to appear outside the visible area or exceed the allocated space.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210146798916027?focus=true) | P2 | DM | Open |
| [[STG][Vision Agent] Requests made with jpeg and png images are not being recorded in the Detailed Logs table in the Usage and Cost page.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210151247788914?focus=true) | P2 | DM | Open |
| [[STG][VA] the page of TPA tools in the VA seem to keep dark theme elements after choosing the light theme.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210119775318737?focus=true) | P2 | DM | Open |
| [[STG][Vision Agent][Mobile Responsive] Previous chats are hidden after starting a new conversation in Vision Agent](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210150627236944?focus=true) | P2 | DM | Open |
| [[STG][Vision Agent][ADE]Quite frequently the ADE is counting and charging the amount of pages as though the document had 1 page less (e.g a document of 50 pages is counted and charged as if it had 49 pages).](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210177813085098?focus=true) | P2 | DM | Open |
| [✓ [STG][Vision Agent] On accounts without available credits, attempting to make a request via the Playground or the API returns the following error message: "Error: Payment Required. User balance is insufficient." This happens even though not all tools consume credits. However, the error appears consistently across all tools.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210145109102209?focus=true) | P2 | DM | Closed |

**Release to PROD:**

During sanity tests in PROD, it was found that pages were not being counted or charged for requests in both ADE and AOD.

**Actions Taken:**

* [Camilo Zapata](mailto:camilo@landing.ai) released vision-tools-rest-api and mi-models-server.
* The issue was resolved for ADE, but AOD was still not charging.

**Additional Patch (by** [**Mingrui Zhang**](mailto:mingrui@landing.ai)**):**

* A small production patch (v2025.05.08.01) was deployed, fixing several issues:
* Fixed AOD not charging credits.
* UI improvements: added scrollbar for document display and removed zoom tool.
* All tools now call va-rest-api-tool directly through proxy, instead of going through the Playground backend.
* AVIF and HEIC image formats are now unsupported.
* Set a rate limit of 4 requests per minute for non-agentic tool endpoints (to prevent abuse, like text-to-object-detection spam).
* Changed prompt generation: example prompts are now generated after document extraction instead of during the chat.

**Sanity Check Results after the Patch:**

* **P1 issue (High priority):**

The problem where AOD requests were not being charged was fixed.

* **Other findings (P2 - Medium priority):**

IMPORTANT [[STG-PROD][Vision Agent][AOD-ADE]Inferences with .TIFF, .BMP and GIFF images are not allowed through the UI, but are allowed and charged via API request.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1209929916081164?focus=true)

IMPORTANT [[STG][Vision Agent][AOD-ADE] Responses for requests made with HEIC and AVIF files through the API are not quite informative as to indicate that the reason of failure was an unsupported file format. In AOD user receives a 500 status response, whereas in ADE, user receives a 200 request status response with a 500 error message within the content of the response.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210191008203494?focus=true)

IMPORTANT [[STG-PROD][Vision Agent][AOD] [Feedback] UI of AOD doesn't let user know the supported image/file formats, as it happens in ADE](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210191008203473?focus=true)

[[STG][VA][ADE]Small imgs, some times don't get charged or listed after being processed either via Playground or via API, especially when they are sent through parallel requests.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210191008203450?focus=true)

[[PROD][VA] We are getting this error in the vision agent chat “KeyError: ‘data’ enabled in stella\_embeddings() due to missing key in the embedding service response”.](https://app.asana.com/1/504311096896991/project/1208747057905216/task/1210199638611778?focus=true)

**Good news:**

* The rate limit for non-agentic tools was successfully reduced to 4 requests per minute.
* The rate limits for ADE and AOD were also confirmed to be working as expected.

**Conclusion:**

During this release testing cycle on the Staging environment, two main new features were thoroughly validated: the implementation of the Light Theme and the Page Limit control for the Playground and API Request. A variety of testing methodologies were applied, including automated backend tests, UI validations, load (RL) testing, and manual smoke testing.

**The overall results were positive:**

* **Automated API and UI Tests:** All automated test cases for the Vision Agent tools passed successfully, confirming the stability and proper functionality of backend and frontend services.
* **RL Tests:** Rate limit (RL) tests for AOD, ADE, and SAM2 tools were completed successfully, with no 500 errors observed and expected request rates achieved for each tier.
* **Manual and Smoke Testing:** Manual validations complemented the automated tests, confirming the correct implementation of the new features and the overall stability of the system.
* **Playground Focused Testing:** Specific tests for the updated ADE Playground, page limit, output generation (Markdown and JSON), latency measurements, responsive design checks, and usage charging validations, were successfully completed.
* **Bug Management:** Several critical (P0 and P1) bugs were identified and fixed, mainly related to error handling, large document visualization, mobile responsiveness, and pricing errors. Some lower-priority (P2) issues remain open and are under monitoring.

Additionally, during the final sanity checks after the latest production patch:

* A P1 issue remains in STG: AOD requests are charged but not recorded on the Usage and Cost page.
* Several P2 findings were noted, including handling of unsupported image formats, API/UI discrepancies, small image charging inconsistencies, and an isolated KeyError observed in PROD.
* The rate limit adjustment to 4 requests per minute for non-agentic tools was successfully verified.

In summary, this release can be considered successful, as all major validations passed as expected, critical defects were resolved before the end of the testing cycle, and the new features meet the established requirements, with minor issues already under tracking.

**Report Prepared By:**

[Luisa Aristizabal](mailto:luisa.aristizabal.external@landing.ai)

**Date of Preparation:**

May 7, 2025