**SOF/C Tactical OPE Toolkit: Custom Error Metric Development**

# **Purpose**

The purpose of this document is to define and justify custom error metrics tailored to the Special Operations Forces / Cyber (SOF/C) Operational Preparation of the Environment (OPE) Toolkit. These metrics provide a more precise evaluation of the system’s core capabilities—specifically, its automated review summarization and image captioning functions—within the unique operational context of Special Operations Forces (SOF). Traditional metrics like accuracy or F1-score offer general benchmarks but fall short in capturing mission-relevant performance. Our custom metrics are designed to reflect real-world usability, robustness, and actionable intelligence, aligning directly with the operational goals and constraints of SOF analysts.

# **Objectives and Key Results**

The primary objective of the SOF/C Tactical OPE Toolkit is to dramatically reduce data processing time—from the traditional 8 to 40 hours per query to under 2 minutes per location—while maintaining high accuracy and usability. To achieve this, the following key results have been prioritized:

* Seamless end-to-end data processing across all tiers (basic, reviews, and photos)
* High-quality review summarization powered by GPT-4o-Mini
* Precise and context-aware image captioning and tagging using GPT-4o Turbo
* Timely export of outputs in KMZ, Excel, and JSON formats, meeting defined performance thresholds based off the customers needs
* Robust error handling and an intuitive user interface requiring no programming expertise

# **Limitations of Standard Metrics**

Standard evaluation metrics such as Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) are insufficient for accurately measuring the effectiveness of the SOF/C Tactical OPE Toolkit's language processing components. While these metrics are commonly used to evaluate the similarity between generated and reference text, they rely heavily on surface-level word overlap and fail to capture the contextual relevance and semantic accuracy required in summarizing real-world, often noisy, user-generated content such as Google Maps reviews. This limitation is especially evident when assessing GPT-4o-Mini, which is optimized for generating concise, context-aware summaries that may differ lexically from reference outputs while still being operationally accurate.

Similarly, traditional metrics like precision, recall, and accuracy fall short when evaluating the output of GPT-4 Turbo, the Vision-Language Model (VLM) used for image captioning and tagging. Visual scenes can be described in multiple valid ways depending on operational context or the user’s prompt. For instance, a street view image tagged as “entrance with security gate” could also reasonably be described as “main access point with barrier.” Standard classification-based metrics would penalize such variation, even though both descriptions may be correct and actionable.

Furthermore, latency metrics that simply measure average response time do not distinguish between user interface delays and actual model inference time—an important distinction when validating system responsiveness in time-sensitive environments. This becomes especially relevant when processing large image batches through the VLM, where asynchronous calls, API rate limits, and token constraints can introduce variability.

Standard metrics also overlook operational constraints unique to our deployment scenario, such as compliance with OpenAI’s content moderation policies and efficient use of token budgets. A failure due to flagged content or token overuse can render an otherwise correct output unusable in the field. Therefore, a more tailored set of metrics is necessary—ones that assess not only correctness and completeness but also robustness, interpretability, and compliance within mission-critical constraints.

# **Custom Metric Ideas and Definitions**

The following custom metrics were developed based on empirical findings from structured end-to-end testing (Test Cases 1–7b, see Appendix for detailed test cases) and iterative feedback from stakeholders. Each metric addresses a specific operational concern identified during test execution, and was refined through real error handling, result inspection, and cross-functional review.

* **Summarization Validity Rate (SVR):** This metric measures the percentage of review summaries that are both structurally complete and free of OpenAI content moderation violations. It was defined following issues encountered in Test Case 6a, where GPT-4o-Mini flagged a review for hate speech and initially returned a failure. The backend was modified to handle such errors gracefully, returning a placeholder message instead of failing outright. SVR reflects the system’s ability to reliably generate mission-suitable summaries under content constraints.
* **Image Prompt Response Accuracy (IPRA):** IPRA evaluates how effectively the Vision-Language Model (GPT-4 Turbo) reflects user-specified prompt keywords in its image captions. This was observed during Test Cases 4 and 5. In the absence of prompt input (Test Case 4), image tags were generic and vague, prompting the implementation of default-safe prompt handling. With keyword input (Test Case 5, e.g., “doors, windows”), we validated that the tags and summaries explicitly referenced the specified elements. IPRA quantifies this alignment.
* **Export Success Rate (ESR):** This metric tracks the success rate of KMZ, JSON, and Excel file generation. Initially, in Test Cases 1a and 1b, missing keys and improperly handled JSON fields (e.g., missing 'street\_view') caused export failures or incorrect formatting. After implementing conditional logic and tier-aware output formatting, all test cases from 2 onward successfully produced both file types. ESR reached 100% in the final validated tests.
* **Conditional Tier Handling (CTH):** CTH is a boolean indicator used to verify that missing or irrelevant API keys do not break application flow. During early testing (e.g., Test Cases 1a, 1b, and 5), unhandled logic caused the frontend to crash when the backend omitted expected fields due to unselected tiers. After identifying these issues, logic was added to dynamically generate Excel fields based on active data tiers, resolving the problem entirely. CTH ensures resilience across partial configurations.
* **Latency Compliance (LC):** LC represents the percentage of locations fully processed within the 2-minute per location performance requirement defined by the customer. This was tracked qualitatively across all test cases. Even for large bounding boxes (e.g., Test Case 2 with 60+ locations), response time remained within acceptable limits after introducing asynchronous VLM processing and token budget optimization. This metric helps validate scalability and field-readiness.

These metrics were not only theoretically defined—they were tested, iterated, and implemented in response to observed failures and customer priorities. Together, they form a tailored performance assessment framework more aligned with SOF operational needs than conventional metrics like BLEU or accuracy.

# **Metric Implementation and Testing**

The custom metrics defined for the SOF/C Tactical OPE Toolkit were not only theoretically designed but also validated through practical testing. Over the course of 27 structured end-to-end test cases, each metric was evaluated in real operational scenarios, capturing edge cases, user errors, and system limitations. The following summarizes how each metric was tested, what issues were encountered, and how they were resolved:

* **Summarization Validity Rate (SVR):**Validated during review summarization tests. Instances where OpenAI’s content moderation flagged input (e.g., hate speech) were successfully handled by inserting placeholder messages in the output. This ensured summaries remained complete and policy-compliant.
* **Image Prompt Response Accuracy (IPRA):**Confirmed through tests with and without user-supplied prompts. Test Case 5 (“doors, windows”) demonstrated that image captions aligned with prompt terms, while default-safe behavior was verified in Test Case 4 when prompts were omitted.
* **Export Success Rate (ESR):**Assessed across 27 end-to-end tests. Initial issues with missing or incompatible data structures (e.g., absent street\_view keys) were resolved through frontend or backend corrections. All exports completed successfully post-fix.
* **Conditional Tier Handling (CTH):**Triggered by failures in early test cases where unselected tiers caused missing keys in the backend response. Resolved by implementing logic that dynamically checks which data tiers were selected before generating output files.
* **Latency Compliance (LC):**Monitored qualitatively across test runs, including scenarios with large bounding boxes (e.g., 60+ establishments). In all cases, processing time remained under the 2-minute per location requirement, confirming the system’s readiness for field use.

# **Comparison with Standard Metrics**

To supplement our custom metrics, we also tracked standard indicators such as average response latency per location and the integrity of exported files, serving as proxies for traditional latency and output completeness metrics. Review summaries were manually validated to assess clarity, coherence, and relevance, complementing what standard NLP metrics alone could not capture. For image captions, we evaluated the inclusion of user-specified prompt keywords to verify alignment with task intent. These practices ensured our results remained interpretable and meaningful to both technical stakeholders and operational end users.

**Reflection on Custom Metrics**

Custom error metrics were defined from the outset to meet the specific operational needs of the SOF/C OPE Toolkit, as conventional metrics like F1-score or accuracy were inadequate for the task. Our use case—summarizing unpredictable, user-generated content and interpreting visual data using AI—demanded evaluation criteria aligned with real-world functionality rather than abstract statistical benchmarks. These tailored metrics allowed us to capture critical aspects such as content validity, prompt alignment, and export reliability, which standard metrics could not meaningfully reflect. However, a key drawback of custom metrics is their limited comparability to other systems or benchmarks, which can make external validation or generalization more difficult. Additionally, some metrics, such as image prompt alignment, required subjective interpretation, introducing potential variability without consistent reviewer guidelines. Despite these limitations, our approach enabled a more accurate and mission-relevant assessment of system performance.

### **Glossary**

* **BLEU**: Bilingual Evaluation Understudy – a metric for evaluating the precision of machine-generated text compared to a reference.
* **CTH**: Conditional Tier Handling – a custom metric indicating whether the system properly adapts output formatting based on selected data tiers and available API keys.
* **ESR**: Export Success Rate – the percentage of successful Excel/KMZ file exports relative to total export attempts.
* **F1-Score**: A standard metric that combines precision and recall, commonly used in classification tasks.
* **GPT-4o / GPT-4 Turbo**: OpenAI’s multimodal models used for language and vision-language processing in the toolkit.
* **IPRA**: Image Prompt Response Accuracy – a custom metric evaluating alignment between user-specified prompts and generated image captions.
* **JSON**: JavaScript Object Notation – a lightweight data-interchange format used for structured data exchange.
* **KMZ/KML**: Keyhole Markup Language (zipped/unzipped) – formats used to display geographic data in Earth browsers.
* **LC**: Latency Compliance – percentage of queries processed under the operational 2-minute per location time requirement.
* **LLM**: Large Language Model – an AI model trained to generate and summarize natural language.
* **ROUGE**: Recall-Oriented Understudy for Gisting Evaluation – a standard metric that measures overlap between generated and reference summaries.
* **SOF/C**: Special Operations Forces / Cyber – referring to units engaged in advanced information and operational preparation.
* **SVR**: Summarization Validity Rate – a custom metric measuring whether generated summaries are complete and policy-compliant.
* **VLM**: Vision-Language Model – AI models capable of interpreting and generating text descriptions based on images.

### **Appendix: Test Case Summaries**

| **Test Case** | **Description** | **Outcome** | **Notes** |
| --- | --- | --- | --- |
| 1a | Basic Tier only, no LLM/VLM keys | ❌ Initial failure (401 error) | Fixed by adding key conditionals |
| 1b | Basic Tier Excel export | ❌ KeyError on missing street\_view | Resolved by conditional field logic |
| 2 | All tiers, no prompt | ✅ Success | Generic VLM prompt used; logic refined |
| 3 | All tiers, with prompt “doors, windows” | ✅ Success | Prompt successfully influenced captions |
| 4 | Valid query with no results | ✅ Success | Added user alert for empty result sets |
| 5 | Basic + Photos Tier | ⚠️ Excel had empty review sheets | Fixed by removing unused tier data |
| 6a | Reviews with hate content | ❌ LLM blocked content | Resolved with error handling and placeholder |
| 6b | Reviews from alternate location | ✅ Success | Validated review-only path |
| 7a | Slightly misspelled query (“hopsital”) | ✅ Success | No degradation in output |
| 7b | Severely misspelled query (“hasptal”) | ✅ Success | Output unaffected |

### **References**

* Microsoft Azure, *Azure OpenAI Service Models*. Microsoft Learn. Retrieved April 2025, from<https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models>
* OpenAI, *OpenAI API Reference*. OpenAI Documentation. Retrieved April 2025, from<https://platform.openai.com/docs/api-reference>
* Google Developers, *Google Maps Platform Documentation*. Retrieved April 2025, from<https://developers.google.com/maps/documentation>
* Docker Inc., *Docker Desktop Documentation*. Retrieved April 2025, from<https://docs.docker.com/desktop/>