Technical Report

Faster, Smarter, User-Aligned: `EvalOps` and the Future of Integrated Evaluation for the IC

Bijesh Shrestha[1], Hilson Shrestha[1], Karen Bonilla[1], Lane T. Harrison[1]* and R. Jordan Crouser [2]

[1] Worcester Polytechnic Institute; {bshrestha; hshrestha; kbonilla; ltharrison}@wpi.edu
[2] Kenyon College; crouser1@kenyon.edu
* Correspondence: ltharrison@wpi.edu

Abstract: The development of AI-enabled tools for the Intelligence Community (IC) faces a persistent challenge: while technical systems and Machine Learning models advance rapidly, ensuring those tools reliably enhance analyst workflows and operational effectiveness remains complex and time-consuming. `EvalOps` addresses this challenge by extending MLOps principles to focus on rapid, scalable, and systematic evaluation of human-machine interfaces. Rather than treating evaluation as an afterthought, `EvalOps` embeds stakeholder feedback, usability assessment, and iterative testing from the start of system development and sustains it throughout. During SCADS 2025, we expanded `EvalOps` through a series of four structured design worksheets that guide project teams through the process of integrating evaluation from the earliest stages on. Combined with multi-level instrumentation and targeted empirical studies, these efforts enabled faster interface refinement, improved alignment with analyst workflows, and accelerated time-to-deployment across operational contexts.

Keywords: Evaluation; design; analyst workflow

1. Introduction

The rapid evolution of AI technologies spans several areas critical to intelligence analysis, including: automated summarization, recommendation, entity extraction, and multimodal analytics, and continues to transform the landscape. However, while technical progress in system architectures and Machine Learning models is well-documented under the umbrella of MLOps, a persistent and consequential gap remains between technically performing systems and operationally effective, user-centered tools. This gap is especially pronounced in the Intelligence Community (IC), where usability, reliability, and alignment with analyst workflows are mission-critical.

The `EvalOps` project addresses this challenge by advancing rapid, scalable, and stakeholder-driven evaluation processes for AI-enabled human-machine teaming. Originally conceptualized and piloted during SCADS 2024, `EvalOps` builds on our prior success in designing and deploying highly instrumented empirical studies that integrate directly into interface development pipelines. The approach complements traditional MLOps practices by focusing not on model performance alone, but on system usability, stakeholder feedback, and alignment with operational needs as well.

Building on this foundation, our efforts during SCADS 2025 expanded `EvalOps` from a prototype into a more comprehensive methodology and toolkit. We piloted a series of four targeted design worksheets aimed at scaffolding the integrated evaluation process, lowering barriers to stakeholder engagement, and embedding evaluation earlier and more consistently within development life cycles. In addition to the worksheets, we generated concrete examples and vignettes illustrating how `EvalOps` has directly contributed to system improvements, including enhanced summarization interfaces, refined user interactions, and accelerated feedback loops. This report documents our SCADS 2025 efforts, presents the evaluated tools and processes, and reflects on lessons learned to guide future operationalization of `EvalOps` across the LAS ecosystem.

This technical report reflects work that was conducted as part of the 2025 Summer Conference on Applied Data Science (SCADS) hosted by the Laboratory for Analytic Sciences (LAS) at North Carolina State University.

2. Background and Motivation

In SCADS 2024, we conducted a preliminary feasibility study aimed at improving trust and usability in AI-generated document summaries. This effort led to the development of *SummShaper*: a prototype that allowed intelligence analysts to verify summaries against source content and adjust the output using contextual, interactive features. During early iterations—beginning with the *SumSifter* prototype—it became clear that aligning AI-generated content with analyst needs required more than effective visual encodings or technique integration. It demanded embedded evaluation practices capable of revealing pain points, supporting verification, and bringing real user input into the design process.

To meet this need, we moved away from traditional linear design-evaluation cycles in favor of a lightweight, rapid feedback model supported by the *reVISit* platform [1,2]. This model enabled us to explore ideas like hallucination detection, layout alternatives, and simplification controls—starting with sketches and mockups, and gradually evolving into interactive, LLM-driven prototypes that matched end-user workflows. We observed that even basic evaluation prompts, introduced early, fostered richer design conversations and more deliberate, user-centered decisions.

The constraints and complexities of intelligence analysis workflows, which require rapid iteration, traceable decisions, and sustained alignment across teams, further underscored the need for structured and integrated evaluation tools. In these environments, evaluation cannot be an isolated phase: it must be embedded throughout the design process to ensure continuity, adaptability, and responsiveness to evolving analytic needs. These observations informed the creation of a reusable set of `EvalOps` design activity worksheets to help teams keep evaluation central throughout the design and development lifecycle.

Our approach draws on and extends several foundational contributions in visualization design research.

McKenna et al.'s *Design Activity Framework* [3] links what designers do (process models) with why they do it (decision models), using the nested model to clarify design decision points. Walny et al. [4] highlight the challenges of transferring design intent and context between stakeholders—particularly between designers and developers—but stop short of providing mechanisms for tracking evolving metrics and decision rationale across a project's lifecycle. We build on these insights by introducing concrete tools to capture, revisit, and communicate design reasoning over time.

Similarly, Oppermann and Munzner's *Data-First* methodology [5] advocates beginning design with deep attention to stakeholder needs and data characteristics. This principle is central to `EvalOps`, which incorporates stakeholder-grounded abstraction and metric considerations from the outset to ensure both feasibility and relevance. To support evaluation that is both task-specific and operationally grounded, `EvalOps` integrates the *Analyst Hierarchy of Needs* (AHON) model by Girona et al. [6]. AHON informs how we structure task-to-metric mappings, prioritize features based on analyst goals and roles, and calibrate support for trust and interpretability in real-world analytic contexts. Together, these foundations shape the structure and intent of the `EvalOps` design sheets: not simply as tools for idea generation, but as practical scaffolds for aligning decisions with domain goals, embedding evaluation across the lifecycle, and maintaining coherence within fast-moving, multidisciplinary teams.

3. EvalOps Approach and Technical Contributions

`EvalOps` is a design framework built around one core idea: *evaluation should not be an afterthought.* Instead of waiting until a system is fully built to gather feedback, `EvalOps` helps teams surface questions early, test ideas often, and align design decisions with real analytic needs. The framework comprises a set of structured activity sheets that guide teams through four key phases: sensing the problem space, co-creating with stakeholders, reflecting on what worked and deciding on the next steps, and ensuring a clear transition and handover (Fig. 1). Each phase integrates design activities with focused evaluation

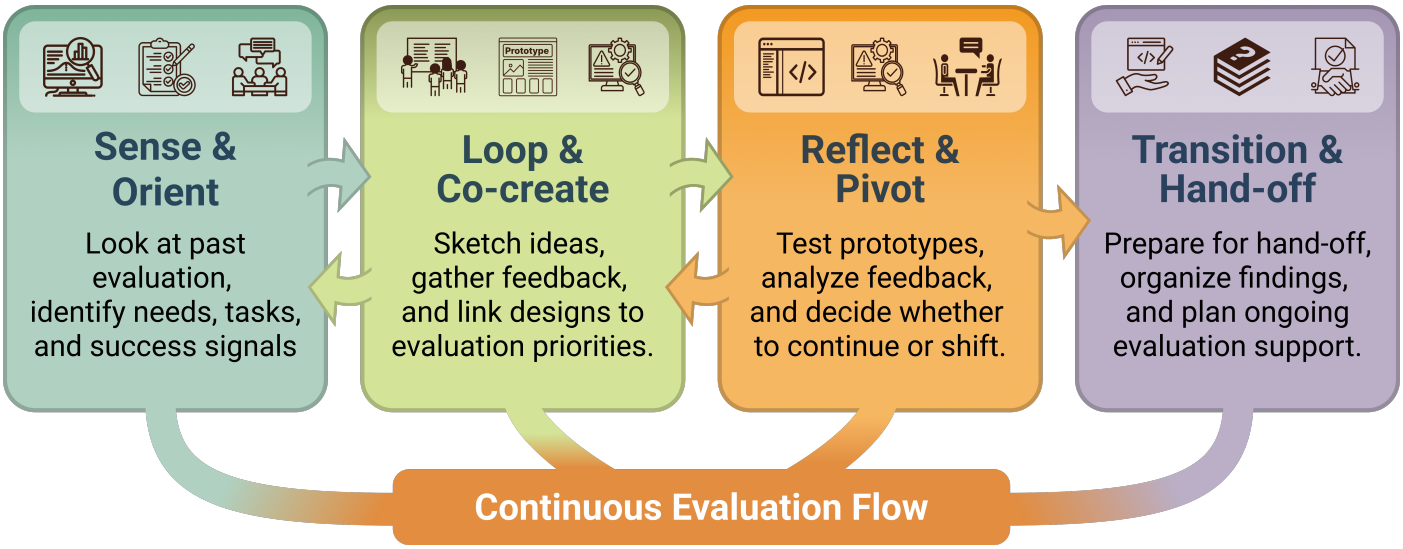


Figure 1. The four `EvalOps` phases, each of which has a corresponding worksheet that can be used in a modular way.

to close the decision loop, enabling teams to stay aligned and responsive throughout the development process.

`EvalOps` design sheets are built to keep things simple and focused, helping teams manage design and development in fast-paced settings like SCADS. The sheets are lightweight, modular, and flexible. They are easy to pick up during a one-hour working session, integrate into existing sprints, or revisit at key decision points. They also help teams stay in sync by making their activities and decisions visible. The design sheets provide a shared structure that helps teams clarify their current status, identify any gaps, and determine next steps.

Additionally, the activity sheets also support teams in identifying what to track, how to connect observations to goals, and where feedback should shape decisions. While each sheet can be used during a focused session, teams often return to them as their work evolves. Some prompts depend on earlier inputs, making the sheets naturally iterative and adaptable to the team's pace.

The `EvalOps` design methodology can help teams stay aligned in the tasks they are supporting and the metrics appropriate to the domain. By capturing feedback at the right moments and tracing how ideas change, teams can close the loop between intention and outcome. When paired with tools like reVISit, additional data such as interaction traces, surveys, and open-ended feedback can enrich the process [2]. Even without formal infrastructure, the sheets offer a practical way to reflect, recalibrate, and move forward with clarity. Because `EvalOps` makes evaluation more visible and intentional, it fits within the process naturally and without slowing teams down.

Extending `EvalOps` to Diverse Operational Contexts

`EvalOps` is intentionally flexible and adaptable to varied mission settings. In contexts like SIGINT, TECHINT, and multimodal data fusion, analytic tasks often involve complex, evolving goals and heterogeneous data types. `EvalOps` can be extended to these domains by tailoring sensing and co-creation activities to surface modality-specific challenges (e.g., signal ambiguity, time-sensitivity, source reliability) and by adapting reflection prompts to support traceability across diverse evidence types. The modular structure also enables insertion of domain-specific evaluation metrics or decision gates, while maintaining the lightweight, iterative nature of the framework. This adaptability positions `EvalOps` as a unifying scaffold across diverse analytic workflows, helping teams reason through uncertainty, reconcile priorities, and remain grounded in operational goals.

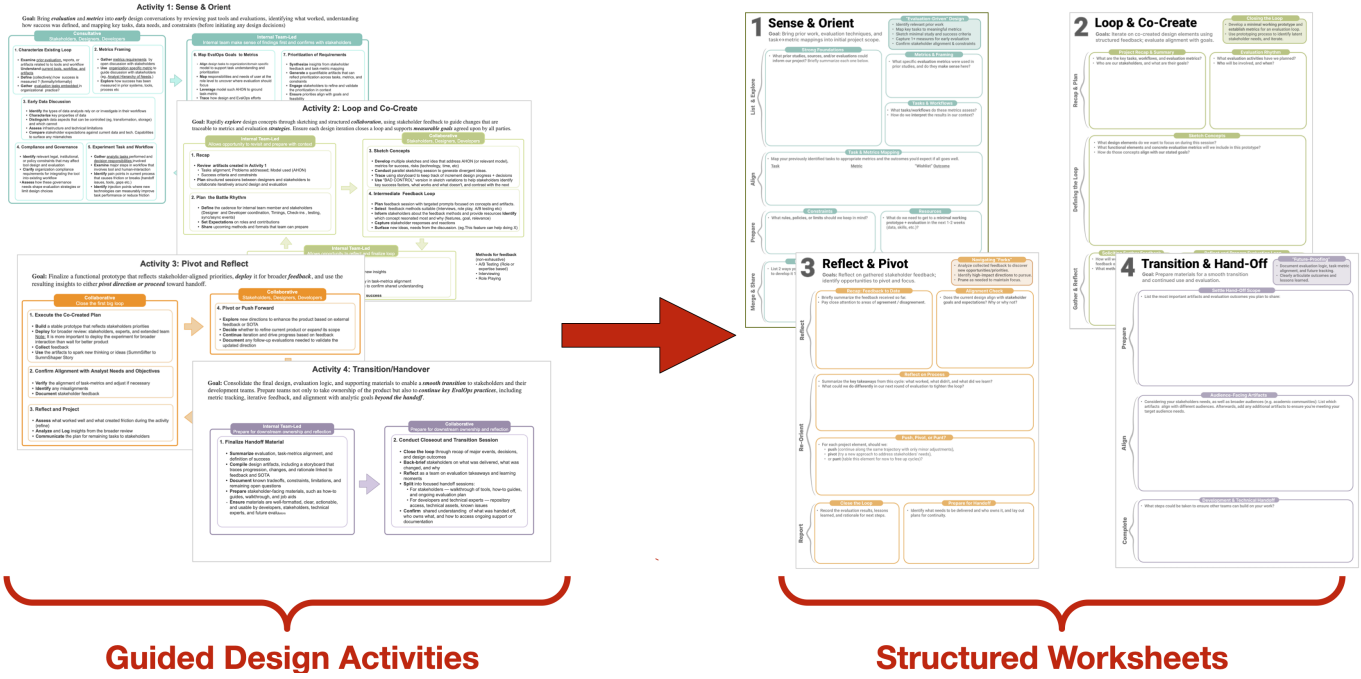


Figure 2. `EvalOps` activity sheets and supporting materials designed to structure and sustain evaluation across the design and development lifecycle.

4. SCADS 2025 Pilot Activities

4.1. Design Worksheet Series

The `EvalOps` design worksheet series was created to help teams without slowing them down. These worksheets were tested through the internal design and development lifecycle and refined in preparation for fielding events at SCADS 2025. The design sheets provide teams with a grounded way to pause, align, and make their decision-making a collaborative effort. Each activity sheet corresponds to a core `EvalOps` phase as shown in Fig. 1.

Activity 1: Sense and Orient This worksheet helps teams map the problem space before jumping into ideas. It focuses on identifying prior evaluations, success criteria, analyst goals, and data or compliance constraints. Teams use it to define what success could look like, grounded in real constraints and stakeholder-defined priorities. It also supports early task-to-metric mapping, helping teams anticipate what should be evaluated later. In SCADS, teams can use this sheet to structure conversations during the first two weeks and to prep for early stakeholder check-ins.

Activity 2: Loop and Co-Create Once teams start sketching concepts and coding early prototypes, this sheet can help the designers connect those ideas back to the goals and metrics defined through stakeholder engagements. The goal of this activity sheet is to enable collaborative sessions among designers, developers, and stakeholders. This process will enable teams to test early design sketches, run A/B studies, and explore "bad control" designs to surface gaps. The sheet can help track how concepts evolve over time, what feedback drove those changes, and whether new ideas remain aligned with intended tasks and metrics.

Activity 3: Pivot and Reflect This worksheet supports reflection and decision-making. It helps teams synthesize feedback, map results to earlier metrics, and decide whether to continue, pivot, or punt some of the efforts. During SCADS, this sheet can help teams pause after early prototypes and recalibrate without losing sight of what matters.

Activity 4: Transition and Handover This sheet helps teams prepare and execute transition and handover as deliberate actions. This includes logistical and procedural steps

that enable a different team or stakeholders to take the tool or product forward. It focuses on packaging key decisions, documentation, code bases, and clarifying what still needs further refinement. It also helps plan how evaluation practices can continue after the transition and handoff events have been completed. At SCADS, this is particularly valuable for teams delivering partial solutions, building on multi-year efforts, or for research handoffs to sponsors.

Early Observations: Initial feedback from SCADS 2025 suggests that teams found the `EvalOps` sheets helpful for organizing their thinking, aligning the project with stakeholders needs, and deciding what to track and revisit. Even when the sheets were not filled out completely during the instructional session, the prompt offered structure that would help teams in guiding the internal discussion and identifying next steps. Teams referenced prior SCADS examples to adapt sheets to their own domains, and several noted that the examples made it easier to understand how to get started. While formal feedback sessions have not yet been conducted, informal conversations suggest that teams appreciated having a tangible structure to return to.

4.2. System Case Studies and Vignettes

SummShaper originated as SummSifter during SCADS 2024, serving as a prototype for verifying AI-generated summaries. From the beginning, we applied key `EvalOps` principles by raising evaluation questions early, sketching alternatives, and using stakeholder input to guide design. Early stakeholder feedback revealed broader needs around the ability to shape tailored summaries across multiple documents, not just validate them. These cycles improved alignment with real analyst workflows and led to quicker adoption of new summary shaping features.

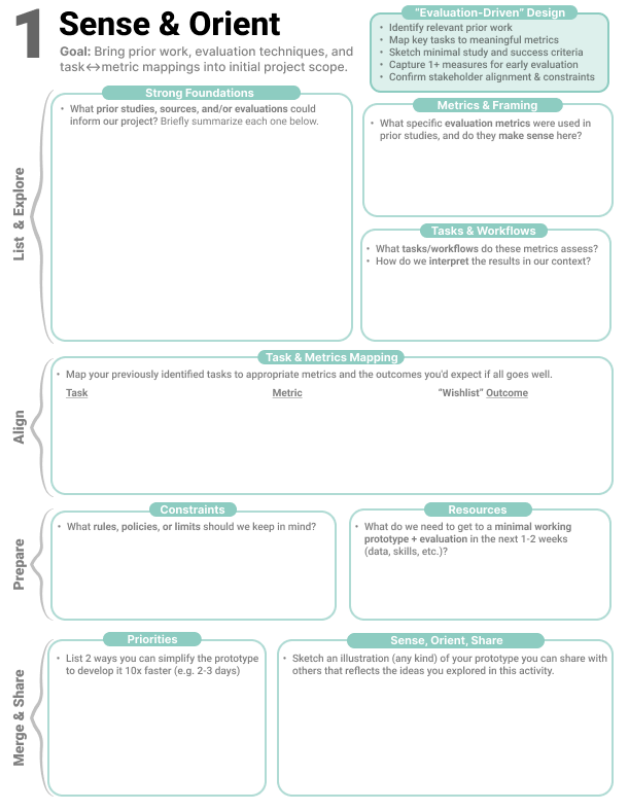
This shift was made possible by the lightweight, feedback-driven practices, which are the core tenet of `EvalOps`. The project highlighted where structure helped and where flexibility mattered, ultimately refining how we framed the sheets for broader audience. The task-metric mapping in SummSifter informed the **Sense and Orient** activity, the sketch-review cycles helped define **Loop and Co-Create**, and the lessons learned around decision to pivot directly shaped **Pivot and Reflect** activity sheet.

PandaJam was the first mature tool where we validated the EvalOps framework in full. Originally developed during SCADS 2022 and refined through prior LAS studies, PandaJam helps intelligence analysts triage noisy, long-form speech-to-text transcripts such as the Nixon tapes. The system includes tools for search, segmented summaries, visual confidence indicators, and structured review workflows. Unlike SummShaper, which evolved alongside `EvalOps`, PandaJam enabled us to test whether the design sheets could support structured evaluation for a larger, more complex system that was already in development. With contributors spread across three institutions and multiple stakeholders, PandaJam served as an ideal testbed for assessing the flexibility, timing, and practical value of the `EvalOps` approach in real-world settings. Using `EvalOps`, we identified usability gaps, refined triage workflows, and aligned better with analyst feedback across institutions

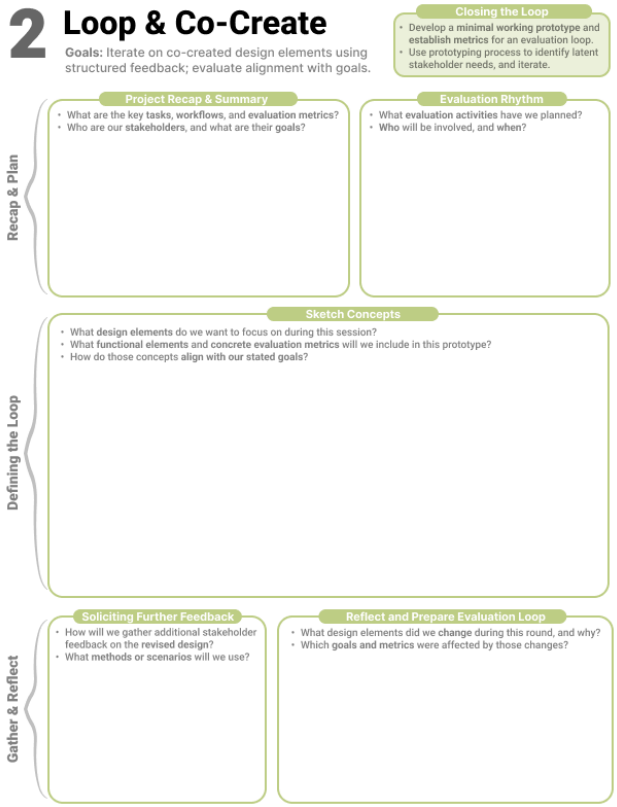
5. Lessons Learned and Best Practices

Implementing the `EvalOps` design framework in real project settings, including SCADS 2025, surfaced several technical, and socio-technical friction points.

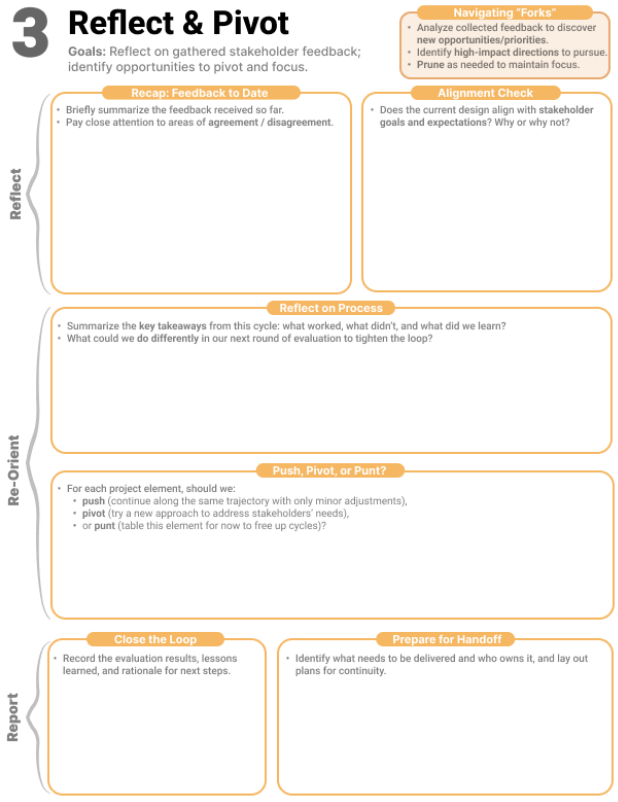
- **Establishing early rapport and rhythm with both the team members and stakeholders matters:** The early weeks involved scoping design challenges, building working relationships, and establishing shared rhythm for events and milestones. During both SummShaper and PandaJam projects, for example, the design team maintained shared document and communication channel to collaborate efficiently. The trust built with analysts early also made it easier for them to speak openly, without hesitation or worrying about being overly polite. The groundwork laid helped team collaborate and respond to feedback more quickly.



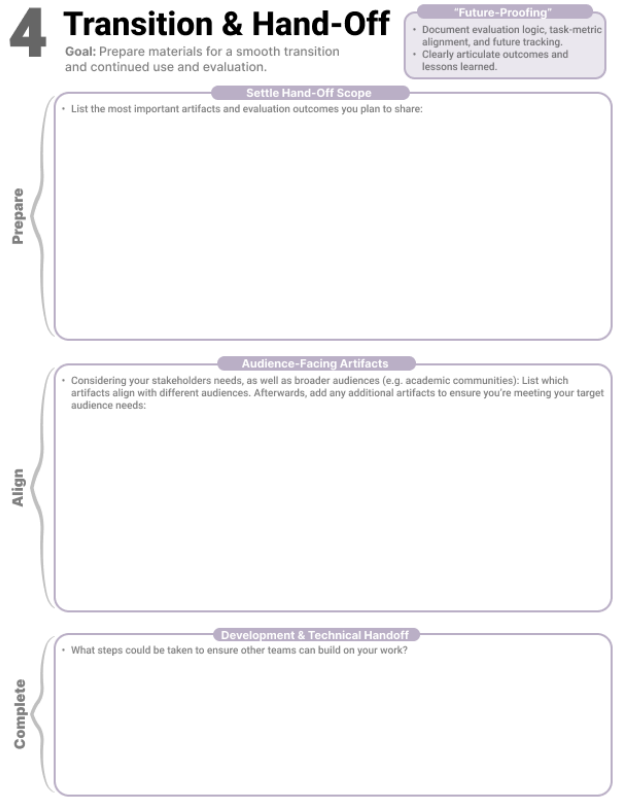
(a) Activity 1: Sense and Orient



(b) Activity 2: Loop and Co-Create



(c) Activity 3: Reflect and Pivot



(d) Activity 4: Transition and Handover

Figure 3. `EvalOps` Activity Sheets (1–4). These sheets help teams plan, test, and reflect during key phases of design and development

- **The reVISit platform enabled faster iteration Early iteration:** By instrumenting the experiment early on the reVISit or similar platform enable early and rich feedback from various participants at speed. During SummShaper, PandaJam, and various other projects, the reVISit platform allowed for quicker iteration of the experiment through structured feedback that included think-aloud notes, surveys, and provenance traces while the design is still being implemented.
- **Evaluation efforts are shaped by social access and not just through infrastructure:** We observed differences across teams in how quickly they were able to work with analysts. Teams that built early touchpoints through informal conversations, shared working sessions, or hallway chats were better positioned to gather useful feedback and adjust course when needed.

Emerging best practices for integrating `EvalOps` into AI development lifecycles:

- **Start small, but start early:** We observed that it is more important to get the idea out early in the process. Events such as *A-Day-in-the-Life-of-Analyst* are very valuable to all participants. However, it is more of a one-way conversation between the presenter and the audience, with some interactions. Lightweight design prompts or artifacts helped teams open a two-way dialogue with stakeholders.
- **Use artifacts to drive discussion:** Using a collection of visual artifacts that convey the idea, problem, and a path to solution jump-starts a creative, two-way communication between stakeholders and the design team. This also helps stakeholders to properly understand the design space for solutions and tailor the feedback to each team.
- **Don't lose sight of task-metric pairing:** Project scoping is a challenging task. Given many interesting problems and feature discussions at SCADS, teams often run into the issue of *project scope creep*. It is essential to consistently return to the initial tasks aligned with metrics. This is particularly important at later stages, when teams need to prioritize and scale down their efforts to meet the milestone deadlines.
- **Make decisions visible and deliberate:** Keeping track of design decisions, factors involved, and tradeoffs made can be very helpful in justifying reasons around *Pivot* or *Punt* and preparing for handoff tasks.

6. Future Directions

The `EvalOps` design framework and design sheets provide a lightweight yet structured approach to keeping evaluation at the center of AI research and development efforts for the analytic workflow. The broader goal is not just to standardize evaluation-centered practices at SCADS, but to foster a cultural shift at SCADS where teams start with stakeholder goals, iterate with intent, and arrive at solutions aligned with shared analytic problems.

- Opportunities for scaling `EvalOps` across LAS projects: The validation of the `EvalOps` during SCADS 2025 showed that even a modest structure the sheets and framework provided could meaningfully shape project direct and final outcomes. To build on this observation, the `EvalOps` method could be introduced during the on-boarding week, which not only discusses the sheets, technology for feedback but also to set a shared expectations around evaluation and collaboration with stakeholders.
- Proposed technical extensions: [Brainstorming]
  - Are we thinking beyond sheets .... like a platform that can integrate different projects easily for the same iterative cycle that we did with reVISit?
  - and/or design traceablity infrastructure –> that can maybe track factors to design decision –> which can then map the evolution of the design,
  - and/or tie evalops sheets to reVISit where the evaluation questions asked are based on metrics discussed and agreed in previous activity?
- Pathways for deeper stakeholder collaboration and continuous evaluation: The `EvalOps` activity sheets helps ensure stakeholder input is not only gathered, but carried forward in ways that remain actionable. Updated versions of the worksheets can surface key decisions, rationale, and unresolved questions from earlier phases,

supporting midstream check-ins and stakeholder re-engagement. Revised activity and supporting sheets with built-in prompts and placeholders can provide teams with natural hooks for follow-ups and maintaining continuity, even through team changes over time.

7. Conclusions

The EvalOps framework demonstrates how lightweight, structured design activities can keep evaluation central to AI-augmented system development in high-stakes environments of IC. Developed through SummShaper project and further refined with new PandaJam project, the activity sheets helped teams identify evaluation needs early, document decisions, and stay aligned with stakeholder goals. At SCADS 2025, even limited use of the sheets appeared to support planning, reflection, and iteration. As EvalOps continues to grow, the sheets offer a repeatable and adaptable way to embed continuous evaluation across diverse analytic contexts.

Acknowledgments: This work was supported in part by NSF#2213757

References

1. ReVISit. ReVISit. https://revisit.dev/, n.d. Accessed: 15 July 2025.
2. ReVISit Team. ReVISit Study Platform. https://github.com/revisit-studies/study, 2025. Accessed: 15 July 2025.
3. McKenna, S.; Mazur, D.; Agutter, J.; Meyer, M. Design Activity Framework for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics* **2014**, *20*, 2191–2200. https://doi.org/10.1109/TVCG.2014.2346331.
4. Walny, J.; Frisson, C.; West, M.; Kosminsky, D.; Knudsen, S.; Carpendale, S.; Willett, W. Data Changes Everything: Challenges and Opportunities in Data Visualization Design Handoff. *IEEE Transactions on Visualization and Computer Graphics* **2020**, *26*, 12–22. https://doi.org/10.1109/TVCG.2019.2934538.
5. Oppermann, M.; Munzner, T. Data-First Visualization Design Studies. In Proceedings of the 2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV), 2020, pp. 74–80. https://doi.org/10.1109/BELIV51497.2020.00016.
6. Girona, A.E.; Peters, J.C.; Wang, W.; Crouser, R.J. The Analyst's Hierarchy of Needs: Grounded Design Principles for Tailored Intelligence Analysis Tools. *Analytics*, *3*.