Problem Statement and Goals
Software Engineering

Team #1, Sanskrit Ciphers
Omar El Aref
Dylan Garner
Muhammad Umar Khan
Aswin Kuganesan
Yousef Shahin

Table 1: Revision History

Date	Developer(s)	Change
09/13/2025	Omar El Aref	Added the problem statement, Inputs and Outputs
09/14/2025	Omar El Aref	Added Stakeholders, Environment, Goals, Stretch Goals and Extras
09/22/2025	Omar El Aref	Added Reflection
09/22/2025	Omar El Aref	Applied Feedback Changes

1 Problem Statement

Throughout this document we aim to clearly define the challenges, inputs, outputs, and goals of our proposed system. By outlining the problem context and the environment in which our solution will operate, we aim to establish a foundation for development and evaluation. This problem statement sets the stage for the subsequent sections, where we describe the problem in greater detail, identify the stakeholders, and specify the technical and scholarly goals of the project.

1.1 Problem

The textual history of Indian Buddhism is fragmented across thousands of manuscript folios, preserved only in partial, damaged, or scattered forms. Traditionally, scholars reconstruct these texts manually through paleographic study, transcription, and content comparison; a slow and error-prone process. While

some existing computational tools exist for pattern recognition and Optical Character Recognition (OCR) none can tackle this problem as they are not designed for irregular, damaged, or arbitrarily oriented manuscript fragments.

The lack of computational tools available for this problem significantly limits progress in reconstructing Buddhist textual history. Scholars need a tool that automates the detection, matching, and transcription of manuscript fragments, thereby reducing manual effort and time; enabling large-scale reconstruction.

1.2 Inputs and Outputs

- **Inputs:**
  - High quality images of manuscript fragments (approximately 21,000 images from collections)
  - Metadata (collection identifiers, orientation, partial transcriptions if available)
- **Outputs:**
  - Probabilistic matches between fragments based on shape/edge/damage features (i.e., list of fragments most likely related to an input fragment image)
  - Enhanced Metadata: Suggested transcriptions of fragment text
  - Searchable/sortable database containing fragment attributes and relationships. This will allow the user to search the database and will return the fragments and its attributes and relationships.

1.3 Stakeholders

- **Primary Stakeholders:**
  - Scholars of Buddhist textual history (religious studies, philology, palaeography). They are a primary user of the system as they will be the ones interacting with the system the most to help them with their work.
  - Supervisors and domain experts (e.g., Dr. Shayne Clarke, McMaster Religious Studies). They will directly benefit from this product as they will be using it in their everyday work. Which is why they are a primary stakeholder.
- **Secondary Stakeholders:**
  - Computer science researchers in Machine Learning (ML), image processing. They might be interested in how an ML model works to tackle similar problems to this and so that is why they are a secondary stakeholder and not primary.

  - Humanities researchers studying textual transmission and manuscript culture that aren't Buddhist texts. This would be mainly people who are more so interested in the manuscript research rather than the reconstruction of manuscripts. They aren't the intended user for this system but they can definitely use it and therefore they are secondary users.

- **End Users:**
  - Academic researchers using the software to assemble fragments (Researchers who aren't primary or secondary stakeholders). Mainly anyone interested in any aspect of manuscripts and their history can use this. This is not the intended use for the system but it could be used that way and so that is why they are end users.
  - Graduate students seeking computational support in philological research. This again wasn't the intended use for the system and so that is why they are end users.

1.4 Environment

- **Hardware:**
  - Development laptops/workstations with GPU support for ML tasks
- **Software:**
  - VScode
  - Coding Libraries
  - Database
  - GitHub repository for version control and CI/CD

2 Goals

- Develop a tool that:
  - Detects edges, shapes, and damage patterns in fragments with at least a 95% accuracy.
  - Matches fragments based on similarity measures (probabilistic scoring). We will measure this by seeing how many fragment matches it gets right over how many attempts it had overall. This number needs to be over 90%.
  - Identifies paleographic script of fragments. This can be measured by comparing the results with the actual script from a manuscript and seeing how well it does. This will be at least 90% accurate.

- Performs preliminary transcription using OCR tuned to Sanskrit scripts. This can be cross referenced with the existing confirmed translation we have and through that we can measure its accuracy which will be over 90%
- Builds a searchable database linking fragments with Metadata and probable matches. This can be measured by seeing if it returns a match when its supposed to or not. This needs to be correct all the time.
- Provides a user interface for scholars to view suggested fragment matches and confirm/annotate them. This will be measured on the basis of if the UI works or not.

3 Stretch Goals

- Expand support to other languages such as Tibetan and Chinese manuscript fragments.
- Incorporate semantic content matching with other fragments.
- Improve transcription accuracy with AI-assisted error correction.

4 Extras

- **User Documentation:** Write user-friendly guides for scholars with limited technical expertise.
- **Use Case Video:** Make a Video on how to use the tool so that we can minimize the learning curve for the intended user

Appendix — Reflection

1. What went well while writing this deliverable?
2. What pain points did you experience during this deliverable, and how did you resolve them?
3. How did you and your team adjust the scope of your goals to ensure they are suitable for a Capstone project (not overly ambitious but also of appropriate complexity for a senior design project)?

What went well the most would be that everyone was on the same page for this project and so coming up with what we wanted to accomplish was pretty easy and we didn't really have many disagreements. This made it easy to focus on the task of writing the problem statement and goals rather than having to worry about different opinions on the team. This project was also one of the projects that was on the list of projects pdf given to us so it was easy to figure our the problem statement and goals since we didn't really have much to come up with anything on our own.

One challenge was avoiding either too much technical detail or too much abstraction. We really had to focus on what we wanted to convey and add too much detail as to not contrain our selves but also not have too little detail that it isn't clear as to what we are doing. We initially struggled to find the right balance between scholarly needs (manuscript context) and technical specifications (ML algorithms, environment). We resolved this by starting broad, then refining with feedback and checking against the POC and problem statement checklists. That way we made sure that our POC and problem statement were in line with each other. Another pain point was uncertainty about which machine learning techniques would realistically be feasible; to address this, we distinguished between core goals and stretch goals to avoid overcommitting. This way we also didn't contrain oursleves later down the line when we start implementing the solution.

Initially, we considered a full end-to-end system covering Sanskrit, Tibetan, and Chinese manuscripts. We recognized this was quite ambitious especially given the time and resources that we had. Instead, we narrowed our core scope to Sanskrit fragments only, focusing on orientation correction, edge/damage-based matching, and preliminary script identification. Transcription and cross-lingual extensions were moved into stretch goals as they are not the core goals that we are trying to achieve with this project. If we are ahead of schedule then they would be great additions to add to the project but again as mentioned they are not the core focus of this project. This adjustment ensures the project is challenging enough, but achievable within the Capstone timeline.