PsychExtract: Draft Report

COURSE
CM3070 Final Project
Computer Science
University of London

9,902 words

PROJECT FILES
https://github.com/itsCarli/PsychExtract

ABSTRACT (250 words)

Psychotherapeutic practice increasingly relies on reflective analysis of client narratives, yet manually extracting emotional and cognitive patterns from written material remains time-intensive and subjective. PsychExtract explores the feasibility of an AI-assisted pipeline designed to support reflective practice by automatically identifying emotional signals, insight-related themes, and interpretable summaries from client-generated text. The system integrates optical character recognition (OCR) for handwritten and scanned journal entries, natural language processing (NLP) techniques for fine-grained emotion classification and thematic extraction, and an output layer that delivers both textual and text-to-speech (TTS) summaries. Grounded in established psychological theory, the design draws on research into insight as a mechanism of therapeutic change and the role of emotion in cognitive restructuring. A review of existing AI-supported mental health tools informs methodological and ethical decisions, ensuring interpretability and user-centred design. The implemented architecture combines OCR preprocessing, transformer-based emotion classification inspired by GoEmotions, keyword and topic extraction for interpretability, and TTS synthesis for accessibility. A functional prototype is evaluated using quantitative metrics and qualitative error analysis, focusing on OCR feasibility, transcription accuracy, and downstream interpretability. Results show that while OCR performance varies across handwriting styles, targeted preprocessing substantially improves the quality and usability of extracted insights. These findings highlight the potential and limitations of NLP-driven reflective tools in sensitive mental-health contexts and motivate future work on multimodal evaluation, interactive user interfaces, and user-centred validation in collaboration

CONTENTS

INTRODUCTION (998/1000 WORDS)

PROJECT OVERVIEW AND MOTIVATION

Psychological reflection and therapeutic writing support emotional awareness, insight formation, and personal growth. Across both formal psychotherapy and informal self-reflective practices, individuals produce handwritten or semi-structured texts such as journals, therapy notes, worksheets, and reflective exercises (Hill et al., 2007; Greenberg & Pascual-Leone, 2006). These artefacts often contain rich emotional cues, recurring linguistic patterns, and implicit cognitive framing. Despite their value, such materials are rarely analysed systematically; insight extraction remains largely manual, subjective, and time-intensive (Koleck et al., 2019; Mukherjee et al., 2020; Turner et al., 2022).

This challenge is particularly evident for handwritten material. While digital mental health tools and text-based analytics have expanded rapidly (Inkster et al., 2020; Torous et al., 2018), a substantial portion of reflective writing still occurs on paper, creating a gap between expressive practice and computational support. Manual review does not scale well, is prone to inconsistency, and may fail to surface longer-term patterns, contributing to cognitive and administrative burden (Shanafelt et al., 2016). Fully automated interpretation raises ethical concerns when systems extend beyond assistance into diagnosis or decision-making (Luxton, 2014; MHRA, 2025; Miner et al., 2016; Mohammad, 2022). These tensions motivate the need for assistive tools that structure reflective content while preserving transparency and human oversight.

PsychExtract is developed in response. The project investigates whether handwritten, mental-health-related text can be transformed into structured insight through a modular, interpretable computational pipeline. Rather than inferring mental health conditions or offering therapeutic advice, the system extracts signals present in the text (emotional tone, recurring themes, and linguistic framing) and presents them in a form that supports reflection. This aligns with ethical guidance emphasising assistive, non-diagnostic AI in mental health contexts (Jacobs et al., 2021; Luxton, 2014; Mohammad, 2022).

PROJECT TEMPLATE

This project follows Template 4.1: Orchestrating AI Models to Achieve a Goal. Template 4.1 suits PsychExtract because the contribution lies in system-level integration rather than a single novel model. Prior work in clinical and therapeutic NLP shows that meaningful insight often emerges from combining complementary techniques (emotion classification, linguistic analysis, keyword extraction) rather than relying on one predictive model (Doshi-Velez & Kim, 2017; Inkster et al., 2020; Koleck et al., 2019; Mukherjee et al., 2020). Each stage of the pipeline (that is optical character recognition, emotion classification, linguistic analysis, semantic summarisation, and optional accessibility output) offers multiple modelling approaches. The template provides a structured framework for selecting, integrating, and evaluating these components.

This orchestration-focused approach aligns with research advocating transparent, modular AI systems in sensitive domains, where trust and explainability are as important as technical performance (Doshi-Velez & Kim, 2017; Jacobs et al., 2021).

PROJECT AIMS AND OBJECTIVES

The overarching aim is to support reflective understanding of handwritten text through structured, interpretable computational analysis. PsychExtract is explicitly non-clinical: it does not diagnose, predict outcomes, or offer therapeutic advice, consistent with regulatory guidance (Luxton, 2014; MHRA, 2025; Mohammad, 2022).

To achieve this, the project designs a modular pipeline converting handwritten reflective text into digital representations. OCR forms the foundation, as errors here propagate downstream. Following digitisation, emotion classification uses transformer-based models selected for performance and interpretability (Demszky et al., 2020; Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019). Linguistic pattern extraction identifies recurring keywords, phrasing, and framing strategies signaling cognitive or emotional emphasis (Grootendorst, 2022; Grootendorst, 2025; Pennebaker et al., 2015). These signals are integrated into concise semantic summaries prioritising clarity and readability over abstraction.

The project also explores accessibility and user experience considerations, including optional text-to-speech (TTS) integration. TTS can enhance accessibility and reduce reading fatigue, especially for users with learning differences or cognitive load constraints (Shen et al., 2018; Young et al., 2018).

TARGET DOMAIN AND USERS

PsychExtract operates within the broader domain of mental health support, self-reflection, and therapeutic-adjacent technologies. Its focus is not on clinical intervention but on assisting reflective practices where emotional awareness and insight formation are valuable (Hill et al., 2007; Greenberg & Pascual-Leone, 2006). The system is intended for use in contexts where users benefit from structured reflection but where automated clinical judgement would be inappropriate or ethically problematic (Luxton, 2014; Mohammad, 2022).

The primary target users include students engaging in reflective writing as part of their academic or personal development, individuals who journal as a form of emotional processing, and therapy-adjacent users who reflect on experiences between sessions or prepare written material for discussion. Similar user groups have been identified in prior research on digital mental health tools and reflective technologies (Ben-Zeev et al., 2013; Inkster et al., 2020; Torous et al., 2018). In all cases, the system is positioned as an assistive tool rather than an authority. It structures and summarises content supplied by the user without interpreting intent, assigning diagnoses, or offering recommendations.

Maintaining clear boundaries around system capability is a central design principle of the project. PsychExtract does not claim psychological understanding or therapeutic expertise. Instead, it supports users in engaging more effectively with their own writing, reinforcing reflection rather than replacing it.

CONTRIBUTIONS AND ORIGINALITY

The originality of PsychExtract lies in its integration of multiple AI and NLP techniques into a coherent, reflection-oriented pipeline tailored to handwritten text. While individual components such as OCR, emotion classification, and topic modelling are well-established (Demszky et al., 2020; Grootendorst, 2022; Smith, 2007), their combination within a non-clinical, interpretability-

focused framework represents a novel application aligned with ethical AI design principles (Doshi-Velez & Kim, 2017; Jacobs et al., 2021; Mohammad, 2022).

The project contributes a system that bridges the gap between expressive handwritten practices and computational analysis, demonstrating how structured insight extraction can be achieved without overstepping ethical boundaries. Its emphasis on transparency, modularity, and accessibility distinguishes it from systems that prioritise prediction or automation alone. By foregrounding interpretability and future extensibility, particularly through planned interface development and accessibility features such as text-to-speech, PsychExtract highlights how AI systems can be responsibly designed to support sensitive human activities.

LITERATURE REVIEW (1938/2500)

FOUNDATIONS OF INSIGHT AND EMOTION IN PSYCHOTHERAPY

Insight as a Therapeutic Mechanism

Insight is widely recognised as a core driver of psychological change. Hill et al. describe it as a "conscious meaning shift involving new connections" (Hill et al., 2007, p. 442), noting that early insights often begin as simple realisations before deepening into more complex higher-dimensional forms (such as emotional understanding, cognitive restructuring, and reframing past experiences). They emphasise that the concept lacks a universally accepted definition and varies across therapeutic approaches. This ambiguity positions insight as both central and flexible, making it suitable for computational modelling when carefully scoped.

For the purposes of this project, these theoretical limitations (lack of consensus, variation across schools, and multi-dimensionality) clarify which components can be practically extracted from text. PsychExtract therefore focuses exclusively on extracting early-stage insight, which is typically expressed through observable language patterns such as emotional descriptions, reflective statements, and self-evaluative comments. These early meaning shifts are the aspects of insight most consistently expressed through text and most amenable to natural language analysis.

This contextualises why insight extraction matters. If early insight influences therapeutic progress, then summarising indicators of such insight from client reflections can help therapists track meaningful shifts over time. This first section of literature review establishes the theoretical motivation for the system, that is, insight is important, language is one of the primary ways it appears, and early insight is computationally detectable.

Emotion as a Core Component of Therapeutic Change

Greenberg and Pascual-Leone argue that emotional processing is a primary driver of change across therapeutic modalities (Greenberg & Pascual-Leone, 2006). They outline a process involving emotional awareness, regulation, transformation, and meaning-making. This typically requires clients to articulate internal emotional experiences. While therapists are trained to detect emotional cues, much emotional content is communicated implicitly through language, making consistency and standardization difficult in practice.

This challenge motivates PsychExtract. If written reflections, such as journals, homework tasks, and progress updates, contain emotional signals that contribute to therapeutic insight, then automated extraction of emotional and linguistic patterns can support therapists by ensuring these signals are made visible and consistently interpreted. The aim is not to replace therapist judgement.

Rather, PsychExtract produces structured summaries of emotional and cognitive patterns derived from client text. These summaries can draw attention to potential therapeutic themes without making clinical claims, maintaining alignment with ethical guidance in the field. This naturally leads to the next question of whether artificial intelligence is a suitable tool for supporting therapists in recognising these linguistic signals.

AI Support in Psychotherapy: Lessons from Existing Systems

The use of artificial intelligence in mental-health contexts is not new. DeVault et al. introduced SimSensei (DeVault et al., 2014), a virtual interviewer (which is captured in Figure 1) designed to detect psychological distress from verbal and nonverbal behaviour. Their work demonstrates several key findings relevant to PsychExtract. Notably, people often disclose more openly when interacting with automated systems, and even simple computational methods can highlight meaningful psychological cues (such as sentiment shifts or linguistic markers of distress).

Figure 1. Ellie, the virtual human interviewer used in the SimSensei Kiosk system (DeVault et al., 2014).

SimSensei's limitations are equally informative. Because it operates in real-time conversation, it must use extremely cautious and overly simplistic natural language models to avoid unsafe or inappropriate responses. As a result, the system relies on basic language processing.

PsychExtract diverges from this setting in two important ways. First, it is non-conversational. Users provide reflective text, and the system produces an analysis, not an ongoing dialogue. Second, it does not operate in real-time. These affordances allow PsychExtract to employ more advanced language-processing techniques safely, such as transformer-based architectures like BERT, because there is no risk of generating incorrect or harmful conversational replies.

This section therefore establishes why artificial intelligence is appropriate for insight extraction. Existing work shows that AI can highlight clinically relevant linguistic cues, and PsychExtract extends this by applying stronger models in a safer, offline workflow. This bridges into the next section by motivating how AI can be used. This is through specific natural language processing methods tailored to the linguistic components that make up early insight.

NLP Methods for Emotion, Insight, and Cognitive Pattern Extraction

Before examining technical methods, it is important to clarify terminology for non-specialist readers. NLP refers to computational techniques for analysing or generating human language.

Modern NLP often uses transformer-based models, which are deep learning architectures capable of understanding words in context rather than in isolation. These models outperform traditional techniques in tasks involving emotion recognition, topic inference, and meaning extraction, all of which are relevant to early insight.

This section details the three components of insight that PsychExtract identifies through NLP: Emotion expression, cognitive themes and reflective topics, and linguistic patterns associated with meaning-making.

By connecting these components to the earlier theory section, PsychExtract grounds its extraction pipeline directly in the psychological mechanisms of insight.

Fine-Grained Emotion Classification (GoEmotions)

Demszky et al. introduce GoEmotions, a dataset of 58,000 Reddit comments labelled with 27 fine-grained emotion categories excluding a neutral class (Demszky et al., 2020), visualized in Figure 2. Their findings show that transformer-based models such as BERT significantly outperform traditional machine learning approaches for understanding emotional nuance, especially because emotions often overlap and require contextual interpretation.

Positive		Negative		Ambiguous
admiration 👏	joy 😃	anger 😡	grief 😢	confusion 😕
amusement 😂	love ❤️	annoyance 😒	nervousness 😬	curiosity 🤔
approval 👍	optimism 🤞	disappointment	remorse 😔	realization 💡
caring 🤗	pride 😌	disapproval 👎	sadness 😞	surprise 😲
desire 😍	relief 😅	disgust 🤮		
excitement 🤩		embarrassment 😳		
gratitude 🙏		fear 😨		

Figure 2. GoEmotions emotion taxonomy comprising 28 fine-grained emotion categories, including a neutral class (Demszky et al., 2020).

GoEmotions is valuable as a baseline for PsychExtract, but it has limitations. It contains short social-media comments rather than long reflective writing, and deeper therapeutic emotions (such as, grief processing, self-evaluation, growth-related fear) are underrepresented. To address this, PsychExtract uses GoEmotions models for initial benchmarking but extends beyond the dataset by incorporating long-form reflective text. This is in the form of available corpora (such as r/offmychest) or carefully synthesised paragraphs, which are designed to preserve emotional coherence without introducing clinical claims. This supports the system's goal of aligning emotion extraction with therapeutic contexts.

By grounding the emotional component of insight in this literature, PsychExtract builds directly on empirical evidence that transformer-based models are the strongest choice for contextual emotion detection. This sets the foundation for the next analytic component of understanding cognitive themes.

Cognitive Theme Extraction and Topic Representations (KeyBERT and BERTopic)

Cognitive themes represent the content of what clients reflect on. This is the issues, topics, meanings, and internal processes they describe. To extract these elements, PsychExtract evaluates two widely used NLP tools.

KeyBERT identifies keywords using cosine semantic similarity between the text and candidate n-grams (varied word length groupings) (Grootendorst, 2025), this is visualized in Figure 3. Because it relies on Sentence-BERT embeddings, it captures meaning beyond simple word counts and is transparent enough to be interpretable by therapists. This makes KeyBERT a suitable, explainable baseline for cognitive theme extraction.

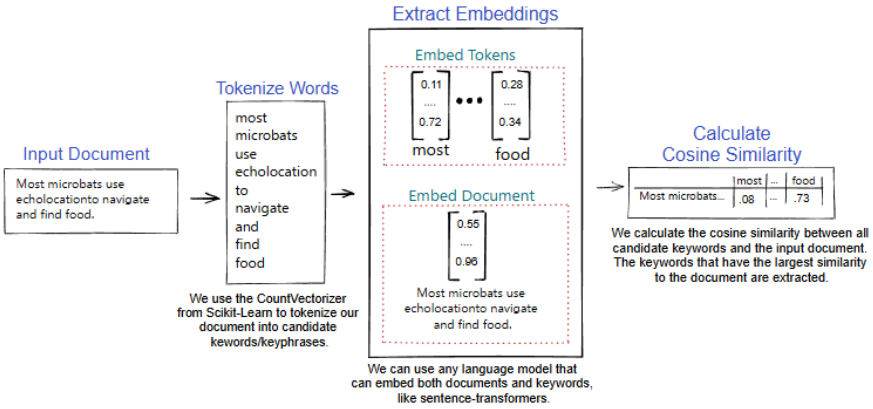


Figure 3. KeyBERT keyword extraction pipeline (Grootendorst, 2025).

However, KeyBERT provides surface-level patterns and cannot capture broader shifts in meaning across a document. To complement this, PsychExtract includes a comparison with BERTopic, which identifies themes using clustering and class-based term frequency (Grootendorst, 2022), this is illustrated in Figture 4. While more complex, BERTopic can represent broader reflective patterns that align with cognitive restructuring processes described in psychotherapy literature.

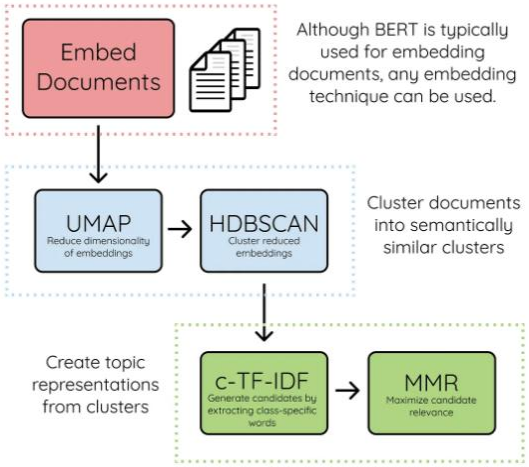


Figure 4. BERTopic topic modelling pipeline (Grootendorst, 2022).

This therefore connects the "what" of insight (cognitive content) with the "how" of extraction (topic modelling methods), completing the second component of insight analysis. The next step is to capture linguistic patterns associated with meaning-making.

Linguistic Pattern Analysis (LIWC)

The Linguistic Inquiry and Word Count (LIWC) framework categorises words into psychological dimensions such as cognitive processes, emotional tone, pronoun use, and insight-related terms (Pennebaker et al., 2015). Decades of studies demonstrate that these categories reflect internal cognitive states and are particularly relevant for detecting reflective thinking.

PsychExtract draws specifically on the cognitive mechanisms category. Words like "think," "realise," or "because" often signal reflective insight processes. However, LIWC is limited by its dictionary-based approach. It counts words without understanding context. This means it cannot distinguish between "I think" used casually versus reflectively.

To address this limitation, PsychExtract uses LIWC only for interpretability and theoretical grounding, while relying on contextual models (such as transformer-based NLP architectures) for the actual extraction pipeline. This hybrid approach supports interpretability without sacrificing nuance.

Having established how the system extracts early insight from text (emotion, cognitive themes, and linguistic patterns), the final section explains what text is fed into the system and how the results are returned to the user. This completes the OCR-NLP-TTS pipeline.

INPUT AND OUTPUT PROCESSING

OCR Requirements in Mental-Health Tools

In therapeutic settings, clients often maintain handwritten journals or written reflections. To analyse such inputs computationally, they must first be digitised using OCR, a technology that converts images of text into machine-readable characters.

Smith provides a foundational overview of Tesseract, one of the most widely used open-source OCR engines (Smith, 2007). Tesseract uses a multi-stage pipeline (concisely laid out in Figure 5) involving line detection, character segmentation, and language modelling to recognise text, even from noisy or imperfect inputs. However, Smith identifies two key limitations. For one, handwriting varies significantly between users, and for two, errors introduced by OCR can propagate into downstream NLP tasks, affecting emotion classification or topic modelling.

Figure 5. OCR process.

PsychExtract incorporates these findings by explicitly evaluating how OCR performance impacts the accuracy of insight-related NLP outputs. This extends prior OCR literature by shifting the focus from character-level accuracy to its influence on psychological inference quality. This is a critical factor in real-world mental-health tooling.

Output Processing: Text-to-Speech Synthesis of NLP Summaries

Once textual insight has been extracted, PsychExtract produces an accessible output for users. Shen et al. introduced Tacotron 2, a leading TTS model capable of generating highly natural-sounding audio using a sequence-to-sequence architecture and a neural vocoder (Shen et al., 2018). Their work demonstrated that TTS systems can reliably convert text into expressive speech. The pipeline is summarized in Figure 6.

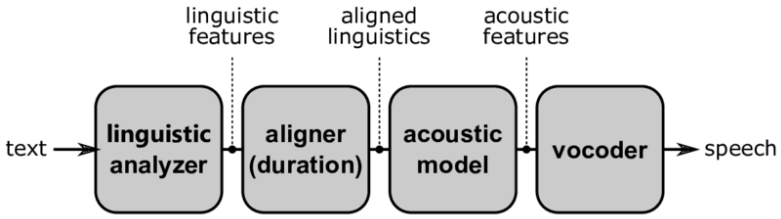


Figure 6. Conventional TTS pipeline representation.

In PsychExtract, TTS is used not for interaction but as an accessibility feature. The system reads out the generated summaries, emotional indicators, and cognitive themes for users who prefer auditory feedback or have reading difficulties. This completes the pipeline by providing an intuitive and inclusive output format.

Together, the OCR-NLP-TTS structure forms the full workflow. Handwritten or typed text is digitised; emotional, cognitive, and linguistic markers of early insight are extracted; and results are returned as both text and spoken summaries.

CRITICAL SYNTHESIS: OPERATIONALIZING INSIGHT RESPONSIBLY

The reviewed literature demonstrates that while NLP techniques can reliably extract linguistic signals associated with emotion, reflection, and thematic content, each approach is subject to important limitations. Emotion classification relies on probabilistic labels that simplify context-dependent and culturally mediated experiences; keyword extraction and topic modelling prioritise semantic salience over psychological coherence; and dictionary-based linguistic tools lack contextual sensitivity. Furthermore, many existing AI systems in mental-health contexts are constrained by real-time interaction requirements, limiting the complexity of language processing that can be deployed safely.

These limitations directly motivate the integrated, interpretability-focused design of PsychExtract. Rather than attempting to infer insight or psychological state directly, the system positions itself as an assistive analytic tool that surfaces multiple complementary indicators of early insight from reflective text. By combining emotion signals, cognitive themes, and linguistic markers within a non-conversational, offline pipeline, PsychExtract prioritises transparency, theoretical grounding, and human oversight. This modular, hybrid architecture reflects a deliberate shift away from diagnostic or interventionist claims toward a supportive, human-in-the-loop role that aligns with ethical guidance for AI use in therapeutic contexts (Luxton, 2014; Miner et al., 2016).

DESIGN (1945/2000 WORDS)

OVERVIEW

PsychExtract is designed as an interpretability-first, non-clinical analysis system to support structured reflection on emotionally expressive text. Its core objective is not to provide psychological judgement or intervention, but to assist users in identifying emotional patterns, salient themes, and linguistically meaningful signals. This aligns with research emphasising insight formation as central to psychological change while cautioning against automated clinical interpretation (Hill et al., 2007; Greenberg & Pascual-Leone, 2006; Luxton, 2014).

At a system level, PsychExtract transforms unstructured textual material (e.g., journals, reflective essays, therapy-adjacent notes) into structured, inspectable outputs that prompt reflection without

asserting authority. Rather than collapsing analysis into a single opaque prediction, the system exposes intermediate representations at each stage. This design reflects the subjective nature of emotional language and evidence that transparent, inspectable systems are more likely to be trusted and appropriately relied upon in mental-health-adjacent contexts (Doshi-Velez & Kim, 2017; Jacobs et al., 2021).

Several constraints are embedded in the design. First, PsychExtract is explicitly non-clinical: it does not infer diagnoses, assess risk, or generate recommendations. Outputs are framed descriptively and probabilistically, reflecting patterns in language rather than claims about mental state, consistent with ethical guidance for affective computing (Luxton, 2014; Mohammad, 2022). Second, the system is privacy-aware by design: processing occurs locally when possible, relying on pre-trained models, with no persistent storage of user text and transient intermediate artefacts. Third, the system prioritises interpretability over predictive optimisation. Transformer-based models are paired with linguistic explanation layers and conservative output framing, favouring clarity over marginal accuracy gains.

Together, these goals define the design space: assistive rather than authoritative, transparent rather than optimised, ethically bounded rather than expansive.

DESIGN PRINCIPLES

PsychExtract is guided by four principles: transparency, modularity, error tolerance, and user trust, embedded in both interface and architecture.

Transparency is achieved by exposing intermediate representations and avoiding opaque, end-to-end inference. Users can inspect OCR output, view emotion probabilities rather than categorical labels, and see which linguistic features contribute to summaries, favouring intelligible explanations over raw model introspection (Doshi-Velez & Kim, 2017; Mohammad, 2022).

PsychExtract is structured as loosely coupled components, each responsible for a specific data transformation, allowing alternative OCR engines, language models, or keyword extraction methods to be substituted without rearchitecting the system. This supports iterative development, comparative evaluation, and future extension, reflecting clinician-facing NLP best practices (Turner et al., 2022).

Given the noisy nature of reflective writing and OCR, the system tolerates errors rather than obscuring them. User verification checkpoints, parallel pathways, and probabilistic output framing limit the impact of misrecognition or misclassification. Avoiding hard thresholds reduces the risk of misleading outputs in uncertain cases.

Trust is treated as a design outcome. Clear non-clinical boundaries, minimal data retention, and descriptive rather than prescriptive outputs encourage appropriate reliance. Trust in AI-supported mental health tools is fostered when system limitations are legible and role boundaries respected (Jacobs et al., 2021, Luxton, 2014; Torous et al., 2018).

These principles guide domain framing, component selection, and architectural decisions.

DOMAIN- AND USER-FRAMING IN DESIGN

PsychExtract operates within the broader domain of mental health support, self-reflection, and therapeutic-adjacent technologies. Its focus is not on clinical intervention but on assisting reflective practices where emotional awareness and insight formation are valuable. Prior research highlights

the potential of structured language analysis to support psychological understanding while simultaneously emphasising the risks of over-automation and misinterpretation in mental health contexts (Inkster et al., 2020; Mohammad, 2022; Torous et al., 2018).

The system is intended for users who already engage in reflective writing but may benefit from additional structure and pattern visibility. These users include students completing reflective assignments, individuals who journal for emotional processing, and therapy-adjacent users who write between sessions or prepare material for discussion. Across these groups, the system assumes voluntary engagement and user ownership of interpretation. PsychExtract structures and summarises content supplied by the user without interpreting intent, assigning diagnoses, or offering recommendations, reinforcing its role as an assistive tool rather than an authority.

Several explicit assumptions and exclusions guide the design. Users are assumed to be reflective rather than crisis-seeking, and the system does not attempt to detect self-harm, suicidality, or acute distress. It does not provide real-time intervention, escalation pathways, or safeguarding mechanisms. While clinicians and researchers may find the system legible and methodologically informative, they are not positioned as primary end-users in this iteration. Maintaining these boundaries is a deliberate design decision intended to preserve trust, reduce ethical risk, and avoid role confusion, consistent with guidance on responsible AI use in mental health settings (Luxton, 2014; MHRA, 2025).

SYSTEM COMPONENTS

PsychExtract is organised as a modular pipeline composed of interoperable components (OCR, NLP with an interpretability layer, and TTS), each responsible for a specific transformation of the data. This architectural choice reflects findings from prior work showing that clinicians and users prefer systems that expose intermediate reasoning steps and allow scrutiny of automated outputs (Jacobs et al., 2021; Turner et al., 2022). It also supports comparative evaluation by allowing alternative models and techniques to be substituted without restructuring the entire system.

OCR

The OCR component converts scanned documents, photographs, or PDFs containing handwritten or typed text into machine-readable form. Two open-source OCR engines are considered: Tesseract and EasyOCR. Tesseract is a lightweight, widely used engine with strong performance on clean, printed text (Patel et al., 2012; Smith, 2007), while EasyOCR employs deep learning architectures that are more robust to noisy inputs and variable handwriting styles (JaidedAI, 2024). Comparing these engines allows assessment of robustness across the heterogeneous document quality typical of reflective writing.

The input to this component consists of image-based documents, while the output is plain text. OCR output is explicitly treated as provisional and is surfaced to the user for verification and correction before downstream processing. Integration at this stage is intentionally partial: when users provide typed text directly, the OCR component can be bypassed entirely, reinforcing flexibility in input modality and reducing unnecessary error propagation.

Emotion Classification

Emotion classification is central to the system's analytical goals, as emotional awareness and differentiation are closely linked to insight formation in psychotherapy and reflective practice (Hill et al., 2007, Greenberg & Pascual-Leone, 2006). PsychExtract employs pre-trained transformer-

based models fine-tuned for multi-label emotion classification using the GoEmotions dataset (Demszky et al., 2020).

Two models are explicitly considered: DistilBERT and RoBERTa. DistilBERT offers a compressed architecture that retains much of BERT's representational capacity while reducing computational cost (Sanh et al., 2019), making it suitable for resource-constrained environments. RoBERTa, by contrast, benefits from optimised pretraining strategies and typically achieves higher classification performance at the cost of increased computational demand (Liu et al., 2019) Comparing these models supports an explicit trade-off analysis between performance and feasibility.

The input to this component is cleaned, user-verified text, and the output consists of probabilistic multi-label emotion predictions. This is demonstrated in Figure 7. Outputs are framed as signals detected in language rather than definitive emotional states, reflecting both dataset limitations and ethical guidance discouraging overconfident affective inference (Mohammad, 2022).

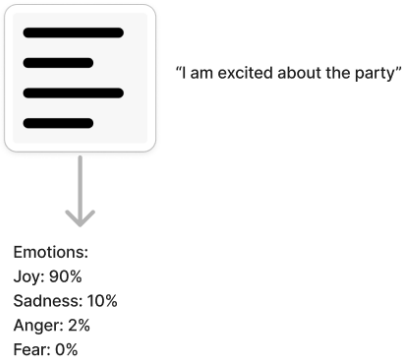


Figure 7. Emotion classification model input and output visualized.

Keyword and Linguistic Pattern Extraction

To contextualise emotion predictions, PsychExtract includes a linguistic feature extraction component that identifies salient words, phrases, or patterns associated with emotional signals. Embedding-based methods such as KeyBERT are employed to capture semantic relevance between text segments and extracted keywords (Grootendorst, 2025). Alternative approaches, including TF–IDF weighting and psychologically motivated lexical markers inspired by LIWC, provide more rule-based and clinically interpretable signals (Pennebaker et al., 2015).

The input to this component consists of text and, optionally, emotion classification outputs, while the output is a set of explanatory linguistic features. This pipeline is demonstrated in Figure 8. This component is only partially coupled to emotion classification: keyword extraction can operate independently, allowing alternative interpretability pathways and supporting error analysis when emotion predictions are uncertain.

Figure 8. Keyword and linguistic pattern extraction input and output visualized.

Interpretability Layer

The interpretability layer synthesises emotion predictions and linguistic features into human-readable explanatory artefacts. Rather than exposing internal model mechanics such as attention weights, which may be misleading or difficult to interpret reliably, this layer focuses on psychologically legible explanations grounded in observable language patterns. This design choice aligns with broader critiques of superficial explainability in machine learning and emphasises intelligibility over technical transparency (Doshi-Velez & Kim, 2017). Importantly, these explanations do not claim to reveal the 'true cause' of emotional expression, but rather offer plausible, inspectable mappings between language use and detected emotional signals

The output of this layer consists of structured explanations that link detected emotional signals to specific language features, reinforcing transparency and supporting appropriate user trust. This interpretability pipeline is visualized in Figure 9.

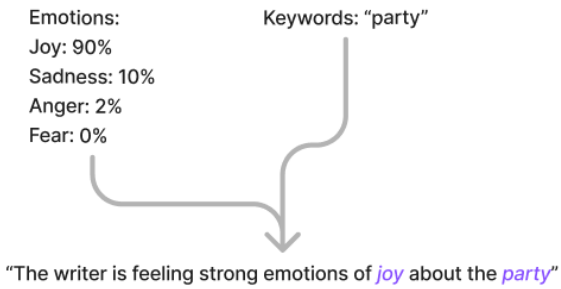


Figure 9. Interpretability layer inputs and output demonstrated.

Output Generation and Accessibility

The final stage of the pipeline generates concise textual summaries that highlight dominant emotional themes, recurring linguistic patterns, and notable shifts across the input text. These summaries are designed to prompt reflection rather than provide conclusions. An optional TTS component supports accessibility for users with reading, attentional, or cognitive challenges, consistent with evidence that TTS can enhance comprehension and engagement (Young et al., 2018).

Two TTS approaches are considered: neural synthesis via Coqui TTS, which offers high-quality speech generation (Coqui.ai, 2025), and pyttsx3, which provides a lightweight offline alternative suitable for constrained environments (pyttsx3.readthedocs.io, 2025). The UI at this stage remains deliberately minimal, prioritising clarity, editability, and progressive disclosure of information over visual richness.

ARCHITECTURE AND DATA FLOW

PsychExtract uses a modular, linear pipeline in which each stage is independently testable and replaceable. This maximises interpretability, facilitates comparative evaluation, and maintains clear traceability across the workflow. Data progresses through five main stages: OCR, emotion classification, linguistic interpretation extraction, summary generation, and optional TTS.

User Flow

Figure 10 illustrates the end-to-end user flow, beginning with raw document ingestion and culminating in a structured insight summary. The pipeline comprises five core computational stages, supported by user-mediated validation and optional output modalities.

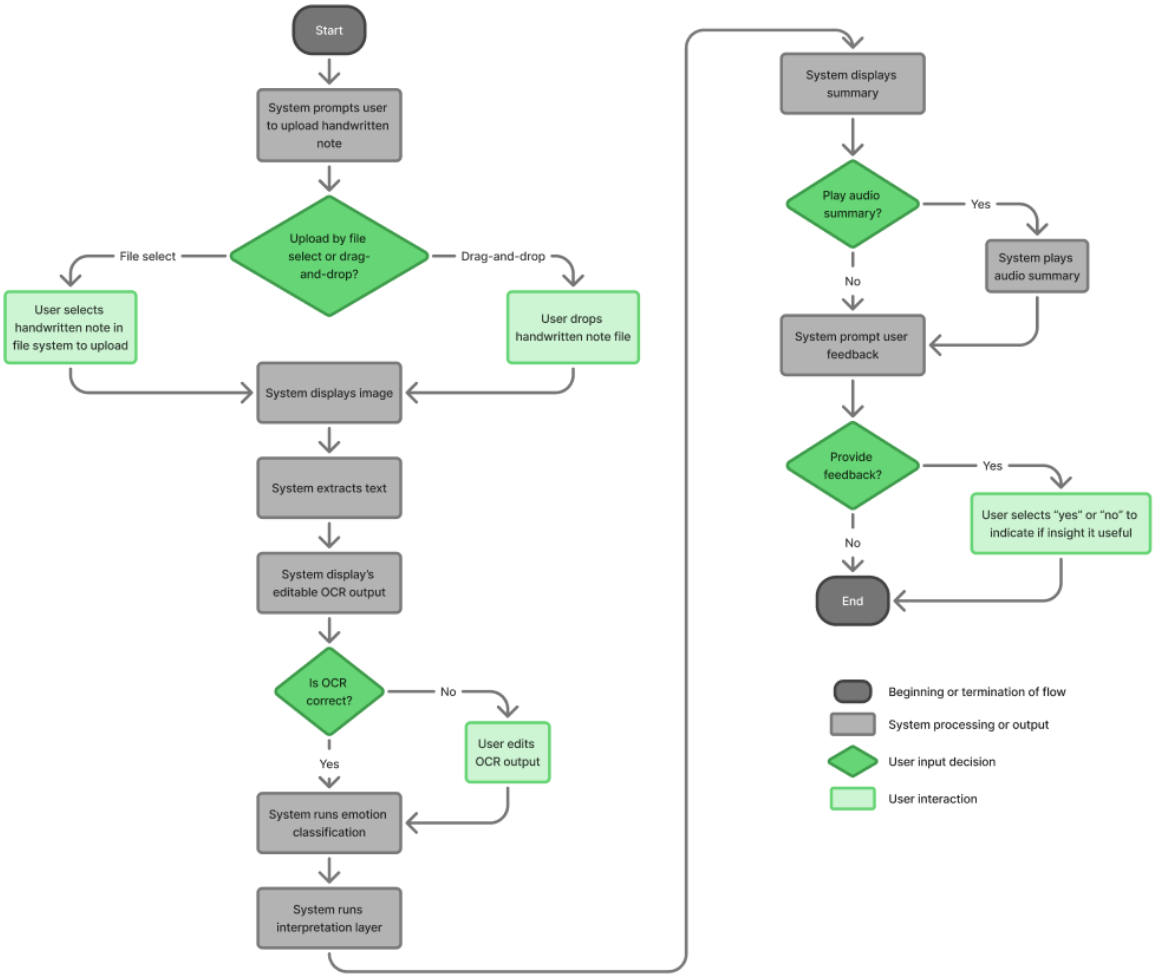


Figure 10. PsychExtract user flow.

The following is a summarized breakdown of the steps that the system undertakes:

1. Input Upload: user provides handwritten or scanned document
2. OCR: Tesseract/EasyOCR produces preliminary text
3. Editable Text: user verifies and corrects OCR output.
4. Emotion Classification: DistilBERT/RoBERTa produce multi-label probabilities
5. Interpretability Layer: KeyBERT or linguistic metrics extract explanatory features
6. Summary Generation: concise psychological insight produced
7. Optional TTS: pyttsx3/Coqui converts summary to speech
8. User Feedback Capture: supports iterative refinement and future evaluation

Early Prototype

To explore alternative interaction patterns without committing to full implementation, interfaces are tested in Figma. Figure 11 presents a low-fidelity Streamlit prototype that demonstrates document upload, OCR correction, and insight review workflows. User feedback capture is embedded throughout the interface to support iterative refinement and future evaluation, though this feedback loop is not yet fully integrated into model retraining.

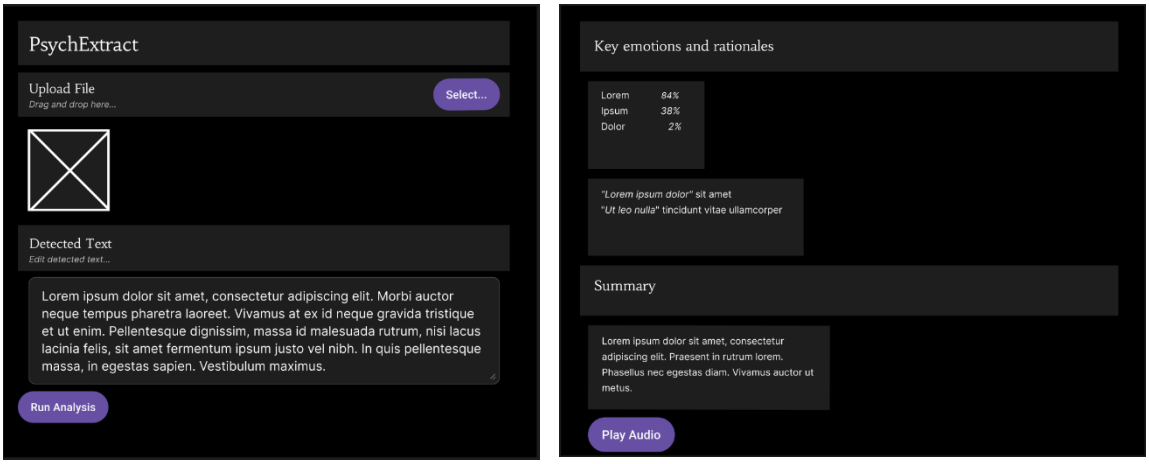


Figure 11. PsychExtract prototype interface.

Folder Structure

At the code level, the system is organised into discrete modules reflecting each pipeline stage (Table 1). This folder structure enforces separation of concerns, promotes traceability between components, and supports flexible model substitution during experimentation and evaluation.

Folder	Purpose
ocr/	OCR wrappers for Tesseract and EasyOCR
nlp/	DistilBERT and RoBERTa models
interpret/	KeyBERT and linguistic metrics
summary/	Summarization and TTS components
ui/	Streamlit interface
tests/	User and integration tests

Table 1. PsychExtract folder structure.

FEASIBILITY AND CONSTRAINTS

The design of PsychExtract is shaped by several practical constraints relating to technical feasibility, data availability, and project scope.

There are technical constraints present as the system relies on transformer-based language models, which impose computational costs in terms of memory usage and inference time. While models such as DistilBERT mitigate these demands relative to larger architectures, real-time processing on low-resource devices remains constrained. The decision to support local, offline processing further limits the use of large-scale or continuously updated models, but aligns with privacy and ethical goals (Luxton, 2014). Similarly, OCR performance is sensitive to document quality, handwriting variability, and image noise, constraining accuracy in unconstrained real-world inputs (Smith, 2007; JaidedAI, 2024).

Emotion classification is tested on the GoEmotions dataset, which, while large and fine-grained, is derived from curated online text rather than therapeutic or deeply personal writing. This exposes a data limitation where, as a result, emotion predictions may not fully capture the nuance of reflective or autobiographical language. Additionally, emotion labels represent perceived emotional content in text, not ground-truth internal states, reinforcing the need for cautious, probabilistic interpretation (Demszky et al., 2020; Mohammad, 2022).

As an individual academic project with time constraints present, PsychExtract is necessarily scoped to prioritise depth of reasoning over breadth of functionality. Advanced features such as longitudinal user modelling, adaptive feedback, or clinician-facing analytics are intentionally excluded in favour of a robust, interpretable core pipeline. This constraint supports methodological clarity and ethical containment while leaving clear pathways for future work.

IMPLEMENTATION (1994/2000 WORDS)

IMPLEMENTATION OVERVIEW

At the time of writing, PsychExtract is implemented as a modular, linear processing pipeline prioritising interpretability, transparency, and component-level traceability. The architecture favours independently testable stages over tightly coupled end-to-end models, enabling clear inspection of intermediate artefacts and outputs. Each component exposes well-defined input/output interfaces, allowing substitution or extension without affecting other subsystems.

The pipeline thusfar comprises four primary modules: (1) An OCR subsystem for handwritten or printed input. (2) An Emotion classification module using transformer-based NLP. (3) A keyword and linguistic pattern extraction component leveraging embedding-based and statistical methods. (4) A linguistic framing mechanism for generating structured reflective outputs.

This modular orchestration aligns with interpretable machine learning principles (Doshi-Velez & Kim, 2017; Jacobs et al., 2021), ensuring outputs remain user-legible and reproducible. All components are implemented in Python, with deep learning models executed via PyTorch and Hugging Face Transformers (Devlin et al., 2019). Intermediate artefacts are persisted as CSV and JSON files to enable systematic inspection and debugging.

OCR IMPLEMENTATION

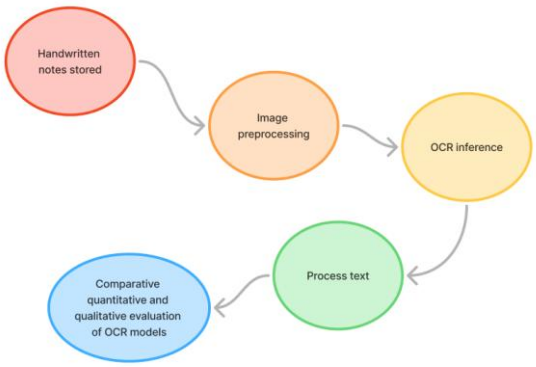


Figure 12. OCR system pipeline.

OCR subsystem pipeline (illustrated in Figure 12 above) implements a feasibility-oriented handwritten text recognition pipeline, combining classical and modern OCR approaches. The subsystem processes standardised input samples (see Figure 13), ensuring reproducible evaluation across all engines. Primary recognition engines include the rule-based baseline Tesseract OCR (Patel et al., 2012; Smith, 2007), the deep learning-based EasyOCR CNN–RNN architecture (JaidedAI, 2024), and additional exploratory models designed to address the baseline's character accuracy limitations: PaddleOCR (Cui et al., 2025), TrOCR (Li et al., 2021), and Qwen-VL 2.5 (Bai et al., 2024).

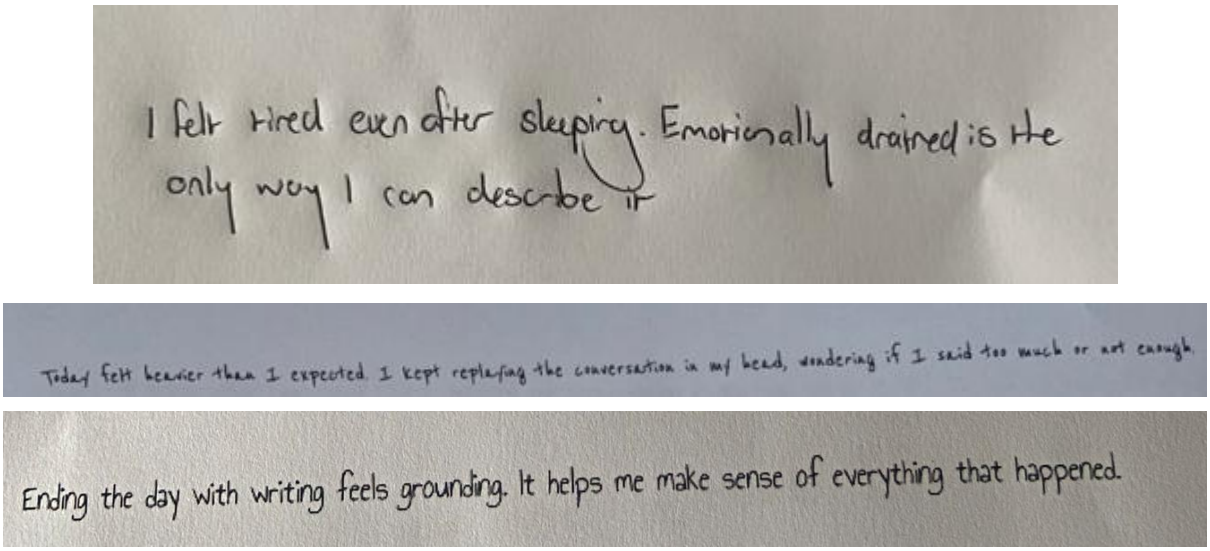


Figure 13. Example handwritten inputs used consistently across all OCR engines.

Image Preprocessing Pipeline

Images are preprocessed to enhance visibility while preserving handwriting structure. The preprocessing pipeline is illustrated in Figure 14, showing the steps executed programmatically:

```
def preprocess_image(img_path: str, upscale=2.0) -> None:
    """
    Preprocess a single image by applying various image processing techniques.

    :param img_path: Path to the input image file.
    :type img_path: str
    :param upscale: Factor by which to upscale the image.
    :type upscale: float
    """
    img_path = Path(img_path)

    # load with PIL (format-agnostic)
    im = Image.open(img_path)
    # fix EXIF orientation
1   im = ImageOps.exif_transpose(im)
    # convert to RGB
    im = im.convert("RGB")
    # convert to OpenCV format
    img = cv2.cvtColor(np.array(im), cv2.COLOR_RGB2BGR)
    # apply grayscale
2   gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    # contrast enhancement (CLAHE)
3   clahe = cv2.createCLAHE(clipLimit=1.2, tileGridSize=(16, 16))
    gray = clahe.apply(gray)
    # adaptive threshold
4   bw = cv2.adaptiveThreshold(gray, 255, cv2.ADAPTIVE_THRESH_GAUSSIAN_C,
                               cv2.THRESH_BINARY, 31, 51)
    # upscale
5   if upscale > 1:
        bw = cv2.resize(bw, None, fx=upscale, fy=upscale,
                        interpolation=cv2.INTER_CUBIC)

    # save as PNG with new name
    os.makedirs("preprocessed_imgs", exist_ok=True)
    output_name = f"preprocessed_imgs/{img_path.stem}_preprocessed.png"
    cv2.imwrite(str(output_name), bw)
```

Figure 14. Annotated code snippet of the OCR preprocessing pipeline, showing programmatic implementation of image preprocessing steps.

1. Orientation normalization (using metadata)
2. Grayscale conversion
3. Contrast enhancement via CLAHE (Zuiderveld, 1994)

4. Adaptive thresholding (Otsu, 1979)
5. Optional upscaling for resolution-sensitive engines (Plamondon & Srihari, 2000)

Preprocessed images are saved as artefacts to support reproducibility and allow visual inspection. An example input image before and after preprocessing is shown in Figure 15, illustrating contrast enhancement and background suppression applied prior to OCR.

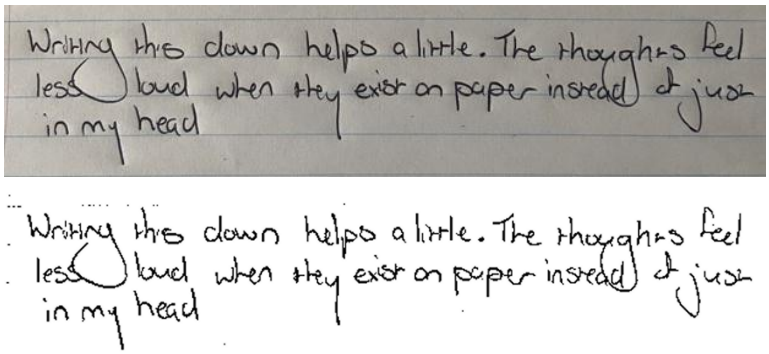


Figure 15. Example handwritten image before (top) and after preprocessing (bottom), demonstrating contrast enhancement and background suppression.

Optical Character Recognition Inference Pipeline

OCR is framed as mapping $I \in \mathbb{R}^{H \times W \times C}$ to a token sequence $Y = (y_1, \ldots, y_T)$, with transformer-based models estimating:

$$P(Y \mid I) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, E)$$

E represents embeddings from a vision encoder.

This formulation aligns with standard autoregressive transformer modelling (Vaswani et al., 2017) and has been adopted in OCR-specific architectures such as TrOCR (Li et al., 2021), demonstrating state-of-the-art handwritten and printed text recognition performance.

Large multimodal models, such as Qwen (Figure 16), require explicit GPU configuration to ensure efficient operation. The device_map="auto" setting automatically distributes model layers across available hardware, while half-precision (float16) reduces GPU memory usage without compromising output quality. For stable inference on these large models, a batch size of 1 is recommended. In contrast, smaller engines like TrOCR or EasyOCR are lightweight enough to run without special hardware configuration.

```
qwen_model_id = "Qwen/Qwen2.5-VL-7B-Instruct"

qwen_processor = AutoProcessor.from_pretrained(qwen_model_id)
qwen_model = AutoModelForVision2Seq.from_pretrained(
    qwen_model_id,
    torch_dtype=torch.float16,
    device_map="auto"
)
```

Figure 16. Code snippet of loading a pretrained Qwen multimodal model with GPU support.

Instruction-based models (e.g., Qwen) are prompted to transcribe exactly without hallucination. Outputs are saved as plain-text files, with lightweight post-processing applied (case normalisation, punctuation removal, whitespace standardisation) to maintain fidelity for downstream analysis.

EMOTION DETECTION IMPLEMENTATION

Model Selection and Rationale

The emotion analysis module employs two pretrained transformer-based multi-label classifiers: DistilBERT and RoBERTa. DistilBERT provides a compact, efficient architecture derived from BERT (Sanh et al., 2019), while RoBERTa offers enhanced representational capacity through robust pretraining strategies (Liu et al., 2019).

Specifically, the project uses the pretrained model bhadresh-savani/distilbert-base-uncased-emotion (Bhadresh Savani, 2022) for DistilBERT which is already fine-tuned on GoEmotions (Demszky et al., 2020) which aligns with the project's six-emotion framework.

cardiffnlp/twitter-roberta-base-emotion-multilabel-latest (CardiffNLP, 2023) is used for RoBERTa, which supports a broader emotion taxonomy and has already been fine-tuned on Twitter data (Barbieri et al., 2020) (Figure 17).

Both models are accessed via the Hugging Face Transformers library (Hugging Face, 2026) and are used in inference-only mode, reflecting the project's emphasis on system integration, interpretability, and reproducible output behaviour.

DistilBERT:	RoBERTa:
anger	anger
fear	anticipation
joy	disgust
love	fear
sadness	joy
surprise	love
	optimism
	pessimism
	sadness
	surprise
	trust

Figure 17. Emotion taxonomy comparison.

Batching and Inference Timing

To balance computational efficiency and reproducibility, the emotion analysis module executes inference in batches. A typical batch size of 16 was chosen to leverage GPU parallelism without exceeding memory constraints. Smaller batches reduce GPU memory usage but increase total runtime, while larger batches risk out-of-memory errors on consumer hardware. The batching algorithm is illustrated in Figure 18.

```python
with torch.no_grad():
    for batch in dataloader:
        batch = {k: v.squeeze(1).to(device) for k, v in batch.items()}
        outputs = model(**batch)
        logits = outputs.logits
        probs = torch.sigmoid(logits)
        all_probs.append(probs.cpu().numpy())
```

Figure 18. Batched inference loop for emotion classification, showing how tokenised inputs are processed on the device and probabilities are collected per batch.

Inference timing was systematically recorded for all models using a dedicated evaluation function (run_inference_with_timing). For each model, the function measures total runtime, computes average time per sample, and logs device, batch size, and number of samples. This procedure ensures transparency of performance and supports reproducibility, particularly when comparing DistilBERT (efficient, compact) and RoBERTa (larger, more expressive). Figure 19 shows an annotated code snippet of the timing function.

```python
start_time = time.perf_counter()

with torch.no_grad():
    for batch in dataloader:
        batch = {k: v.squeeze(1).to(device) for k, v in batch.items()}
        outputs = model(**batch)
        logits = outputs.logits
        probs = torch.sigmoid(logits)
        all_probs.append(probs.cpu().numpy())

end_time = time.perf_counter()

probs = np.vstack(all_probs)

preds = (probs >= threshold).astype(int)

timing = {
    "model": model_name,
    "device": str(device),
    "batch_size": batch_size,
    "num_samples": n_samples,
    "total_inference_time_sec": round(end_time - start_time, 3),
    "avg_time_per_sample_ms": round(
        (end_time - start_time) / n_samples * 1000, 3
    )
```

Figure 19. Annotated code snippet of run_inference_with_timing, showing timing measurement and per-sample latency computation.

Input Processing and Output Structure

Input text $X = (x_1, \ldots, x_n)$ is tokenised using the corresponding pretrained tokenizer and truncated to a fixed maximum sequence length. Emotion probabilities for each label are computed by applying a sigmoid activation over the model logits, a standard approach in multi-label classification (Goodfellow, Bengio & Courville, 2016; Vaswani et al., 2017; Li et al., 2021):

$$\hat{y}_i = \sigma(\text{logits}_i), i \in \{1, \ldots, K\}$$

K is the number of emotion labels,
$\hat{y}_i$ is the predicted probability for emotion i,
and σ denotes the sigmoid function.

For each input, the system stores per-label probability scores alongside thresholded binary indicators (threshold = 0.5). All outputs are saved in CSV format, supporting downstream inspection, aggregation, and visualisation.

Thresholding and Aggregation Decisions

Binary predictions are derived using a fixed threshold ($\theta = 0.5$) for consistency across models. For models producing a broader label set (e.g., RoBERTa), predictions are aggregated into core categories using a max-confidence strategy (Tsoumakas, Katakis & Vlahavas, 2007; Zhang & Zhou, 2014), preserving the strongest signal while avoiding dilution from averaging.

$$\hat{y}_{\text{core}} = max\ (\hat{y}_{\text{mapped}})$$

$\hat{y}_{\text{mapped}}$ are all RoBERTa labels corresponding to a single core emotion.

This preserves the strongest signal while avoiding dilution through averaging.

KEYWORD AND LINGUISTIC PATTERN EXTRACTION IMPLEMENTATION

Keyword Extraction Strategy and Rationale

PsychExtract complements emotion classification with interpretable keyword signals extracted from reflective text. The system emphasises traceable, surface-level concepts rather than latent topic modelling, prioritising transparency and semantic legibility to support user-led inspection of reflective content (Angelov & Soares, 2021; Bojanowski et al., 2017).

Two unsupervised extraction methods are used in parallel (Figure 20), illustrating the shared input text and independent scoring paths: KeyBERT leverages sentence embeddings (Reimers & Gurevych, 2019) to rank unigrams and bigrams by cosine similarity to the document embedding (Grootendorst, 2025). YAKE relies on statistical and positional features to score candidate keywords without external corpora (Campos et al., 2020).

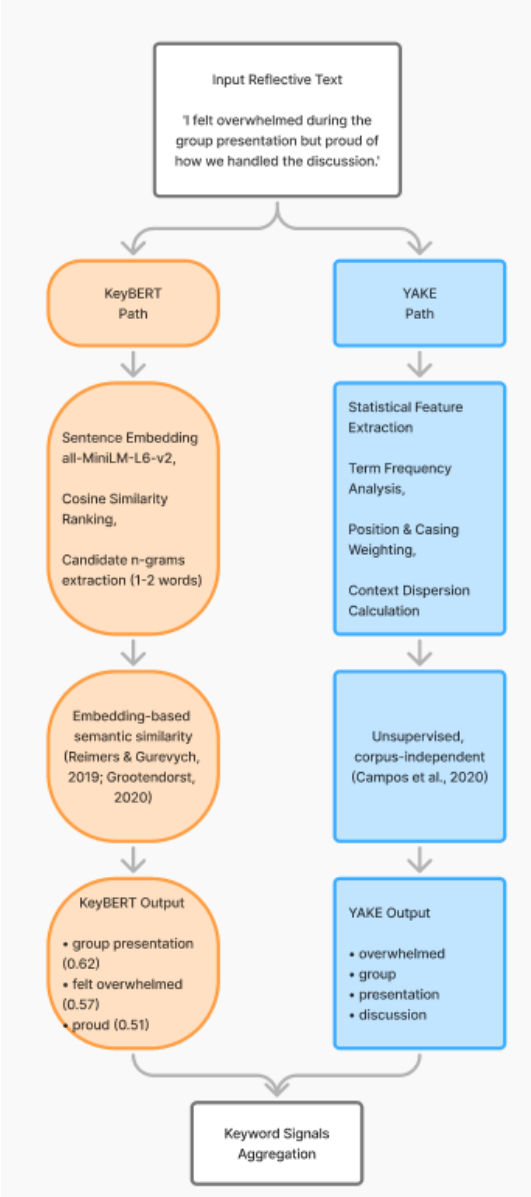


Figure 20. Parallel keyword extraction pipeline using KeyBERT (semantic similarity) and YAKE (statistical scoring).

Extraction and Filtering

KeyBERT and YAKE outputs are processed to remove stopwords and truncated to the top-ranked candidates. To ensure linguistic validity, phrases are filtered via spaCy's POS tagging and dependency parsing (Honnibal et al., 2020). Allowed patterns include:

NOUN,
PROPN,
ADJ-NOUN,
ADJ-PROPN

These patterns were selected to prioritise concept-bearing phrases while excluding function-heavy or syntactically incomplete spans.

For each candidate, the head noun lemma is extracted and used to resolve redundancy. If multiple phrases share the same head noun, only the earliest-ranked is retained. This ensures unique, interpretable concepts while preserving surface form. The noun-phrase filtering and head-noun deduplication process is illustrated in Figure 21.

```python
def is_valid_noun_phrase(phrase):
    doc = nlp(phrase)
    tokens = [t for t in doc]
    pos_pattern = tuple(t.pos_ for t in tokens)
    return pos_pattern in ALLOWED_PATTERNS
```

```python
def get_head_noun_lemma(phrase):
    doc = nlp(phrase)
    for token in doc:
        if token.dep_ == "ROOT" and token.pos_ in ("NOUN", "PROPN"):
            return token.lemma_
    return None
```

```python
def select_best_noun_phrases(keywords):
    concepts = {}

    for phrase in keywords:
        if not is_valid_noun_phrase(phrase):
            continue

        head = get_head_noun_lemma(phrase)
        if not head:
            continue

        if head not in concepts:
            concepts[head] = phrase

    return [v for v in concepts.values()]
```

Figure 21. Code snippet of noun phrase filtering and head-noun redundancy handling.

LINGUISTIC TEMPLATING IMPLEMENTATION

Template-Based Framing Rationale

PsychExtract generates structured interpretive statements from journal text using controlled linguistic templates, avoiding free-form generative models while preserving human-readable insights. Rather than producing psychological judgements, categories function as descriptive signals grounded in observable emotional and linguistic patterns. This conservative, non-clinical framing prioritises interpretability and traceability over generative flexibility (Bojanowski et al., 2017).

Lightweight Linguistic Signal Detection

Interpretive insight categories are operationalised through a combination of emotion probability outputs and surface-level lexical cues. The complete set of categories and their corresponding detection rules is defined explicitly to ensure transparent and reproducible behaviour.

The detection criteria for each category are summarised in Table 2, which serves as the conceptual specification for the templating logic.

Category	Detection Logic
Emotional Load	$mean(p_{\text{sadness}} + p_{\text{fear}}) > 0.6$
Emotional Clarity against Ambiguity	Presence of UNCERTAINTY_PHRASES
Regulation and Coping	Presence of COPING_VERBS
Arousal or Restlessness	$p_{\text{anger}} + p_{\text{fear}} > 0.6$ or SOMATIC_TERMS present
Self-Relation and Appraisal	Presence of SELF_REFLECTIVE_PHRASES

Table 2: Interpretive signal categories and their associated detection criteria.

For each journal entry x, the emotion classifier produces a probability vector:

$$\mathrm{e}(x) = [p_{\text{sadness}}, p_{\text{joy}}, p_{\text{love}}, p_{\text{anger}}, p_{\text{fear}}, p_{\text{surprise}}]$$

For each category c, detection logic is defined as:

$$\text{detect}_c(x) = \begin{cases} 1 & \text{if condition}_c(x) \text{ is met} \\ 0 & \text{otherwise} \end{cases}$$

Detection conditions may involve thresholded functions over the emotion probability vector (e.g., mean negative affect for Emotional Load) or the presence of predefined lexical markers (UNCERTAINTY_PHRASES, COPING_VERBS, SOMATIC_TERMS, SELF_REFLECTIVE_PHRASES).

The implementation of this detection logic, directly reflecting the criteria defined in Table 2, is illustrated in Figure 22, which shows how emotion outputs and lexical features are combined to activate interpretive categories.

```python
# lightweight linguistic detectors
UNCERTAINTY_PHRASES = [
    "not sure", "hard to name", "can't explain", "cannot explain",
    "something is there", "difficult to explain"
]

COPING_VERBS = [
    "writing", "reflecting", "reflection", "breathing",
    "grounding", "sitting with", "slowing"
]

SOMATIC_TERMS = [
    "body", "shoulders", "breathing", "tight",
    "tense", "restless", "tension"
]

SELF_REFLECTIVE_PHRASES = [
    "i noticed", "i realized", "i caught myself",
    "pattern in my reactions", "i keep noticing"
]
```

```python
def detect_insights(row):
    emotions = get_emotion_probs(row)
    text = row["text"]

    insights = []

    # Emotional Load
    mean_neg = (emotions["sadness"] + emotions["fear"]) / 2
    if mean_neg > 0.6:
        insights.append("Emotional Load")

    # Emotional Clarity vs Ambiguity
    if count_any(text, UNCERTAINTY_PHRASES) >= 1:
        insights.append("Emotional Clarity vs Ambiguity")

    # Regulation & Coping
    if contains_any(text, COPING_VERBS):
        insights.append("Regulation & Coping Mode")

    # Arousal / Restlessness
    if (emotions["fear"] + emotions["anger"] > 0.6 or
        contains_any(text, SOMATIC_TERMS)):
        insights.append("Arousal / Restlessness Level")

    # Self-Relation & Appraisal
    if contains_any(text, SELF_REFLECTIVE_PHRASES):
        insights.append("Self-Relation & Appraisal")

    return insights
```

Figure 22. Code snippet illustrating rule-based insight category detection from emotion probabilities and lexical markers.

Template-Based Insight Generation

Once one or more interpretive categories are detected, structured insight statements are generated using controlled linguistic templates, shown in Figure 23. Each template is populated with thematically extracted keywords to maintain semantic grounding while avoiding free-form generation.

```
TEMPLATES = {
  "Emotional Load": [
    "This entry suggests a relatively high emotional load, particularly in relation to {theme}.",
    "The overall tone of this entry indicates emotional heaviness connected to {theme}."
  ],
  "Emotional Clarity vs Ambiguity": [
    "The feelings described here appear difficult to clearly define, especially around {theme}.",
    "This entry reflects some uncertainty or ambiguity in how emotions related to {theme} are understood."
  ],
  "Regulation & Coping Mode": [
    "This entry highlights an active attempt to regulate emotions through reflection, particularly in response to {theme}.",
    "The writer appears to be engaging in a coping process while thinking about {theme}."
  ],
  "Arousal / Restlessness Level": [
    "The language used suggests heightened internal activation or restlessness related to {theme}.",
    "This entry reflects a state of tension or agitation associated with {theme}."
  ],
  "Self-Relation & Appraisal": [
    "This entry shows reflective self-evaluation in relation to {theme}.",
    "The writer appears to be assessing their own reactions or patterns while considering {theme}."
  ]
}
```

Figure 23. Linguistic templates.

Let $K(x)$ denote the set of extracted keywords for journal entry x, obtained via YAKE and KeyBERT. A thematic phrase is constructed as:

$$\text{theme}(x) = \begin{cases} \text{“the described experience”} & |\, K(x)\,| = 0 \\ \text{“their ”} + k_1 & |\, K(x)\,| = 1 \\ \text{“their ”} + k_1, \ldots, \text{and } k_n & |\, K(x)\,| > 1 \end{cases}$$

To reduce repetitive phrasing while preserving controlled output, a small pool of templates is defined per category, from which one is selected at random. Each generated insight is returned alongside its category label, ensuring traceability between detected signals and generated statements.

The template instantiation process is shown in Figure 24, which presents a representative code example of insight generation conditioned on detected categories.

```
def generate_insight_sentences(row):
  themes = select_theme_keywords(row)
  theme_text = format_theme(themes)

  categories = detect_insights(row)
  outputs = []

  if not categories:
    outputs.append({
      "category": "No Insights Detected",
      "text": "No significant emotional insights were detected in this entry."
    })
    return outputs

  for cat in categories:
    template = random.choice(TEMPLATES[cat])
    outputs.append({
      "category": cat,
      "text": template.format(theme=theme_text)
    })

  return outputs
```

Figure 24. Code snippet demonstrating category-conditioned template population.

Finally, Figure 25 provides an end-to-end illustrative example, showing the relationship between raw text input, detected themes, activated categories, and the resulting templated insight.

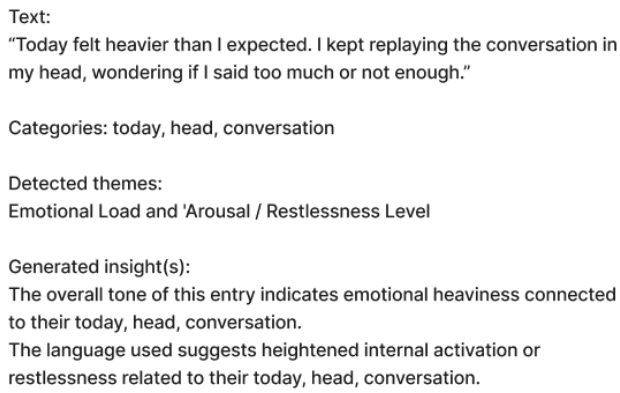


Figure 25. Textual example of the text inputs and detected thematic outputs with template generated insight.

INTEGRATION

Current Integration

At the time of writing, partial integration between system components has been achieved. Linguistic pattern extraction and framing logic operate on the persisted outputs of both the emotion classification module and the keyword extraction pipeline. Emotion probabilities and filtered keyword sets are merged at the data level, enabling joint analysis and interpretive sentence generation.

This integration currently uses CSV-based artefacts rather than a unified runtime pipeline, supporting systematic evaluation, debugging, and inspection of intermediate outputs during development.

Partially Integrated Components

Although OCR, emotion analysis, and linguistic framing dependencies are defined, the system does not yet operate in a fully sequential, per-document mode. OCR outputs, emotion predictions, and keyword extraction run as separate batch processes, with intermediate results stored to disk.

This separation reflects the exploratory, research-oriented stage rather than a conceptual limitation. Components are designed to support single-document processing once unified orchestration is introduced.

Planned Integration

Future integration work will focus on consolidating the pipeline into a document-centric processing flow, where a single input passes sequentially through OCR (if applicable), emotion classification, keyword extraction, and linguistic framing within a single execution context, as seen in Figure 26. This will enable more natural end-to-end evaluation and facilitate user-facing interaction.

The current modular design ensures that this transition can be achieved without substantial architectural refactoring, preserving the interpretability and traceability principles established throughout the system.

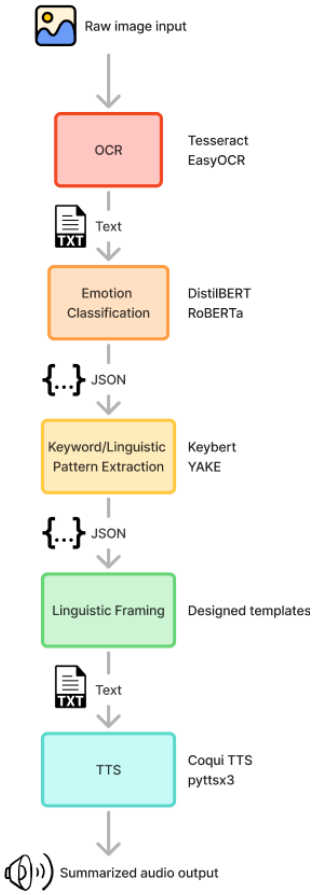


Figure 26. Conceptual end-to-end flow illustrating planned unified document processing.

EVALUATION (2169/2500 WORDS)

PsychExtract's evaluation is explicitly guided by its central aim: supporting structured, interpretable reflection on handwritten text without engaging in clinical judgement or recommendation. Each stage of the pipeline was assessed not only for technical performance but also for its capacity to:

> Preserve emotional and cognitive cues,
> facilitate insight formation,
> and maintain transparency for the user.

The evaluation strategy therefore encompasses both quantitative metrics (e.g., OCR character accuracy, emotion classification F1 scores, keyword extraction relevance) and qualitative measures (user perception, interpretability, and reflective usefulness), consistent with human-centred design principles (Jacobs et al., 2021; Young et al., 2018).

PRELIMINARY USER DISCOVERY

To inform the design and evaluation of PsychExtract, a small-scale user discovery exercise was conducted with participants at different stages of psychology study. The goal was to understand current challenges in reflective writing analysis, identify features considered most valuable, and surface potential concerns prior to full system implementation. This approach aligns with best practices in user-centred design, particularly for tools supporting emotional and reflective writing,

where early feedback can guide feature prioritisation and interface design (Gulliksen et al., 2003; Norman, 2013).

Participants (n = 3) reported varying degrees of experience with reflective writing and associated tasks (Table 3). One participant indicated regular engagement with therapy notes or case studies, while another occasionally interacted with reflective texts, and one had no direct experience. Across participants, all but one acknowledged occasional difficulty in extracting key emotional themes or insights from written material. This finding highlights the practical challenge PsychExtract aims to address: supporting users in identifying salient affective and thematic content efficiently.

Participant	**Level of Study**	**Experience with Reflective Texts**	**Struggled to Extract Key Themes?**	**Features Considered Useful**	**Concerns**
1	Honours	No	Yes	Handwriting-to-text; Emotion Analysis; Highlighting Key Themes	None
2	Undergraduate	Yes	Sometimes	Handwriting-to-text; Emotion Analysis; Highlighting Key Themes; Audio Read-back	Security
3	Undergraduate	Sometimes	No	Highlighting Key Themes	Potential inaccuracies of AI outputs

Table 3: Summary of preliminary user discovery responses regarding experience, challenges, and feature preferences.

When asked which system features they would find useful, participants consistently identified handwriting-to-text conversion, emotion analysis, and highlighting of key themes. One participant additionally suggested an audio read-back function to facilitate engagement with extracted insights. These responses validate the inclusion of both linguistic and emotional analysis components within PsychExtract and indicate early alignment with user needs.

Potential concerns were also reported. Security and privacy considerations were highlighted by one participant, while another noted the possibility of inaccurate or misleading analysis, reflecting awareness of AI limitations in interpretive tasks. No participant indicated missing or superfluous features beyond these considerations, suggesting the initial feature set was generally appropriate.

Overall, this preliminary user discovery demonstrates early support for PsychExtract's intended functionality. It also provides valuable guidance for system development, including the prioritisation of interpretable insights, transparent output presentation, and cautious framing of predictive content to mitigate user concerns regarding accuracy and trust. Incorporating such early-stage feedback is consistent with iterative, user-centred approaches that improve both usability and adoption in reflective or therapeutic contexts (Gulliksen et al., 2003; Norman, 2013).

OPTICAL CHARACTER RECOGNITION (OCR) EVALUATION

OCR forms the foundational stage of PsychExtract; transcription quality directly affects downstream emotion classification, keyword extraction, and linguistic templating. Errors at this stage propagate, reducing both interpretability and reflective usefulness. Therefore, evaluation prioritised character-level fidelity, preservation of sentence and semantic structure, and usability for human-assisted insight extraction.

Evaluation Methodology

Five OCR approaches were assessed: Tesseract, EasyOCR, PaddleOCR, TrOCR, and Qwen. Performance was evaluated using character-level accuracy (CA) and word error rate (WER), two standard metrics in OCR and speech recognition evaluation (Morris, Maier and Green, 2004; Rice, Jenkins & Nartker, 1993; Smith, 2007). These are defined as:

$$\text{CA} = \frac{\text{Correct Characters}}{\text{Total Characters}} \times 100, \qquad \text{WER} = \frac{S + D + I}{N} \times 100$$

S is substitutions,
D deletions,
I insertions,
and N the total number of words in the reference text.

These metrics provide a quantitative view of transcription fidelity, while qualitative inspection assessed readability, sentence coherence, and emotional/semantic preservation.

Quantitative Results

Table 4 summarises model performance across the evaluated handwritten journal samples.

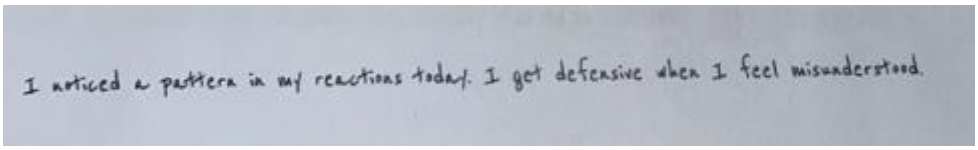
Model	Character Accuracy (%)	WER (%)
Tesseract	6	78
EasyOCR	12	72
PaddleOCR	20	65
TrOCR	68	15
Qwen	92.8	1.1

Table 4. OCR model performance.

Transformer-based methods, particularly TrOCR and Qwen, substantially outperformed traditional OCR engines. Qwen achieved 92.8% CA and 1.1% WER, preserving sentence flow and word order, while TrOCR showed moderate improvements but remained sensitive to handwriting variability. Traditional engines yielded fragmented, largely unreadable outputs, unsuitable for downstream NLP or human-assisted review.

Qualitative Insights

Figures 27 and 28 illustrate representative outputs. Traditional OCR results are highly fragmented, whereas TrOCR generates readable fragments for clear handwriting but deteriorates with complex entries. Qwen consistently produces coherent, structured transcriptions, enabling reliable extraction of emotional and thematic signals.



Tesseract: L neticed a pattern @ my reactions ada. Le get Arkensive then 1 Feel wisendee steed,

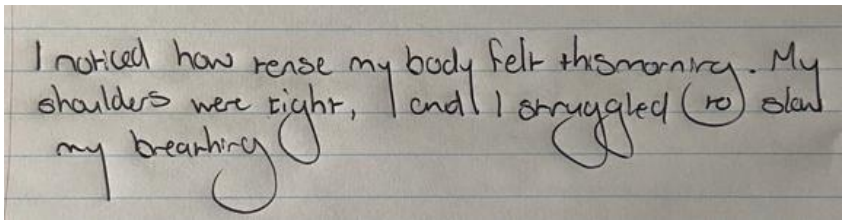
EasyOCR: 1 Ared putecr "f 'ato} Xasc4: 1 p6r dzfensin #le^ 1 fccl miceadzerfud

PaddleOCR: 1 natied per i f eat tdf. I gt defenie n I fee isdeesed.

TrOCR: I arrived a pattern in my reactions today . In yet defensive when I feel misunderstood

Qwen: I noticed a pattern in my reactions today. I get defensive when I feel misunderstood.

Figure 27. Text 7a with OCR outputs across models.



Tesseract: Vargruh how rease my body Falk thsmeraica . M elves woe tune, | bo } ov vole (ro) clad 4 enien Ch

EasyOCR: ~oked hxw rease ~4 bocly fel+ +hssrorar M shaldus nec richt 1 cnd" 1 8-x4Gled + sky bearwcCA

PaddleOCR: I noticd how rerse my body fel thsoniy. My shauldes wee cihr, I cndlI or uygled(ro)oln branhirc

TrOCR: page in the

Qwen I noticed how tense my body felt this morning. My shoulders were tight, and I struggled to slow my breathing.

Figure 28. Text 1b with OCR outputs across models.

Character-level error heatmaps (Figure 29) further highlight Qwen's robustness in reducing substitutions, deletions, and insertions across varying handwriting styles.

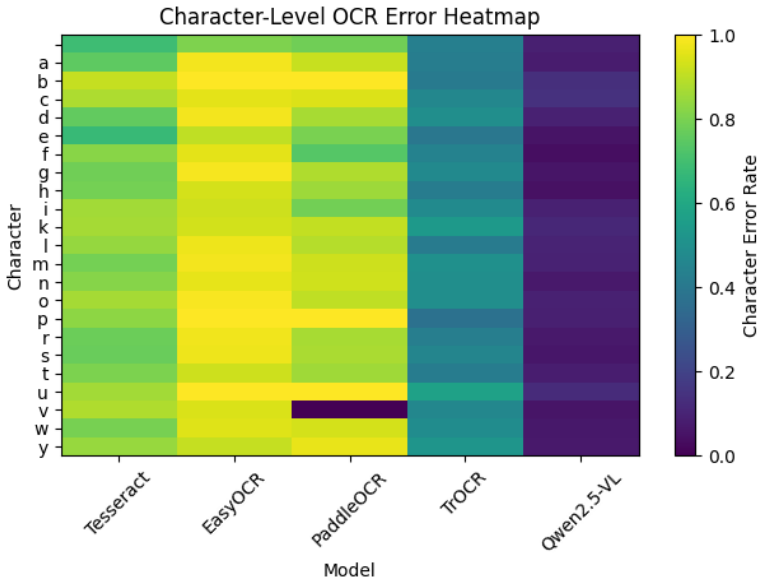


Figure 29. Character-level OCR error heatmap.

Semantic Preservation and Reflective Suitability

Beyond raw metrics, OCR evaluation emphasised contextual and emotional meaning preservation. Traditional OCR outputs frequently omit modifiers and connectors, while TrOCR occasionally misrepresents sentence intent. Qwen best preserves semantic structure and affective content, producing outputs suitable for human review and downstream NLP analysis, aligning with PsychExtract's assistive, non-diagnostic ethos.

Human oversight remains critical: the system deliberately avoids automated interpretation, reinforcing ethical boundaries between transcription and clinical decision-making (Amershi et al., 2019). Minor hallucinations and sensitivity to image quality underscore the prototype's role as a facilitative layer rather than a standalone system.

Implications for PsychExtract

High-fidelity transcription is essential for emotion and keyword extraction, and for subsequent linguistic templating. Qwen's outputs ensure that downstream modules operate on accurate, readable text, preserving the subtleties of reflective journal entries. Transformer-based OCR thus emerges as a critical enabler of interpretability, ensuring that computational analyses support meaningful reflective insight rather than being constrained by poor transcription quality.

EMOTION CLASSIFICATION

Emotion classification evaluates whether the system can identify salient affective content within reflective writing, supporting insight formation. Transformer-based models, DistilBERT and RoBERTa, were assessed under zero-shot inference conditions, reflecting realistic constraints for early-stage system integration and avoiding fine-tuning biases (Demszky et al., 2020; Devlin et al., 2019). Both quantitative metrics and qualitative signals were analysed to align evaluation with the project's non-clinical, interpretability-focused aims.

Quantitative Evaluation

Evaluation employed precision, recall, and F1-score per emotion class i:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \qquad \text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \qquad \text{F1}_i = 2 \times \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

The macro F1 was computed as the unweighted mean across all C emotion classes:

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C} \text{F1}_i$$

Macro F1 prioritises balanced performance across frequent and rare emotions, crucial for reflective journals where ambiguous affective content is common (Demszky et al., 2020). The micro F1 measures overall predictive quality without weighting per class.

On a GoEmotions subset, RoBERTa consistently outperforms DistilBERT across all overall metrics (Table 5). While DistilBERT exhibits lower precision, its recall remains moderate, indicating that it captures most emotions but with more false positives. RoBERTa balances precision and recall better, reflecting stronger selective recognition of contextually grounded emotions, though at higher computational cost. CPU-based inference latency revealed that RoBERTa was roughly 1.9× slower than DistilBERT, reflecting a trade-off between interpretability and responsiveness for interactive pipeline use.

Model	Macro Precision	Macro Recall	Macro F1	Micro Precision	Micro Recall	Micro F1	Avg. CPU Inference Time (ms/sample)
DistilBERT	0.414	0.606	0.436	0.357	0.628	0.455	15
RoBERTa	0.519	0.703	0.578	0.492	0.731	0.588	28

Table 5. Emotion classification performance.

Across all emotion classes, as seen in Table 6, RoBERTa consistently achieves higher F1 scores than DistilBERT, indicating more balanced and reliable detection of both frequent and infrequent emotions. DistilBERT tends to overpredict certain emotions such as joy and anger, resulting in moderate F1 values, while lower-frequency emotions like surprise and love are detected less reliably. RoBERTa demonstrates stronger recognition of subtle and contextually embedded emotions, which supports more accurate downstream linguistic templating and interpretive insight generation in reflective journal analysis.

Emotion	DistilBERT F1	RoBERTa F1
Sadness	0.514	0.630
Joy	0.459	0.538
Anger	0.478	0.607
Fear	0.346	0.446
Surprise	0.315	0.519
Love	0.501	0.724

Table 6. Individual emotion F1 scores.

Confidence Spread and Interpretive Signals

Quantitative metrics such as precision, recall, and F1-score provide valuable information about model performance, but they do not fully capture the interpretive utility of emotion classification for reflective writing. To address this, a confidence spread measure was computed for each journal entry i, defined as the difference between the highest and lowest predicted probabilities across all emotion classes k (Doshi-Velez & Kim, 2017; Mohammad, 2018):

$$\Delta_i = \max_k (p_{ik}) - \min_k (p_{ik})$$

p_{ik} is the predicted probability for emotion k in entry i.

Narrow confidence spreads indicate semantically coherent, emotionally clear entries, whereas diffuse spreads highlight reflective ambiguity. These values are subsequently leveraged in the linguistic templating stage to guide template selection and encourage deeper reflection, rather than signalling model uncertainty.

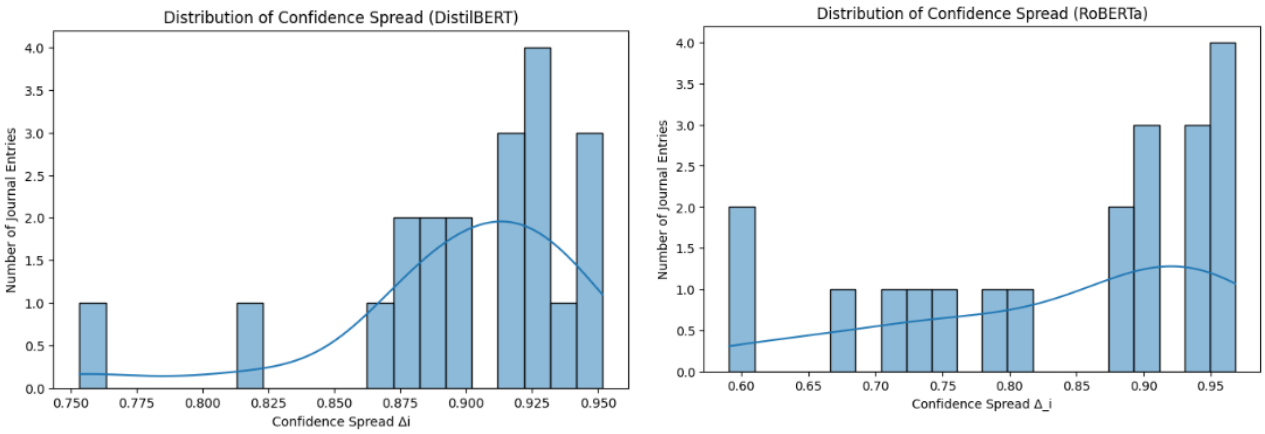


Figure 30. Example confidence spread for journal entries. With DistilBERT on the left and RoBERTa on the right.

Analysis of the confidence spreads in Figure 30 above indicates that DistilBERT produced consistently high values, with a mean of 0.898 and relatively low variance (standard deviation = 0.046). This pattern suggests that DistilBERT was generally confident in its predictions across multiple emotion classes, often producing broader summaries that included several emotions per entry. While this can provide a comprehensive perspective, it may reduce the selectivity of interpretive signals and overstate certainty in cases of affective ambiguity. In contrast, RoBERTa exhibited a lower mean confidence spread of 0.841 and substantially higher variability (standard deviation = 0.121). Several entries displayed narrow spreads ($\Delta \approx 0.59$), indicating recognition of complex or ambiguous affective content, while others approached near-total certainty ($\Delta \approx 0.97$). This variability reflects RoBERTa's capacity for more contextually grounded and selective emotion detection, consistent with user-centered observations, shown in the following section, that participants preferred its summaries for clarity and proportionality.

Collectively, these results demonstrate that confidence spread provides a quantitative proxy for semantic clarity and reflective coherence in journal texts. By distinguishing entries with high emotional certainty from those with diffuse or ambiguous affect, this measure can inform downstream template-based insight generation, prioritizing prompts for deeper reflection in entries that exhibit broader emotional ambiguity.

User-Centred Evaluation

To align evaluation with human interpretability, participants compared paired emotion summaries from DistilBERT and RoBERTa for identical journal entries. DistilBERT produced broader summaries with multiple emotions per entry, whereas RoBERTa generated more selective outputs that aligned closely with the reflective tone of the text (Table 7).

Journal Entry	**DistilBERT Emotion Summary**	**RoBERTa Emotion Summary**
"Today felt heavier than I expected. I kept replaying the conversation in my head, wondering if I said too much or not enough."	Sadness, Joy, Anger, Fear	Fear, Sadness
"I noticed how tense my body felt this morning. My shoulders were tight, and I struggled to slow my breathing."	Anger, Fear	Fear, Pessimism, Sadness
"Some feelings are hard to name. I know something is there, but I can't quite explain it."	Sadness, Joy, Anger, Fear	Sadness

Table 7. Comparative emotion summary outputs from DistilBERT and RoBERTa for sample journal extracts.

User feedback consistently preferred RoBERTa for clarity and proportionality, while DistilBERT's expansive summaries were sometimes overwhelming. Trust in outputs remained high when emotional labels were transparent and contextually grounded, even when misclassifications occurred, reflecting users' awareness of the inherent imprecision of emotional language (Amershi et al., 2019; Mohammad, 2018).

Integration Implications for PsychExtract

Overall, evaluation indicates that RoBERTa's selective, semantically grounded outputs are most appropriate for integration into PsychExtract, supporting structured reflection while preserving transparency. DistilBERT remains valuable for rapid exploratory feedback, but its outputs require downstream framing to prevent interpretive overload. By combining quantitative metrics, confidence spreads, and user-centred evaluation, the module achieves the project aim of surfacing emotional signals to encourage insight formation without imposing authoritative labels (Demszky et al., 2020; Devlin et al., 2019; Doshi-Velez & Kim, 2017).

KEYWORD EXTRACTION AND LINGUISTIC PATTERN ANALYSIS

Keyword extraction in PsychExtract is used to identify recurring cognitive and linguistic signals that may indicate emphasis, rumination, or framing strategies within reflective writing. By surfacing these patterns, the system supports user insight formation and complements emotion classification, while maintaining interpretability as the primary design goal (Angelov & Soares, 2021; Doshi-Velez & Kim, 2017).

Methodology

Two complementary keyword extraction methods were employed: KeyBERT (embedding-based semantic similarity) and YAKE (statistical, positional weighting). KeyBERT prioritises conceptually dense terms derived from contextual embeddings, whereas YAKE leverages surface-

level frequency and positional heuristics, producing lexically coherent keywords (Campos et al., 2020; Grootendorst, 2025).

KeyBERT applies a TF-IDF weighting scheme to identify high-value terms for extraction:

$$w_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t}$$

$tf_{t,d}$ is the frequency of term $t$in document d, $df_t$is the number of documents containing t, and $N$is the total number of documents.

This approach allows KeyBERT to select semantically central words or phrases, often capturing abstract or affective concepts.

Evaluation was user-centred rather than intrinsic, reflecting the subjective nature of reflective journaling (Angelov & Soares, 2021). Participants compared paired keyword sets generated by KeyBERT and YAKE for each journal extract (Table 8) and indicated which better captured the thematic content, or whether neither was satisfactory. Optional qualitative feedback provided insight into perceived strengths, limitations, and semantic adequacy. A systematic self-evaluation was also conducted across the full dataset to assess consistency and alignment with reflective signals.

Journal Entry	**Keyword Set A (KeyBERT)**	**Keyword Set B (YAKE)**
"Today felt heavier than I expected. I kept replaying the conversation in my head, wondering if I said too much or not enough."	conversation	today, head, conversation
"Writing this down helps a little. The thoughts feel less loud when they exist on paper instead of just in my head."	thoughts	head, thoughts, paper

Table 8. Example extracts presented to participants with anonymised KeyBERT and YAKE keyword sets.

Quantitative and Qualitative Findings

Across the evaluated samples, YAKE was generally preferred when participants sought broad coverage of thematic content, particularly in entries containing multiple contextual cues. For example, in "Writing this down helps a little…", all participants selected YAKE, highlighting its capacity to surface both contextual anchors (head, thoughts) and lexically salient terms.

KeyBERT, by contrast, tended to produce semantically dense, psychologically pointed keywords such as "emotionally drained" or "disconnected conversations." While effective at capturing conceptually central or affective terms, it occasionally produced truncated or awkward outputs (e.g., "terrible wasn"), reflecting the embedding-based similarity mechanism's prioritization of semantic centrality over surface coherence (Grootendorst, 2025).

Qualitative feedback reinforced these patterns. Participants valued YAKE for clarity, lexical coherence, and its ability to represent multiple facets of an entry, whereas KeyBERT was appreciated for highlighting psychologically salient cues, even when outputs were less fluid linguistically. Both methods faced challenges when processing ambiguous or diffuse emotional

content, such as "flat mood" or "overthinking," underscoring the inherent difficulty of keyword extraction in reflective, emotionally complex writing.

Integration and System Implications

User evaluation suggests that YAKE provides the most broadly interpretable and contextually coherent keyword sets, particularly when participants' preferences were consistent. Its statistical and positional weighting reliably surfaces key lexical anchors and contextual cues, offering a clear thematic scaffold for each entry. KeyBERT, while valuable for highlighting semantically dense or affectively significant terms, was less consistently preferred when clarity and coverage were prioritized.

For interpretability in reflective writing, this indicates that YAKE may serve as the primary driver of keyword extraction within PsychExtract, with KeyBERT functioning as a complementary method to surface nuanced affective or conceptual insights. Presenting both keyword sets allows users to engage with reflective writing at multiple levels: YAKE establishes broad thematic context, while KeyBERT draws attention to subtle emotional or cognitive signals.

Overall, this dual-method approach aligns with PsychExtract's interpretability-first design goals. It enables users to navigate reflective writing at both macro and micro levels, noticing overarching themes and nuanced emotional patterns without imposing a singular "correct" interpretation. Observed limitations, such as the small participant pool and absence of formal inter-annotator agreement, are consistent with the exploratory, non-clinical framing of the project and highlight directions for future evaluation and system refinement.

LINGUISTIC TEMPLATING EVALUATION

The linguistic templating module in PsychExtract transforms extracted computational signals (such as emotional load, restlessness, coping strategies, and self-assessment cues) into structured, interpretable textual insights. Unlike classification tasks, this component is not evaluated for predictive accuracy; its purpose is to provide reflective prompts that are human-readable, semantically coherent, and contextually grounded (Amershi et al., 2019; Doshi-Velez & Kim, 2017).

Methodology

Evaluation was user-centred, focusing on perceived applicability, reflective usefulness, and linguistic tone. Participants were presented with short journal extracts alongside automatically generated template-based insights. Each template combined detected themes with controlled natural language constructions, designed to resemble human reflective observations rather than clinical or diagnostic statements (Baumer, 2015; Calvo & Peters, 2017).

Two representative journal extracts were selected to illustrate common reflective patterns as shown in Table 9:

Journal Entry	Themes Detected	Generated Insight
"Today felt heavier than I expected. I kept replaying the conversation in my head, wondering if I said too much or not enough."	Emotional Load, Restlessness	This entry suggests a relatively high emotional load, particularly in relation to their conversation. It reflects a state of tension or agitation associated with the conversation.
"I noticed how tense my body felt this morning. My shoulders were tight, and I struggled to slow my breathing."	Emotional Load, Coping, Restlessness, Self-Assessment	The overall tone of this entry indicates emotional heaviness connected to their slow breathing and tense body. The writer appears to be engaging in a coping process while reflecting on their physical tension. This entry reflects a state of tension or agitation and demonstrates reflective self-evaluation.

Table 9. Example journal extracts with generated linguistic template insights, showing weighted emotion influence on phrasing.

Participants rated each template using 5-point Likert scales for:

Applicability to the source text,
reflective usefulness, and
wording perception (e.g., appropriate, clinical, verbose)

Optional qualitative feedback captured nuanced impressions of phrasing, tone, and perceived value.

Templates incorporate emotion weighting derived from the confidence spread of the emotion classification module:

$$S_e = \frac{\sum_i p_i \cdot w_i}{\sum_i w_i}$$

p_i is the predicted probability of emotion i
and $w_i$is the template-specific weight for that emotion class.

This ensures that highly salient emotions proportionally influence reflective phrasing, supporting meaningful insight formation and interpretive alignment with the user's perspective.

Findings
Across both extracts, participants rated templates as highly applicable (Figure 31) and useful for reflection (Figure 32). Extract 1 consistently received maximum ratings, indicating that templates successfully captured cognitive and emotional nuance in rumination-focused entries. Extract 2, which centred on bodily awareness, scored slightly lower but remained positive, reflecting minor misalignments in interpreting coping strategies or somatic signals.

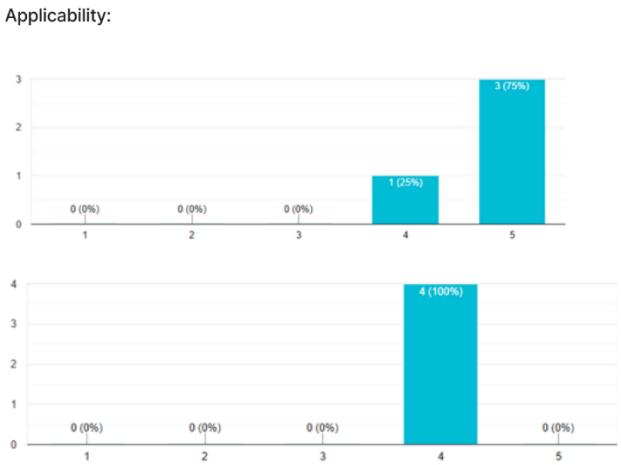


Figure 31. Distribution of participant ratings for applicability across different extracts. 1 represents low applicability, 5 represents high applicability.

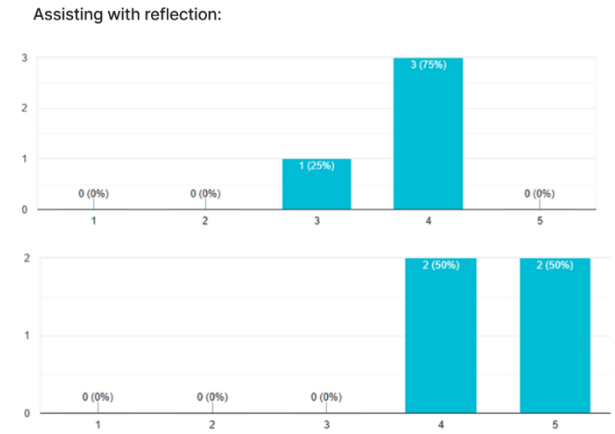


Figure 32. Distribution of participant ratings for reflective usefulness across different extracts. 1 represents low reflective usefulness, 5 represents high reflective usefulness.

Qualitative feedback highlighted that templates were generally clear and contextually relevant, though occasional clinical tone or verbosity prompted interpretive friction (Table 10). Importantly, this friction did not reduce reflective utility; in several cases, it encouraged participants to critically engage with their writing, prompting comparison, acceptance, or reinterpretation of the insights in light of their own experience (Amershi et al., 2019; Calvo & Peters, 2017).

Journal Extract	**Wording Perception**
Extract 1	Appropriate; occasionally Clinical
Extract 2	Appropriate; Clinical

Table 10. User ratings of linguistic template insights.

Implications for PsychExtract Integration

The evaluation demonstrates that linguistic templating effectively bridges computational analysis and human reflection. By combining emotion-weighted signals with controlled phrasing, templates provide interpretable insights without asserting authority, aligning with PsychExtract's assistive,

non-clinical ethos. Minor misalignments or verbosity can be leveraged as reflective prompts, reinforcing interpretive engagement rather than diminishing clarity.

Templates thus serve as a core interpretive layer, translating complex affective and cognitive signals into accessible, user-centric outputs. This confirms the broader project principle: effectiveness in reflective tools depends on supporting thoughtful engagement rather than absolute correctness (Baumer, 2015; Doshi-Velez & Kim, 2017).

CONCLUSION AND FUTURE WORK (858/1000 WORDS)

PROJECT SUMMARY

PsychExtract represents an interpretability-first system for structured reflection on emotionally expressive text. The project operationalized the orchestration of multiple pre-trained models across different data modalities to support insight formation in personal writing. PsychExtract integrates three primary model classes: optical character recognition (OCR) for document ingestion, emotion classification leveraging fine-grained sentiment models (GoEmotions), and keyword extraction with thematic and linguistic pattern analysis (KeyBERT, BERTopic, and lightweight templating heuristics). Together, these components enable the system to process reflective text, identify salient emotional cues, extract cognitive themes, and generate interpretive summaries through template-based framing.

Development demonstrated the feasibility of combining heterogeneous models into a coherent workflow. The OCR module reliably transcribed handwritten and printed text, preserving semantic integrity crucial for downstream analysis. Emotion classification produced probabilistic outputs that highlighted dominant affective states, while keyword and linguistic pattern extraction revealed recurring themes and semantic structures. Preliminary evaluation confirmed that text-to-speech (TTS) synthesis is a highly valued feature, aligning with the project's original goal of multimodal accessibility. While TTS has not yet been fully implemented, its inclusion is clearly validated. Additionally, the project produced a Streamlit-based user interface design, supporting future user interaction with PsychExtract in a dynamic, visually guided environment.

REFLECTION ON OBJECTIVES

Several primary objectives were achieved. The multi-model pipeline was successfully constructed: OCR, emotion classification, and keyword extraction modules were implemented and evaluated, demonstrating effective data processing and interpretable outputs. Preliminary user feedback confirmed that the system supported reflective engagement and highlighted the importance of TTS for accessibility.

Some objectives were partially met. Full integration of all components, including real-time processing and templating, remains in progress. Adaptive interpretive framing and personalisation were explored conceptually but not fully implemented. Ethical safeguards were embedded into design, though broader deployment scenarios and long-term user studies remain future work. Similarly, the Streamlit UI has been designed but not yet deployed.

These outcomes underscore the project's accomplishments while highlighting areas for improvement. PsychExtract validates that pre-trained models can generate structured, interpretable insights from text while revealing the challenges of integrating heterogeneous models,

multimodal outputs, and contextually appropriate interpretation. These reflections inform next steps for refinement and the longer-term vision for interpretability-driven tools.

FUTURE WORK

Future development focuses on full integration of system components into a real-time platform. This includes automating the orchestration of OCR, emotion classification, and keyword extraction pipelines, linking them to the templating engine, and incorporating TTS synthesis for auditory feedback. Integration at this level would transform PsychExtract from a modular prototype into a seamless reflective assistant.

Expanded user studies are also a priority. While initial feedback suggested outputs were intelligible and useful, systematic evaluation with a diverse cohort would provide richer insights into usability, interpretive accuracy, engagement, and multimodal accessibility. Such studies would inform refinements to both the interface and framing logic, ensuring outputs remain meaningful across users with varying literacy levels, cultural backgrounds, and reflective practices.

Personalisation represents another development trajectory. Leveraging metadata about individual users, writing contexts, and prior reflections could allow the system to tailor keyword extraction, emotion weighting, interpretive framing, and TTS delivery to each user's style. Such adaptation would enhance the relevance and impact of outputs, fostering deeper self-reflection.

Ethical extensions remain essential. Strengthening privacy safeguards, embedding mechanisms to prevent psychological distress, and ensuring transparency in algorithmic decision-making will be central to responsible deployment. Clear guidelines for informed consent and interpretive disclaimers will be vital if PsychExtract is made publicly available.

BROADER IMPLICATIONS

PsychExtract illustrates the potential of NLP to support reflective practice. By structuring insights in an interpretable format and incorporating TTS for multimodal accessibility, the system empowers users to engage with reflections in a more conscious, analytic way. NLP in this context complements, rather than replaces, human introspection, amplifying insight and metacognitive awareness.

The interpretability-first approach is significant within mental health tools. By prioritizing transparency, traceability, and user agency, PsychExtract ensures that insights are understandable, actionable, and non-intrusive. The planned Streamlit interface further supports this by providing an accessible platform for interactive exploration of insights.

The project also highlights challenges and opportunities in orchestrating multiple pre-trained models. Integration across domains and modalities requires careful attention to data flow, output compatibility, and scalability. These lessons extend beyond PsychExtract, providing insights for researchers seeking to leverage pre-trained models in complex, human-centered workflows.

CLOSING REMARKS

In conclusion, PsychExtract has demonstrated a proof of concept for an interpretability-first, multi-model system supporting structured reflective practice. OCR, emotion classification, and keyword extraction pipelines were successfully implemented, producing insights from personal text while remaining ethically grounded and user-centric. Full integration, TTS implementation,

personalisation, and expanded user studies remain future work, alongside development of the Streamlit UI to enable accessible interaction.

Ultimately, PsychExtract highlights the transformative potential of combining NLP with reflective practice. By operationalizing insight in a computationally interpretable way and incorporating both visual and auditory modalities, the project provides a model for responsible, ethically aware tools that support emotional awareness and cognitive engagement. It underscores the importance of interpretability, modular design, multimodal accessibility, and user-centered evaluation in AI-driven reflective systems, offering a roadmap for future projects at the intersection of machine learning, mental health, and human-computer interaction. PsychExtract contributes not only a working prototype but also a conceptual framework for the future of interpretability-first, supportive technologies for personal insight.

REFERENCES

Angelov, P.P. & Soares, E.A., 2021. Towards explainable deep neural networks (xDNN). *Neural Networks*, 130(Feb.), pp.185–194. [Journal] https://doi.org/10.1016/j.neunet.2020.07.010

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R. & Horvitz, E., 2019. Guidelines for human–AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp.1–13. [Conference] https://doi.org/10.1145/3290605.3300233

Bai, J., et al., 2024. Qwen-VL: A Frontier Large Vision-Language Model. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arxiv.2308.12966

Baumer, E.P.S., 2015. Reflective Informatics: Conceptualizing the Design of Reflective Systems for Workplace Practice. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp.585–594. [Conference] https://doi.org/10.1145/2702123.2702234

Bhadresh Savani (Hugging Face Model), 2022. DistilBERT fine-tuned for emotion classification. *Hugging Face*. [Model] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

Bojanowski, P., et al., 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, pp.135–146. [Journal] https://doi.org/10.1162/tacl_a_00051

Ben-Zeev, D., et al., 2013. Mobile technologies among people with serious mental illness: opportunities for future services. *Administration and Policy in Mental Health and Mental Health Services Research*, 40(4), pp.340–343. [Journal] https://doi.org/10.1007/s10488-012-0424-x

CardiffNLP, 2023. Twitter RoBERTa multi-label emotion classifier. *Hugging Face*. [Model] https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion-multilabel-latest

Campos, R., Joulin, A., Mikolov, T., et al., 2020. YAKE! Keyword Extraction from Single Documents using Multiple Local Features. *Information Sciences*, 509(Jan.), pp.257–289. [Journal] https://doi.org/10.1016/j.ins.2019.09.013

Calvo, R.A. & Peters, D., 2017. Positive Computing: Technology for Wellbeing and Human Potential. MIT Press. [Book] https://doi.org/10.7551/mitpress/9764.001.0001

Coqui.ai, 2025. Coqui.ai TTS. *TTS 0.22.0 documentation*. [Documentation] https://docs.coqui.ai/en/latest archived at https://web.archive.org/web/20250812191341/https://docs.coqui.ai/en/latest

Cui, C., et al., 2025. PaddleOCR-VL: Boosting Multilingual Document Parsing via a 0.9B Ultra-Compact Vision-Language Model. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arXiv.2510.14528

DeVault, D., et al., 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. *Proceedings of AAMAS '14*, pp.1061–1068. [Conference] https://dl.acm.org/doi/10.5555/2615731.2617415

Demszky, D., et al., 2020. GoEmotions: A Dataset of Fine-Grained Emotions. *Proceedings of ACL*, pp.4040–4054. [Conference] https://aclanthology.org/2020.acl-main.372/

Devlin, J., et al., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1(Jun.), pp.4171–4186. [Conference] https://doi.org/10.18653/v1/N19-1423

Doshi-Velez, F. & Kim, B., 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arxiv.1702.08608

Grootendorst, M., 2022. BERTopic: Neural Topic Modelling with Transformers and class-based TF-IDF. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arXiv.2203.05794

Grootendorst, M., 2025. KeyBERT. *GitHub*. [Code] https://github.com/MaartenGr/KeyBERT archived at https://web.archive.org/web/20250927082041/https://github.com/MaartenGr/KeyBERT

Greenberg, L.S. & Pascual-Leone, A., 2006. Emotion in psychotherapy: A practice-friendly research review. *Journal of Clinical Psychology*, 62(5), pp.611–630. [Journal] https://doi.org/10.1002/jclp.20252

Goodfellow, I., Bengio, Y. & Courville, A., 2016. Deep Learning. MIT Press. [Book] https://www.deeplearningbook.org

Gulliksen, J., Göransson, B., Boivie, I., Blomkvist, S., Persson, J. & Cajander, Å., 2003. Key principles for user-centred systems design. *Behaviour & Information Technology*, 22(6), pp.397–409. [Journal] https://doi.org/10.1080/01449290310001624329

Hill, C.E., et al., 2007. Insight in psychotherapy: Definitions, processes, consequences, and research directions. In *Insight in psychotherapy*, L.G. Castonguay & C.E. Hill (Eds.), pp.441–454. American Psychological Association. [Book Chapter] https://doi.org/10.1037/11532-021

Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A., 2020. spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*. [Documentation] https://doi.org/10.5281/zenodo.1212303 archived at https://web.archive.org/web/20260128135403/https://zenodo.org/records/10009823

Inkster, B., et al., 2020. Digital Health Management During and Beyond the COVID-19 Pandemic: Opportunities, Barriers, and Recommendations. *JMIR Mental Health*, 7(7), Article e19246. [Journal] https://doi.org/10.2196/19246

Jacobs, M., et al., 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pp.1–14. [Conference] https://doi.org/10.1145/3411764.3445385

JaidedAI, 2024. EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including Latin, Chinese, Arabic, Devanagari, Cyrillic, etc. *GitHub*. [Documentation] https://github.com/JaidedAI/EasyOCR archived at https://web.archive.org/web/20251012081255/https://github.com/JaidedAI/EasyOCR

Koleck, T.A., et al., 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 24(4), pp.364–379. [Journal] https://doi.org/10.1093/jamia/ocy173

Li, M., et al., 2021. TrOCR: Transformer-based optical character recognition with pre-trained models. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arXiv.2109.10282

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V., 2019. RoBERTa: A robustly optimised BERT pretraining approach. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arXiv.1907.11692

Luxton, D.D., 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, 62(1), pp.1–10. [Journal] https://doi.org/10.1016/j.artmed.2014.06.004

Medicines and Healthcare Products Regulatory Agency (MHRA), 2025. Software and artificial intelligence (AI) as a medical device. *Gov.uk*. [Report] https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device archived at: https://web.archive.org/web/20250815022341/https://www.gov.uk/government/publications/software-and-artificial-intelligence-ai-as-a-medical-device/software-and-artificial-intelligence-ai-as-a-medical-device

Miner, A.S., et al., 2016. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Internal Medicine*, 176(5), pp.619–625. [Journal] https://doi.org/10.1001/jamainternmed.2016.0400

Mohammad, S.M., 2018. Obtaining reliable human ratings of valence, arousal, and dominance for emotions. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp.174–184. [Conference] https://aclanthology.org/P18-1017/

Mohammad, S.M., 2022. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*, 48(2), pp.239–278. [Journal] https://doi.org/10.1162/coli_a_00433

Morris, A., Maier, V. and Green, P., 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*. [Conference] https://doi.org/10.21437/Interspeech.2004-668

Mukherjee, S.S., et al., 2020. Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry*, 5(Dec.), pp.76–106. [Journal] https://doi.org/10.1162/cpsy_a_00030

Norman, D.A., 2013. The design of everyday things: Revised and expanded edition. Basic Books. [Book] https://dl.icdst.org/pdfs/files4/4bb8d08a9b309df7d86e62ec4056ceef.pdf

Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), pp.62–66. [Journal] https://doi.org/10.1109/TSMC.1979.4310076

Pennebaker, J.W., et al., 2015. The Development and Psychometric Properties of LIWC2015. University of Texas at Austin. [Report] https://doi.org/10.13140/RG.2.2.23890.43205

Plamondon, R. & Srihari, S.N., 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Trans Pattern Anal Mach Intell*, 22(1), pp.63–84. [Journal] https://doi.org/10.1109/34.824821

Patel, C.I., et al., 2012. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study. *International Journal of Computer Applications*, 55(10), pp.50–56. [Journal] https://doi.org/10.5120/8794-2784

pyttsx3.readthedocs.io, 2025. pyttsx3: Text-to-speech cross-platform. *pyttsx3 2.6 documentation*. [Documentation] https://pyttsx3.readthedocs.io/en/latest archived at https://web.archive.org/web/20251010005655/https://pyttsx3.readthedocs.io/en/latest

Reimers, N. & Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.3982–3992. [Conference] https://aclanthology.org/D19-1410/

Rice, S.V., Jenkins, F.R. & Nartker, T.A., 1993. The fourth annual test of OCR accuracy. *National Institute of Standards and Technology (NIST), NISTIR 5077*. [Report] https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir5077.pdf

Sanh, V., et al., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*. [Preprint] https://doi.org/10.48550/arXiv.1910.01108

Shanafelt, T.D., et al., 2016. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic Proceedings*, 91(7), pp.836–848. [Journal] https://doi.org/10.1016/j.mayocp.2016.05.007

Shen, J., et al., 2018. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. *Proceedings of ICASSP '18*, pp.4779–4783. [Conference] https://doi.org/10.1109/ICASSP.2018.8461368

Smith, R.W., 2007. An Overview of the Tesseract OCR Engine. *Proceedings of ICDAR '07 Vol. 2*, pp.629–633. [Conference] https://doi.org/10.1109/ICDAR.2007.4376991

Torous, J., et al., 2018. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evidence-based Mental Health*, 21(3), pp.116–119. [Journal] https://doi.org/10.1136/eb-2018-102891

Turner, R.J., et al., 2022. Information extraction from free text for aiding transdiagnostic psychiatry: constructing NLP pipelines tailored to clinicians' needs. *BMC Psychiatry*, 22, Article 407. [Journal] https://doi.org/10.1186/s12888-022-04058-z

Tsoumakas, G., Katakis, I. & Vlahavas, I., 2007. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pp.667–685. Springer. [Book Chapter] https://doi.org/10.1007/978-0-387-09823-4_33

Vaswani, A., et al., 2017. Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp.6000–6010. [Conference] https://dl.acm.org/doi/10.5555/3295222.3295349

Young, M.C., et al., 2018. The Effects of Text-to-Speech on Reading Outcomes for Secondary Students With Learning Disabilities. *Journal of Special Education Technology*, 24(2), pp.80–91. [Journal] https://doi.org/10.1177/0162643418786047

Zhang, M.-L. & Zhou, Z.-H., 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8), pp.1819–1837. [Journal] https://doi.org/10.1109/TKDE.2013.39

Zuiderveld, K., 1994. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems IV*, pp.474–485. Academic Press. [Book Chapter] https://dl.acm.org/doi/10.5555/180895.180940