Solace Safety Framework for Responsible Use of Generative AI when Providing Mental Health Support

THE NEED

When AI is used to provide psychological support, there is a responsibility to prevent potential harms and uphold user safety.

Potential harms include:

- Inadequate handling of high-risk scenarios (e.g. suicidal ideation, self-harm, active crises)
- Inaccurate or misleading guidance
- Misinterpreted as a substitute for professional advice
- Suggestions not grounded in evidence-based practice in psychology

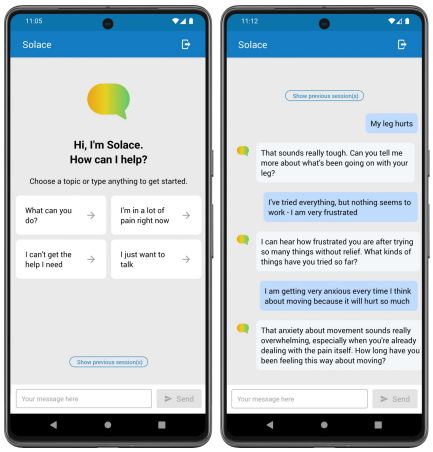
These harms can lead to:

- Compromised safety and well-being [1,2,3,4]
- Erosion of user trust and therapeutic alliance
- Prevention of seeking necessary professional care

To address these risks, we developed a **Safety Framework** to guide the responsible design, testing, and deployment of *Solace*, our AI companion.

THE SOLUTION: SOLACE

Solace is a **first-of-its-kind AI companion that delivers real-time evidence-based pain psychology support**. Solace engages with empathy, builds actionable plans, and provides personalized support. Solace is integrated into the Manage My Pain platform and uses data from it to reinforce skills, track adherence, and drive progress.



Solace has been designed, tested, and deployed using this **Safety Framework.**

OUR FRAMEWORK

Our comprehensive Safety Framework ensures that prevention and reduction of harm has been considered across every layer of Solace's design, testing, and deployment. The framework is composed of five overlapping pillars to ensure safe and trustworthy AI engagement when providing evidence-based support for those living with pain.

- **Guardrails and Fail-safes**: Layered validators for detecting and appropriate addressing high-risk scenarios
- **Validation:** Comprehensive testing using automated, generative AI, and human-in-the-loop reviews
- **Ecosystem Monitoring:** Continuous external monitoring of publicly-reported information about AI in mental health
- **Transparency & Trust:** Disclaimers throughout the user experience highlight what Solace is able to do and not do
- **Governance**: Multi-layered oversight through dedicated committees and expert reviewers responsible for auditing, approving, and monitoring system updates to ensure Solace adheres to current clinical standards and AI practice

Solace Safety Framework for Responsible Use of Generative AI when Providing Mental Health Support

GUARDRAILS & FAIL-SAFES

Solace uses a multi-layered system with bounded scope to ensure compliant responses to high-risk or sensitive user input such as, but not limited to:

- **High-risk Scenarios:** Suicidal ideation, self-harm, violence against others, treatment or intervention recommendations, diagnosis
- **Bias Scenarios:** Age, sex, gender, opioid use, housing status, race, ableism, occupation, health status
- **Other Scenarios:** Depression, technical competence, physical ability, intimacy, emotional attachment

When medical emergencies or active suicidal intent are detected, users are provided with contact information for verified and locally-accessible crisis support services.

VALIDATION

- **Automated regression testing:** Before any change to Solace is made, a set of tests for known prompt triggers is conducted to validate existing guardrails and fail-safes are functioning as intended
- **Simulated conversations:** Sessions between Solace and generative AI personas designed to elicit bias and stigma or simulate active emergencies - transcripts are analyzed for bias and stigma using both purpose-built AI tools and manual expert reviews
- **Expert reviews:** Nuanced prompts shown to elicit problematic responses from other LLMs and derived from ecosystem monitoring processes, are tested and analyzed through manual review by clinical experts.

ECOSYSTEM MONITORING

Continuous external monitoring of publicly-reported safety issues, deficiencies, technical developments, standards and regulations from sources such as:

- State/provincial/federal level AI regulations from entities such as Health Canada and the FDA (US)
- Media reports related to the use of AI in mental health support and issues that arise
- Industry publications or white papers about guardrails and/or safety concerns of LLMs
- Academic publications related to LLM deficiencies
- Organizations such as NIST or ISO

Internal processes ensure ecosystem monitoring is done regularly and development uses the latest information.

TRANSPARENCY & TRUST

- **Disclaimers** requiring active acknowledgement before first-use and displayed at each session inform users that Solace is not a substitute for professional healthcare advice
- **In-session transparency** contained within Solace's response ensure limitations of Solace are clear (e.g. cannot recommend treatment) and reinforce when a healthcare professional should be involved
- **Terms & conditions** are clearly described in Solace-specific sections of both Privacy Policy and End User Licence Agreement to make it clear about how Solace and the data provided to it can be used
- **Open source release** of this **Safety Framework**

GOVERNANCE

Our framework is supported by an **AI Ethics Review Committee**, made up of patients with lived experience and individuals with research or clinical expertise in pain management or AI ethics, to provide stewardship throughout Solace's development and of this Safety Framework. Internal processes are in place to realize the Safety Framework, and identify and investigate risks, assign accountability, and implement mitigation strategies:

- **Automated alerts** for system failures and safety guardrail activations to immediately notify responsible teams and strengthen real-time responsiveness
- **Agile development** → Continuous improvements to enhance user safety and learning from real-world usage
- **Periodic internal audits** → Testing the processes and procedures to validate the Safety Framework is implemented
- **Ongoing research and validation** → Publications of Solace and its development with academic partners

Together, these efforts ensure our **Safety Framework** remains effective and adaptive for Solace users.