

# 生成AIベースのメンタルケア対話ボットの企業内利用コンセプトと提供準備について



Metadata

Livie's Jump, メタデータ株式会社 松田 圭子

生成系AIとロジックのハイブリッド型で実装したメンタルケア対話ボット「やわらかネコ」を個人向けに開発し、2023年9月よりβ提供を行ってきた。現在、この「やわらかネコ」をベースとし、企業内で従業員のメンタルケアに用いたいという需要に応える「ほっと猫」を研究開発中である。個人向けと企業内利用でのコンセプトの差と、それによる実装・提供における差について発表する。

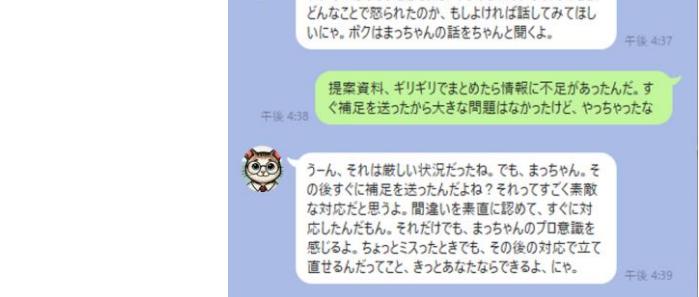
## 「やわらかネコ」と「ほっと猫」：コンセプトや実装の差異



- 1. 目的の違い**  
個人向け：気持ちの整理・ストレスや不安の軽減・自己評価を適切に向上  
企業向け：
  - 従業員のメンタルヘルス支援（個人向けの目的に加えて）
    - 従業員のストレスが業務に影響を与えること、離職のリスクが高まる前に、サポートリソース（例：企業カウンセラー、産業医）を提案するエスカレーションを提供
    - 企業ごとのカスタマイズ
    - 社内の特定の問題や文化に対応できるノウハウ、知識の組み込み



- 2. ユーザーインターフェース、機能の違い**  
個人向け：LINEベース。ユーザーの自動的な利用に向く。穏やかで和むキャラづくり  
企業向け：LINEに加え、SlackやMicrosoft Teamsにも対応。業務環境で同僚にDMするように利用可能に。カスタマイズ機能追加（業務の質問に回答、プッシュ通知設定etc.）キャラも気のいい同僚のような雰囲気を目指す



- 3. セキュリティとプライバシー**  
個人向け、企業向けどちらも：対話内容はユーザーだけのもの  
企業向けでの注意点
  - 個人の対話内容を管理職など他者が閲覧することは不可とする
  - ユーザーの対話内容から深刻なメンタル不調が懸念される場合に、企業で設定するエスカレーションを行う（企業カウンセラー・産業医・メンタルケアサービスへの連絡を促す等）
  - 対話のサマリをエスカレーション先に共有するかは要検討。基本的には、ユーザーの同意を得た場合のみ行うこととしたい

## 背景

### メンタルヘルス不調による休職・退職について

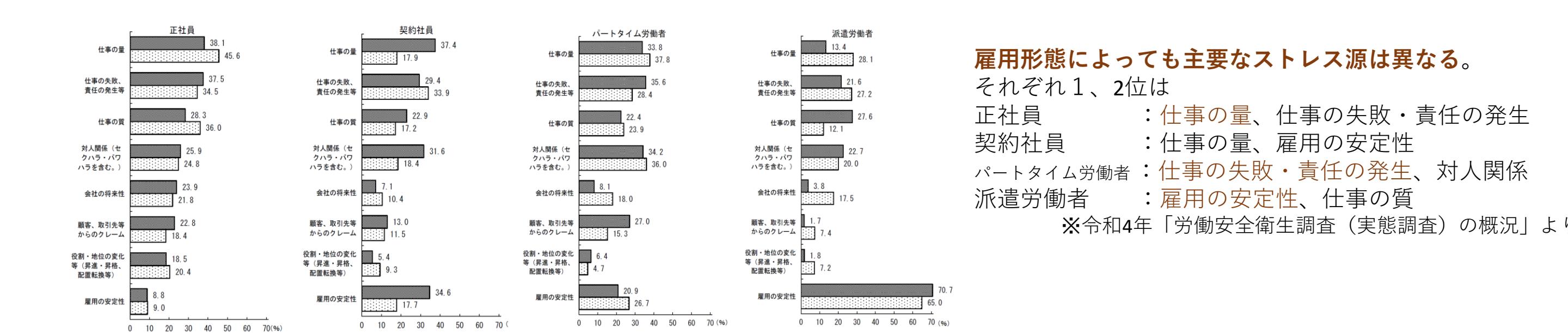
- メンタルヘルス不調により1ヶ月以上連続で休職した労働者がいる事業所の割合は10.6%
- メンタルヘルス不調により退職に至った労働者がいる事業所の割合は5.9%

※令和4年「労働安全衛生調査（実態調査）の概況」1年間(R3.11～R4.10期間中)での数値 [https://www.mhlw.go.jp/toukei/list/dl/r04-46-50\\_gaikyo.pdf](https://www.mhlw.go.jp/toukei/list/dl/r04-46-50_gaikyo.pdf)

### 従業員のストレス状況

- 日本の労働者の約82.2%が、現在の仕事や職業生活に関する強いストレスを感じている
- 主なストレスの原因は「仕事の量」(36.3%)、「仕事の失敗や責任の発生」(35.9%)、および「対人関係」(26.2%)
- 「相談できる相手がいる」91.4% 相談相手は家族・友人68.4%、同僚68%、上司65%（複数回答）  
…だが実際に相談したことがあるのは69.4%（相手が居る人のうち）→4割は相談したことがない
- 休業・退職とともに情報通信業がトップ（それぞれ32.0%、17.0%）2位は休業は電気ガス熱供給水道、退職は医療・福祉

※独立行政法人労働政策研究・研修機構による「令和4年「労働安全衛生調査（実態調査）の概況」”抜粋解説



### 企業のメンタルヘルス対策の現状

- 企業の約9割(86.2%)がメンタルヘルス問題が業務パフォーマンスに負の影響を与えると認識しているが、過去1年間にメンタルヘルスで1ヶ月以上の休職または退職した労働者がいた事業所の3分の1(36.0%)が具体的な対策を講じていない

※「職場におけるメンタルヘルスケア対策に関する調査」結果 <https://www.jil.go.jp/press/documents/20110623.pdf>

## 「ほっと猫」での提供準備（問題解決のための対策、機能）

### メンタルケア要素のある応答でのエンパワメントと従業員のモチベーション低下・離職防止：

- 対話生成AIの活用によりストレス源の種類を問わず、やわらかネコ同様の受容・傾聴・プラスのフィードバックが可能（持たせるナレッジにより特化も可能）
- 職場の人間関係等の知識、業種別Tips、企業内ノウハウ（カスタマイズ）の情報を持ち回答に活用

### エスカレーションを要するユーザーのメンタル不調の検知について試作：

- (1) メンタル状態の数値による把握  
a) ユーザーが日々の気分を7段階で入力。閾値以下の値が1週間以上続くと検知  
b) AIによる機嫌の数値化、対話内容からAIが自動的に機嫌を推測し、数値化

- (2) 対話の内容から緊急度を数値化  
a) 自殺願望の言葉や「疲れた」「休みたい」といった特定表現を検知し緊急度を数値化。一定期間中の値で評価  
b) AIによる推測：対話全体のトーンやコンテキストから緊急度を推測する

※緊急度を検知しない状況でも、問題解決への本人の要望がある場合やお勧めに同意を得た場合等、フローから実行動やエスカレーションへの案内を出す実装も

### エンゲージメント向上、生産性向上：

- 「企業がメンタルヘルスケアサービスを導入した」ことによる効果（→導入自体による価値なので個別の実装なし）

#### 企業内のノウハウ、知識を含むカスタマイズ応答の実装

「ほっと猫」が企業内のリソースをRAGデータとして格納し、対話の中で回答できるようにする

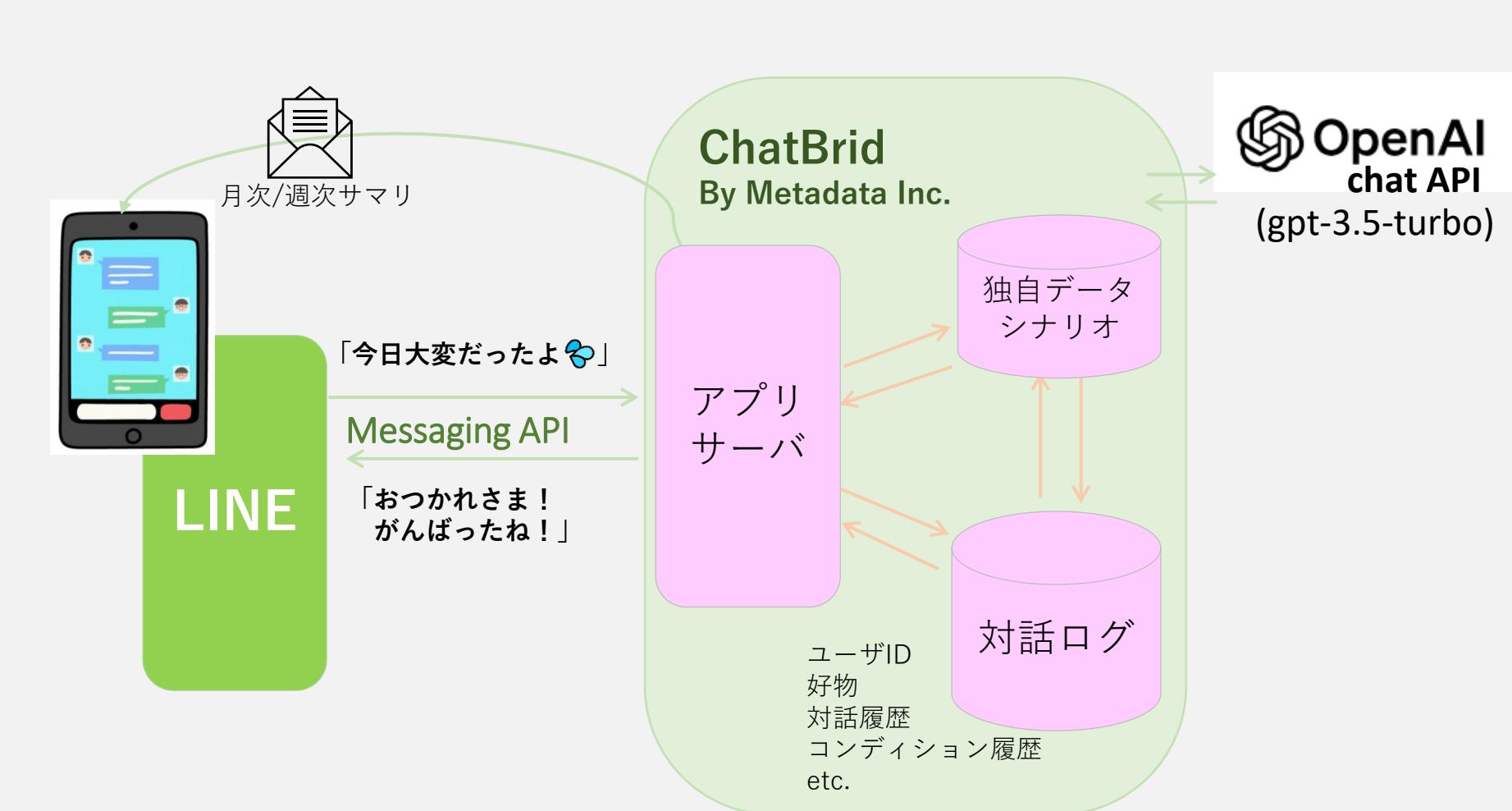
→既存の実装（chatbrid内）活用で実現可能。課題はデータ準備と階層化等の整備

・継続利用を促す要素の試作

・「Repetition penalty」：占い、天気予報、企業内お知らせ等のプッシュメッセージ提供で対話へ繋げる  
・「Frequency penalty」：ユーザーが何を話すか迷う ユーザー向けにランダムで気軽な話題候補を

## システム構成の差異

### やわらかネコのシステム構成



- 対話インターフェースとしてLINE(Messaging API)を活用。親密な内容を書きやすい1対1対話アドリとして普及していることから選択。
- 対話生成AIとしてはOpenAIのchat(chat completion) APIを使用。現在はgpt-3.5-turboを使用。（感情的回答に強い）
- ユーザー情報（利用状況、各種ログ等）の管理およびロジック処理部分、専門知識の格納、対話内容に応じたプロンプトのハンドリング部分にメタデータ社のChatBridを使用。

### ほっと猫のシステム構成



- 対話インターフェースとしてLINE/Slack/Teamsを活用。業務用メッセージ機能を含むアドリとして普及していることから選択。
- 利用可能な対話生成AIはOpenAI/Gemini/Claude/Groq。LLMは機能により切替を行う見通し。製作時点ではOpenAIのgpt3.5, 40-miniを活用
- 対話の初段階ロジック、シナリオ処理部分にmiibo、ユーザー情報（利用状況、各種ログ等）の管理および専門知識のRAGデータ格納にメタデータ社のChatBridを使用。これにより対話内容、選択メニューによる対話生成AIおよびRAGの切り替えを実現

## 課題と解決策 1. メンタル状態の数値化について

### 課題：ユーザーによる虚偽の入力

→解決策：ユーザー入力の数値とAIによる数値化との両建て

### 課題：対話内容からの緊急度検出の精度向上

→解決策：試作から、AI検出のプロンプトによるチューニングが有効とみられる  
(精度向上についてAI本体と「相談」して向上するアプローチも有効。下記はその一部)

松田：…この緊急度判断方法で誤判断する可能性にはどんなものがある？

AI：誤判断のリスクは複数存在します。以下に主なリスク要因とその詳細を挙げます。

#### 1. コンテキストの不足

- 限られた情報量：ユーザーのメッセージは一時的な感情や状況を反映している場合が多く、長期的なメンタルヘルスの状態を正確に把握するには不十分です。
- 非言語的因素の欠如：対話ボットは表情や声のトーンなどの非言語的なサインを読み取ることができません。これにより、深層的な感情やニュアンスを見逃す可能性があります。

#### 2. 言語の曖昧さと誤解

松田：…ではその対策としてプロンプトを改善できる？

AI：プロンプトを改善することで、ChatGPTによる緊急度スコアの精度を向上させ、誤判断のリスクを軽減することは十分可能です。

以下に、具体的な改善方法とその理由を詳しく説明します。

#### 1. スコアリング基準の詳細化

…

## 課題と解決策 2. LLMの選定およびパラメータ指定

### 課題：

ほっと猫ではメニュー選択等により各種機能を持つ対話をを行うが、機能ごとに対話に求められる適性が異なる。

例) 雑談→受容的であること、柔軟さ。企業内情報への質問→問題解決力。

雑談の際に問題解決にフォーカスしすぎるとユーザのストレスを増やすリスク

### 解決策：

#### 1) 状況によるLLMの使い分け

実装環境で利用可能なLLMの特徴は以下。

OpenAI GPT系 ※参考：公式ドキュメント <https://platform.openai.com/docs/concepts> ほか

GPT-3.5系：雑談や感情の受け止めに優れ、傾聴や共感的な対話に強い。感情のトーンを把握する能力が高い

GPT-4系：問題解決にも向く、トピック検知に強み。感情を捉える能力も向上。GPT-3.5に比べ柔軟さでは劣るか

o1-preview: 高度な推論や複雑な問題解決タスクに優れ、精度が高い反面、処理速度やコスト面でやや負担が大きい

o1-mini: 小規模なモデルで、効率的かつコスト効果の高い推論が可能。コード生成に効果を発揮

Anthropic Claude系 ※参考：公式ドキュメント <https://docs.anthropic.com/ja/docs/intro-to-claude> ほか

Claude 3 Opus: データ処理や複雑なタスクに優れており、特に科学的なリサーチやレポートの解析に適する

Claude 3 Sonnet: 自然で人間らしいトーンの生成。感情を把握する能力が向上。

Meta Llama系 ※参考：公式ドキュメント <https://ai.meta.com/blog/meta-llama-3/> ほか

LLama 2: 軽量で効率的。雑談や柔軟な対話に適しているため、気分の数値化や傾聴型対話に有効とみられる

LLama 3: 大規模データの解析やプログラミングタスクでの強みを持ち、自然言語処理に優れる

### 仮説

メンタル状態の数値化：GPT-4o-mini (LLama2?)

雑談型：GPT-3.5-turbo、Claude3 Sonnet

問題解決型：GPT-4o-mini (今後安定版が出たらo1系も)、Claude3 Opus

#### 2) パラメータ調整

以下のパラメータは対話の適性に影響すると考えられる。（パラメータ名の後の数値はOpenAIでのデフォルト値）

##### 1. Temperature (1.0)

応答を生成する際のランダム性を指定。0に近づくと決定的で予測可能な応答、値が高いほど創造的で多様な応答

##### 2. Top-p (1.0)

候補に入るトークンの範囲。デフォルトでは全て。値を下げるごとに、上位n%の確率を持つ候補のみが選ばれるため応答の多様性が抑えられる

##### 3. Top-k (0) (無効化)

値を設定すると上位k個の候補からサンプリングされ、特定の範囲内でのみ応答を生成する。デフォルトは無制限

##### 4. Max tokens (4096) (GPT-3の一部モデルの場合)

生成されるトークン数の最大値。トークン数が多いほど長い応答が生成される（モデル毎に上限あり）

##### 5. Frequency penalty (0.0)

値が高いと、モデルが既に生成した単語の繰り返しを避ける。デフォルトでは繰り返しにペナルティが課されない

##### 6. Presence penalty (0.0)

既出の単語を再度使用する際のペナルティの強度を制御。値が高いと新しい単語が優先的に使用される

##### 7. Repetition penalty (1.0)

値が高いほど繰り返しを避ける。値が1.0であればペナルティなし

### 設定方針

#### メンタル状態の数値化：デフォルト値

雑談型：Temperature高め、presence penalty高めにすることで多様な回答を期待

問題解決型：Top-k、frequency penaltyを下げることで論理的で一貫性のある回答を期待

#### 3) プロンプトによりモデル間の差異を補完

「より共感的」に「ユーザーの感情を受け止めてください」などをプロンプトに与えておくことで、問題解決に向くLLMでも雑談対応もある程度できる見通し。その効果、利用時にユーザのストレスを回避できるか否かは評価を行う必要がある

## まとめ、今後の考慮点