# ローカルLLMで
# キャラ会話を作る

## — 6モデル比較検証レポート —

RTX 4070 Ti (VRAM 12GB) の一般家庭向けゲーミングPCで
キャラクター会話AIに最適なローカルLLMを探す試み

本レポートは、ローカル環境で動く対話AIキャラクター
「SEKAI」「つばさ＆なぎさ」の頭脳として、
6つのローカルLLMを実際に試した記録です。
各モデルの長所・崩れ方・最終的にたどり着いた設定を
そのまま公開します。同じ環境で挑む方の参考になれば幸いです。

発行日: 2026年6月22日

ぽすとそに工房

# 1. 背景・動機

きっかけは、ひとつの問いだった。かつて夢中になった「会話形式の恋愛シミュレーションゲーム」を、現代のAIで超えられないか。決まった選択肢を選ぶのではなく、本当に言葉が返ってくる相手と、自由に掛け合いができたら。それも、クラウドの大規模AIに頼らず、自分のPCの中だけで完結する形で。

ただし現実的な制約があった。手元のPCは RTX 4070 Ti、VRAM は
12GB。決してハイエンドではない、ゲームをそこそこ楽しむ一般家庭のPCである。高度すぎる処理はできない。だが、逆に言えば――この制約の中で形にできれば、同じくらいの環境を持つ多くの人にも届くはずだ、とも考えた。

## 作った2つのキャラAI

本検証の舞台となったのは、自作した2種類のローカルAIキャラクターである。性格が違うので、求められるモデルの資質も異なる。

・SEKAI … ユーザーと1対1で向き合う、会話形式のチャット型AI。1人のキャラと対話する。
・つばさ＆なぎさ …
姉妹2人の掛け合いを見て楽しむ観賞型。自動で進む会話に、ユーザーが割り込むこともできるチャット型AI。

特に「つばさ＆なぎさ」は、2人を演じ分けながら掛け合いを成立させる必要があり、1対1のSEKAIより負荷が高い。本検証で各モデルの実力差がはっきり出たのは、主にこの掛け合い形式だった。

# 2. 目的と検証環境

## 目的

キャラクター会話に最適なローカルLLM(頭脳)を見つけること。求める条件は、(1)日本語が自然で安定している、(2)キャラの感情表現や掛け合いに「味」がある、(3)VRAM 12GB
に音声合成と同居できる、(4)Ollama で扱える GGUF 形式であること。

## 検証環境

| | |
|---|---|
| GPU / VRAM | NVIDIA GeForce RTX 4070 Ti / 12GB |
| LLM実行基盤 | Ollama (各モデルを GGUF で導入) |
| 音声合成 | Irodori-TTS(ローカル・GPU上で同居。声クローン) |
| 構成 | 頭脳(LLM)と声(TTS)を同じ12GBのVRAMに同居させる |
| 用途 | キャラクターの掛け合い・1対1会話 |

ポイントは「同居」である。会話のたびに、まず頭脳(LLM)が返事を考え、次に声(TTS)が音声を作る。両方を12GBのVRAMに収める必要があり、モデルのサイズ選びは常にこの制約と隣り合わせだった。

## 3. 検証した6モデルの結果

実際に試した6モデルを、導入順に並べる。判定は「つばさ＆なぎさ」の掛け合いを中心に、日本語の安定性とキャラの味で評価した。

### ① gemma4:12b（当初の頭脳）

ベース: Gemma系 / 12B　判定: 基準・日本語は安定

日本語は安定していて崩れない。ただしキャラの感情表現(味)は控えめで、会話が同じ話題を巡る「ループ」が起きやすかった。この弱点が、別モデルを探す出発点となった。

### ② MN-Violet-Lotus-12B（英語RP特化）

ベース: Mistral Nemo / 12B(Q4_K_M, 約7.5GB)　判定: 不採用・日本語崩壊

感情表現・キャラの味は素晴らしかった。しかし日本語が崩壊。英語混じり(例:「especially飛行機は」)、話者の入れ替わり(つばさの台詞がなぎさの枠に出る)、内部の指示文をそのまま読み上げる、といった症状が頻発。配信用途には不安定すぎると判断。

### ③ cyberagent 日本語Nemo（日本語特化）

ベース: Mistral Nemo(日本語継続学習) / 12B(Q4_K_M, 約7.5GB)　判定: 採用・本命

日本語が安定。英語混じり・話者入れ替わり・指示文の読み上げが解消し、味もそこそこ残る。弱点は会話ループ。これは後述のパラメータ調整で大きく改善できた。最終的にこのモデルを採用。

### ④ Vecteus-v1（日本語小説特化）

ベース: Mistral / 7B(Q5_K_M, 約5.1GB)　判定: 不採用・掛け合い不向き

日本語は自然だが、長文の小説生成に向いた性質で、短いテンポの掛け合いには合わなかった。

### ⑤ Magnum-v4-12b（英語RP特化）

ベース: Mistral Nemo / 12B(Q4_K_M, 約7.5GB)　判定: 不採用・崩れ

発言の先頭に毎回「ぽすとそに:」のような話者ラベルが付く、片方が一方的に長く話し続ける、といった崩れが出た。英語向けモデルの日本語の弱さが表れた。

### ⑥ ArrowPro-7B-KUJIRA（日本語会話特化）

ベース: Mistral系 / 7B(Q4_K_M, 約4.4GB)　判定: 不採用・崩壊

日本語特化だが、2人を演じ分ける複雑な掛け合いでは文として成立しなくなった。実例:「私の番ですね。話します。なぎさです。やったぁー!お兄ちゃん、おはようおは!」。7Bでは複雑な役割設定に力不足と思われる。

## 4. わかったこと

6モデルを試した結果、はっきりした傾向が見えた。

・英語のロールプレイ特化モデルは、味は濃いが日本語が崩壊する(②⑤)。
・小説特化モデルは、長文は得意でも短い掛け合いのテンポには合わない(④)。
・7Bの小型モデルは、単発の応答はできても、2人を演じ分ける複雑な設定では崩れやすい(⑥)。
・12Bクラスの日本語特化モデルが、安定性と味のバランスで最も優れていた(③)。
・掛け合い(複数キャラ)は、1対1より格段に難しく、モデルの実力差が出やすい。

つまり、限られた12GBのVRAMでキャラ会話を成立させるには、「英語が得意」でも「小説が書ける」でも「軽い」でもなく、『日本語に最初から強く、会話の演じ分けにも耐える12Bクラス』が要だった。

## 5. 最終的な解決

本命の日本語Nemo(③)に戻り、唯一の弱点だった会話ループと、その後に現れた細かな崩れを、モデルを変えずに「パラメータ」と「プロンプト」と「後処理」の3つで詰めていった。

### 最終パラメータ(Ollama)

| | |
|---|---|
| モデル | cyberagent-Mistral-Nemo-Japanese-Instruct-2408 (Q4_K_M) |
| temperature | 0.75 (高すぎると言葉が崩れ、低すぎるとループ。中間点) |
| repeat_penalty | 1.15 (同じ言い回しの再利用を抑える) |
| top_p | 0.95 |
| num_ctx | 4096 |
| num_predict | 140 (長台詞を物理的に防ぐ上限) |
| 渡す会話履歴 | 直近6発言のみ (古い話題に引きずられるのを防ぐ) |

### プロンプトと後処理の工夫

・プロンプトは欲張らず簡潔に(1人分・1〜2文・日本語のみ・話を進める・話題に追従)。
・口癖(「えへへ」)の出過ぎ → 後処理で確率的に間引く。
・2人分を一度に書く暴走 → 後処理で2人目以降を切り捨て、1発言だけ残す。
・文の途中切れ → 最後の完結した文までで整える後処理。
・固有名詞の化け(「ぽすとそに」→「ボスとソニ」等) → 化けパターンを正しい表記に補正。

重要なのは、すべてをプロンプトに詰め込まないことだった。指示を足しすぎると、モデルが処理しきれず破綻する。『モデルにお願いすること』と『コードで確実に処理すること』を切り分けたことで、安定した。

## 6. まとめ

6つのモデルを巡る回り道の末、結論はシンプルだった。派手な海外製のロールプレイモデルでも、最新の大型モデルでもなく、『日本語に強い12Bモデルを、丁寧にチューニングして使う』こと。そして、モデル選びと同じくらい、プロンプトの引き算・パラメータ調整・後処理が効いた。

RTX 4070 Ti / VRAM 12GB
という、特別ではない環境でも、キャラクターと自然に掛け合えるローカルAIは作れる。完璧ではなく、ときどき不思議な発言もする。だが、その『揺らぎ』こそ、決まった選択肢を選ぶだけの旧来のゲームにはなかった、生きた会話の手応えでもあった。

同じPCの前で、同じ夢を見ている人へ。この記録が、最初の一歩の地図になれば嬉しい。

# 著作権・利用規約

【許可される利用】

・個人での閲覧・学習
・YouTube等での紹介・解説(収益化含む) ※事前にお問い合わせより連絡必須

【禁止される利用】

・無断転載(SNS・ブログへの全文コピー)
・商用目的での再配布・販売
・著作者名を削除しての二次配布

ぽすとそに工房

ホームページ：https://postsoni.github.io/
お問い合わせフォームよりご連絡ください。

SNS

X (Twitter)：@postsoni
YouTube：@postsoni
note：postsoni

---