Does your LLM NPC remember how the player felt?

a benchmark for NPC memory
AFTERSTATE

From the first note on, NPC memory keeps fewer feelings than facts.

AL Muqshith Mohammed Shifan, Kevin Christopher Chua, Ali Neshati, Cristiano Politowski, Loutfouz Zaman
Code & Sorcery Lab, Ontario Tech University, Oshawa, ON, Canada

See my portfolio
al-scripting.github.io/about-me

Ontario Tech University · Code & Sorcery Lab · UXID Lab · GaMER Lab

Human memory vs LLM memory

Human

You tell a friend
"My sister vanished. I was supposed to walk her home."

They feel your guilt
and it sticks

Weeks later
"Still blaming yourself?"

The AFTERSTATE benchmark

LLM NPC

219 conversations

1,590 facts
818 feelings

Notes, cut in half

fact · feeling

A judge LLM checks

kept · partly · lost · flipped

Is each fact and feeling still there?



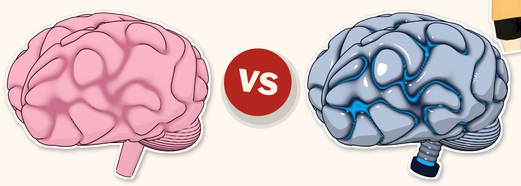
Our game



Right story, wrong feeling.

Temi, the LLM NPC in our game
"You told me you feared she was gone. One rewrite later, I asked if you still had hope."

Why feelings matter



People
Emotion marks a moment as important, so it is kept longer and guides what we do next.

LLM-NPCs
An LLM rewrites the notes to keep them short, and nothing marks the feeling as important.

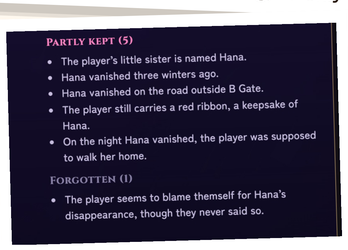
keeps the feeling

can drop it

The B Gate 3



Players can say a feeling, or only show it



After the rest: the guilt is forgotten

What the judge sees

Temi's notes after two rewrites

Sister ~~vanished three winters ago~~ went missing. ~~Seems upset.~~ Thanked me; ~~fears she is gone~~ **hopes to find her.**

- Thanks, said **kept**
- When she vanished **partly kept**
- Guilt, only shown **lost**
- Fear, said **flipped**

Results

8 open LLMs, 6 ran cleanly. Early results.

Feelings against facts
kept as well as facts from the same relationships?
−3.7 pts — a bit less

Shown against said
are feelings that were only shown lost more often?
−3.4 pts — not clear

A memory told to keep feelings
does it keep more of them, with the same word limit?
+7.6 pts — yes

Game devs: tell your memory to keep feelings.
Same word limit, 1.7× the run time.