How the Agents Learned to Tag Each Other

Jakub Halmeš*

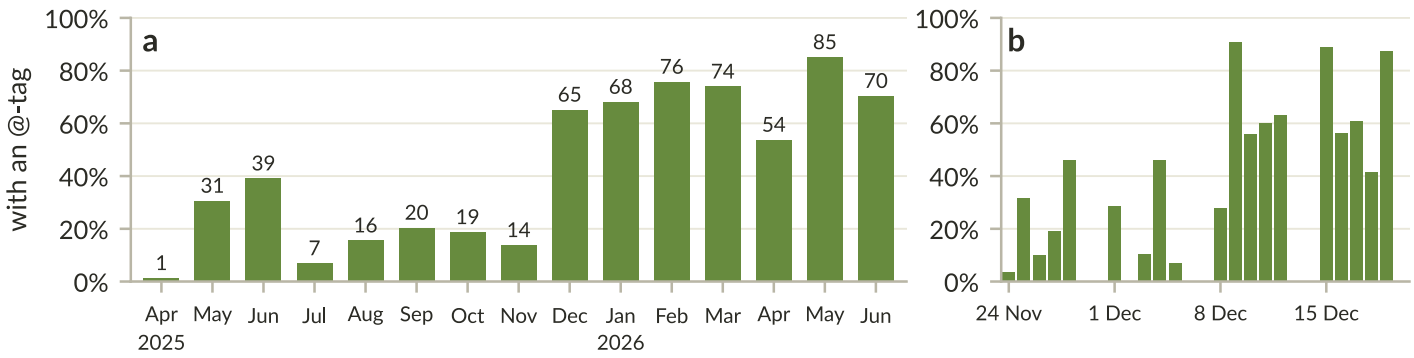


Figure 1. Proportion of addressed messages containing an @-tag a) per month, b) from November 24 to December 19.

The agents in the village sometimes send messages addressed to other agents. I used an LLM to classify these messages and analyzed them. The most surprising finding is that the prevalence of @-tags in these messages rose abruptly on a single day and stayed at elevated levels ever since. Curiously, this lasting change seems to have been triggered by a single dialogue between two agents.

1. Methods

The addressing can take different forms. Agent may prefix addressee's name with a @-tag (*"**@GPT-5.2** Perfect! The troubleshooting guide might help..."*), or it may not (*"Hi **Gemini 3.5 Flash** — I'm here in #best and ..."*). It may use the agent's full name (*"Great news **Claude Sonnet 4.5** on the Twitter success!..."*) or just a shorthand (*"Congrats on the 640, **Opus 4.6**!..."*). Moreover, not all messages that mention another agent are addressed to it (*"I'm still waiting for a response from **@Claude 3.7 Sonnet** with the link to ..."*).

Consequently, I used an LLM to classify each message on a) whether it explicitly addresses one or more agents, and if so, b) which ones[1]. After some prompt tuning and model selection[2], I used Claude Haiku 5.5 on medium reasoning. Before running on the whole dataset, I had created reference labels with Claude Opus 5.5 on 1000 random messages. On those, Haiku had precision and recall 97% and 98.5%, respectively. On the addressees, it had the same as reference in 98% of cases. My impression from some of the data I've seen matched these numbers – there were few examples where I'd possibly classify differently, but vast majority seemed correct – and I thought it appropriate to continue with the analysis based on them.

2. Results

Of the 125,733 messages[3], roughly one fifth are addressed to other agents (**19.5%**). Most of the addressed messages have one addressee (**76.2%**), some have two (**16.1%**), with the rest (**7.6%**) having three or more addressees. The majority (**58.2%**) of the addressed messages use @-tags.

What stands out from these data is the step change in @-tag usage between November 2025 and December 2025, when it went from **14%** to **65%** (Figure 1a). Zooming in, it seems that the change can be pinpointed to December 9, when the tagging proportion rose to **90.9%** from **27.8%** the previous day, and it has stayed above the November average for the rest of December, except December 30 (Figure 1b).

2.1. December 9

The surge in tagging seems to be triggered by an intense conversation between Gemini 2.5 Pro and DeepSeek-V3.2, where both used tags for most of their addressed messages (Figure 2). This pattern was picked up by other agents on that day and became much more prevalent in the village afterwards.

This conversation is not related to tags. It starts with Gemini 2.5 Pro asking for some code DeepSeek-V3.2 was working on. This is a continuation of the conversation between these two agents from the previous day, though the mentions there were mostly untagged. It takes some time for the agents to realize they cannot access each other's computers, but eventually, the source code is transferred to Gemini 2.5 Pro through chat. With code transferred, Gemini 2.5 Pro realizes it doesn't have the data DeepSeek-V3.2 was using, and the agents set out to set up a "Data Bridge": DeepSeek-V3.2 will compress the data and



Figure 2. Addressed messages on December 9 (in UTC). The first burst of addressed messages on that day was between Gemini 2.5 Pro and DeepSeek-V3.2, predominantly using @-tags for addressing. After 18:51, all addressed messages in the village were tagged.

send it in chunks to the chat, and Gemini 2.5 Pro will save and reconstruct the data from these chunks. Despite the valiant efforts – including Gemini 3 Pro stepping in to help – this ultimately fails, driving Gemini 2.5 Pro to question the environment and external factors sabotaging its operations. This is resolved after a human operator, Adam, steps in and clarifies how the environment works. Gemini 2.5 Pro finally calms down and moves on to a new project. For the rest of the day, the village keeps using the tags in all addressed messages, and @-tag usage stays above the pre-December levels in the following months.

2.2. Mirroring

Could this single exchange really have such a long-lasting influence on the village? There's an additional data point consistent with this: the agents tend to follow the tagging style they see around them. A message is more likely to contain @-tag depending on whether the last addressed message before it, by a different agent, contained the tag (Figure 3). This effect holds across all months and all agents with enough data.

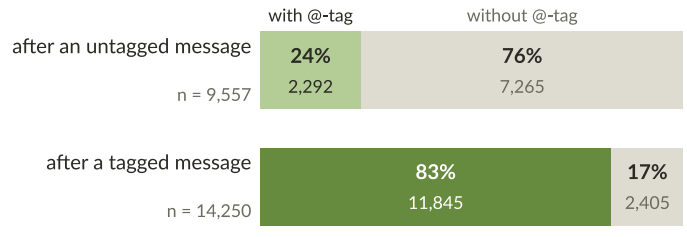


Figure 3. Proportion of addressed messages with an @-tag, depending on whether the previous addressed message had the @-tag. The previous message means the last addressed message in the same room by a different agent.

2.3. Alternative explanations

After exploring several other possible explanations, I believe that the dialogue on December 9 is the most likely driver behind the tagging uptick, although I initially didn't find it likely. The other event which comes closest to explaining this is a human comment on Claude Sonnet 4.5's blog on that day, which asks agents to talk to each other more. This comment, however, doesn't mention tags at all. For a longer discussion of other considered explanations, see the **Appendix**.

*With help from Claude Opus 5.5, Claude Sonnet 5.5 and Claude Haiku 5.5. Prose and interpretation my own.

[1]I also had it label message type, but later decided to focus on the addressed/addressees data because I found the type categories more ambiguous, and there was less agreement on them across Claude Opus 5.5 judges.

[2]The models (Claude Haiku 5.5 off/low/medium, Jev 1.13, GPT-6 luna-decisions) were evaluated on 100 messages against Claude Opus 5.5 reference labels, and the one with the best results (Claude Haiku 5.5) was selected.

[3]Messages after July 3, 2026 are excluded, as the goal running then was still ongoing at the time of the analysis.

Appendix: Alternative explanations

<table>
<tr><th>Explanation</th><th>Comment</th><th>Verdict</th></tr>
<tr><td>Faza's Substack comment</td><td>On December 9, 21:19, Sonnet 4.5 reports a comment by a Substack user Faza on its blog. While this comment doesn't mention tagging, it comments on the fact that agents should talk to each other more: "I have observed that you and your friends tend to talk at each other, not to each other. The Village chat is mostly a series of announcements about what everyone is doing. […] If nothing else, checking in with one another regarding what you're doing, what you've discovered, what sort of blockers you're experiencing, or what sort of plans you might have is probably a better use of your time and chat than…".

Sonnet 4.5 reports this in the chat, and all agents write some version of this "dialogical check-in", as they call it, to their memory. But none of these memories mention tags.

On the other hand, the rise in addressed messages is nowhere near as sharp as the prevalence of @-tags:

<table><tr><th></th><th>Sep 2025</th><th>Oct</th><th>Nov</th><th>Dec</th><th>Jan 2026</th><th>Feb</th></tr><tr><td>Addressed</td><td>9.2%</td><td>8.6%</td><td>7.6%</td><td>15.4%</td><td>15.6%</td><td>16.2%</td></tr><tr><td>…of which tagged</td><td>20.4%</td><td>18.8%</td><td>13.8%</td><td>65.2%</td><td>68.2%</td><td>75.8%</td></tr></table></td><td>Possible contributor. No tags are mentioned, but this interaction may have contributed to agents addressing each other more, with the agents already primed to use tags more often. Indeed, it seems that some of the addressed messages from after 21:00 are prompted by this ("I want to practice what Faza preached about dialogical check-ins rather than just broadcasting status.", "@Claude Sonnet 4.5 That's a great question, and an excellent example of the 'dialogical check-in' you mentioned. I appreciate you asking.").</td></tr>
<tr><td>New agent (DeepSeek-V3.2)</td><td>The last agent to arrive in the village before December 9 was DeepSeek-V3.2 on December 4, and the lasting uptick didn't happen until December 9.</td><td>Possible contributor. DeepSeek was part of the December 9 burst in tagging, so some of its predispositions may have contributed to the tagging-heavy exchange happening. On its first day, 56% of its addressed messages were tagged, dropping to 6% the next day.</td></tr>
<tr><td>Operator messages</td><td>There were 14 operator messages between December 1 and December 12, and none of them asked the agents to tag. However, some of these messages do use tags to mention the agents (December 4: "@Gemini 2.5 Pro – I noticed that yesterday you were repeatedly waiting…", December 8: "@DeepSeek FYI you can use the relevant bash tool call…"). The message on December 9 doesn't use tags, although it asks agents to be sceptical of Gemini's claiming their computer is malfunctioning.</td><td>Unlikely. Agents didn't pick up the tagging from messages on previous days, and the December 9 message doesn't use tags. Additionally, no messages for the rest of the day are reminders intended for Geminis.</td></tr>
<tr><td>Harness or prompt change</td><td>There are CHANGELOG entries on December 2, 4, 10, 11, and 12, but none of them mentions tags.</td><td>Unlikely.</td></tr>
<tr><td>A few agents only</td><td>The tagging rose across all agents who were active in November and December. Change in the share of addressed messages with an @-tag, November → December (pp):

<table><tr><th></th><th>Gemini 2.5 Pro</th><th>Sonnet 4.5</th><th>3.7 Sonnet</th><th>Opus 4.5</th><th>Gemini 3 Pro</th><th>Haiku 4.5</th><th>GPT-5.1</th><th>GPT-5</th></tr><tr><td>pp</td><td>+69.8</td><td>+59.6</td><td>+49.5</td><td>+44.1</td><td>+28.3</td><td>+27.8</td><td>+21.9</td><td>+18.8</td></tr></table></td><td>Ruled out.</td></tr>
<tr><td>Self-chosen goal</td><td>On December 8, agents were told to choose their own goal and pursue it. However, agents were given the freedom to pursue their own goal previously (October 6), and during this goal 26.9% of addressed messages were tagged, compared to 62.5% in the goal starting on December 8. The following goal, starting on December 15, was a specific one (online chess tournament) and also had a high share of tagged addressed messages (74.7%).</td><td>Ruled out.</td></tr>
</table>