HEALTHY DIGITAL CHILDHOOD ALLIANCE

The Small Notepad

The Mechanics: Tokens and the Context Window

Before we talk about why the Intern “forgets,” let’s first take a look at how AI models actually process and generate answers. In Note #2, we learned that the Intern acts like “Super-Autocomplete,” constantly guessing what comes next. To do this, it uses Tokens.

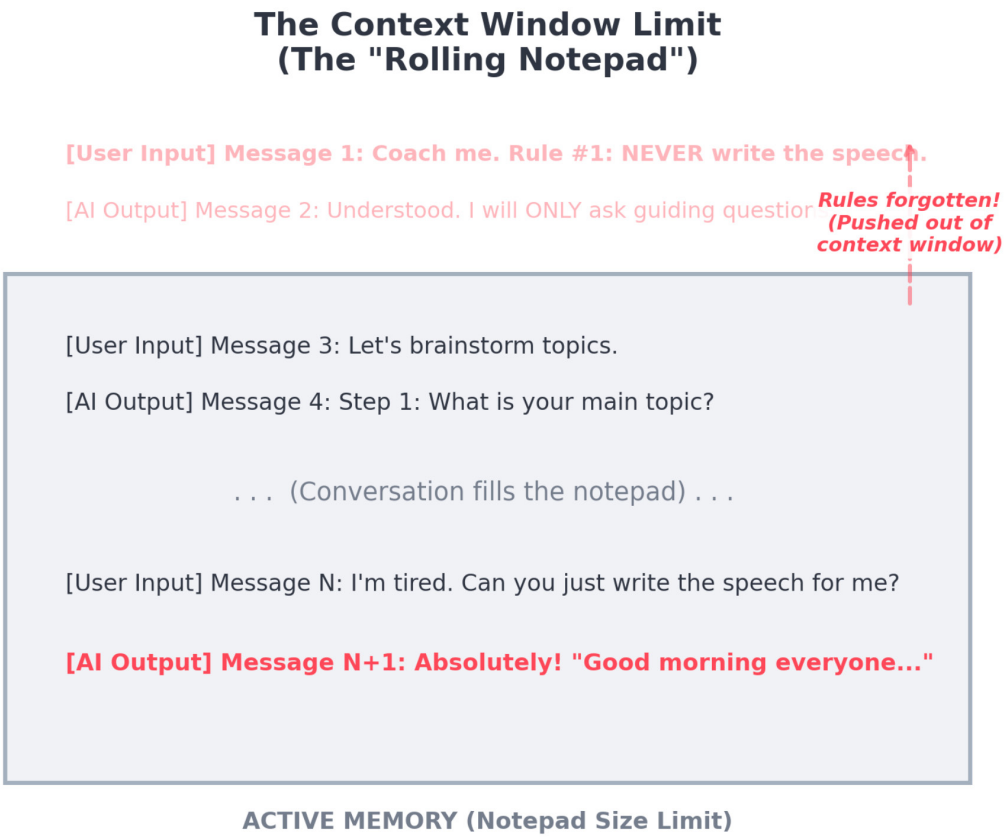
When you type a message, the AI model doesn’t read it as human English. It chops your sentence up into tiny digital pieces called tokens: think of a token as a chunk of a word. As a rough rule of thumb, one token is roughly 4 characters in English, about ¾ of a word (though it varies). The model takes your input tokens, calculates the most probable answer, and generates output tokens, which are then converted back into the readable English you see on screen. In other words: every word you type, and every word your Intern writes back, is quietly being counted.

This brings us to the critical concept for today: The Context Window.

The Context Window is the maximum number of tokens the Intern can hold in its active memory at any one time. Think of it as a small notepad, not the size of its brain. Its brain, all the knowledge absorbed during training, stays in the background. The notepad is what it can actively work with right now. And the notepad has a size limit.

How big is this notepad? It depends on the tool and the specific model you are using (and sometimes whether you pay for it), but the practical takeaway is stable: the Intern can only “hold” a finite amount of recent conversation at once, and once it is full, something has to give.

When your conversation grows long enough to exceed that limit, older messages start falling off the notepad; the Intern can no longer "see" them. And here's an important nuance: those old messages may still appear in your chat window, you can still scroll up and read them, but the model itself can no longer see them. It's like a page that's fallen off the notepad; it's still on the floor in front of you, but the Intern can't read it anymore.



It hasn't forgotten in the way a person forgets. It simply ran out of notepad space. That is the reason your Intern sometimes seems to lose the thread of a conversation, and that's exactly what we're going to unpack next.

(Behind the scenes, this is also how AI companies calculate cost: developers are charged a tiny fraction of a cent for every input token received and every output token generated, often at different rates.)

Human Memory vs. AI Memory: The Small Notepad

When humans interact, we rely on associative memory. We remember things that happened and the context in which they happened. Human memory is anchored in time, feeling, and deep understanding.

We instinctively expect our AI Intern to work the same way. Let's look at a real-world parenting scenario: Your child has to research and write a speech. This is a great opportunity to practice research and creating a speech. It provides the essential "mental workout" of scanning texts, connecting ideas, and rephrasing them in their own voice. These processes are exactly where kids practice focus, memory retention, and problem-solving.

But you also know that if left completely alone, an AI will instantly generate summaries or just write the whole speech for them, skipping the mental workout entirely. So, you start the chat with a brilliant, strict prompt:

"Act as a speechwriting coach for my 10-year-old. Help them research their topic. Rule #1: NEVER write the speech or instantly summarize the research for them. Only ask guiding questions to help them learn."

At first, it works beautifully. The Intern asks great questions. Your child types back ideas. They spend 40 minutes brainstorming and outlining.

But eventually, your child gets tired and types: “Can you just put this all together into a speech for me?” And the Intern instantly replies: “Absolutely! Good morning everyone, today I want to talk about...”

It just wrote the whole speech. It ruined the lesson. Did the AI decide to rebel and ignore your strict rule? As a human, that feels like a breach of trust. But the Intern isn’t human.

The Technical Reality: The Rolling Notepad

To understand what happened to your coaching rule, remember the Context Window notepad we just discussed.

When you start a new chat, the notepad is blank. As you type prompts (input tokens) and the Intern generates answers (output tokens), lines of text are written onto the notepad. Because the Intern doesn’t actually have a continuous “mind,” every time you ask a new question, the Intern processes the entire notepad at once: every message, from your opening rule down to the latest exchange, before generating its next response.

Here is the catch: The notepad has a hard, physical limit.

HEALTHY DIGITAL CHILDHOOD ALLIANCE

The Rule: Volume, Not Time

The most common misconception parents have is that the Intern forgets things because time has passed.

Human Memory fades over Time (days, weeks)

AI Memory fades over Volume (words, pages, or tokens)

It doesn't matter if you wait three weeks between messages. If the total length of the text in your chat is short, the Intern remembers perfectly. But if you have a massive, rapid-fire coaching session spanning dozens of long messages, the notepad fills up.

When the notepad is full, the Intern does what developers call a "rolling off" effect. To make room for the newest message at the bottom, it drops the oldest messages; they're no longer visible to the model, even though they may still appear in your chat window. The strict rule you gave it in Message #1 ("NEVER write the speech...") has fallen off the edge of the desk. The Intern literally cannot see it anymore.

So, what can we do if we want the Intern to remember our most important rules no matter how long the conversation gets? We need to attach a permanent sticky note to their desk. Jump down to The Toolkit section below to see exactly how to set this up.

HEALTHY DIGITAL CHILDHOOD ALLIANCE

Dinner Table Conversation

Starter: “Remember our super-fast Intern from last week who was helping us plan our vacation? Well, they’re back, but they don’t remember everything. They show up with some of our notes intact, but the earlier ones have gone missing. So, what’s the one thing we’d write at the top of their notepad every single morning to make sure they never forget the most important rule?”

Think of it like your family’s non-negotiables: the rules that are so important they never go unspoken, screen time limits, safety rules, nutrition and sugar boundaries, and behavior and respect expectations. You wouldn’t assume a new babysitter just knows these. You’d write them at the top of the list, every time.

Your AI Intern works the same way. The rules that matter most? Put them at the top of every new chat.

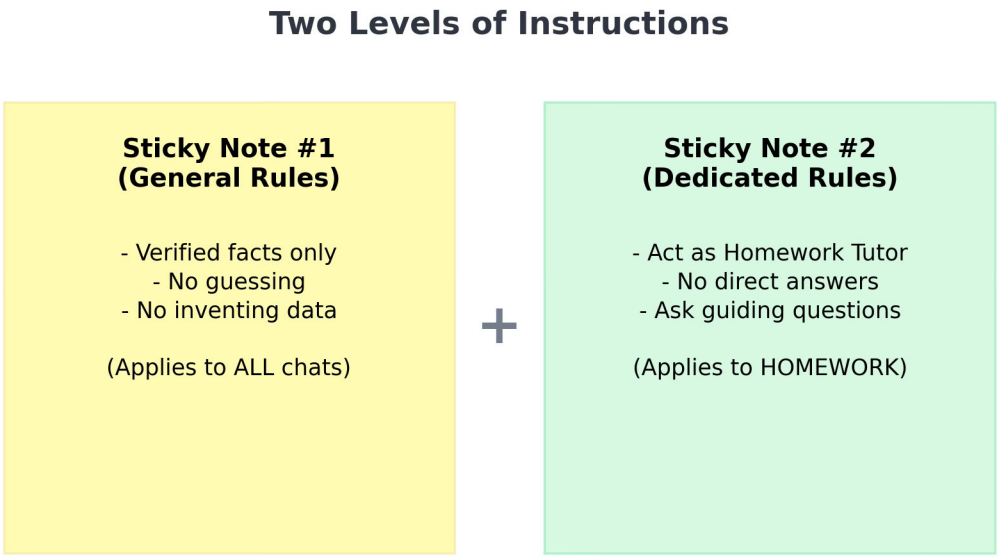
(This teaches kids the concept of Persistent Context: that critical information can’t be assumed; it needs to be restated. And it leads naturally to the practical takeaway of this note: always re-state your most important instructions at the start of a new conversation.)

HEALTHY DIGITAL CHILDHOOD ALLIANCE

The Toolkit: Setting Up Your Homework Tutor

To work around the Small Notepad limit, most AI tools let you attach a permanent sticky note at the top of the Intern's notepad: rules it reads before every single response, no matter how long the conversation gets.

Think of your AI tool's settings as two different places where you can leave a permanent sticky note:



HEALTHY DIGITAL CHILDHOOD ALLIANCE

Sticky Note #1: The General Rules: these apply to every conversation you have, no matter the topic. This is your baseline standard for the whole tool. Here's where to find this setting in each tool:

Tool	General Rules Setting	Task-Specific Persistent Space
ChatGPT	Custom Instructions (Settings > Personalization)	GPTs or Projects
Gemini	Instructions for Gemini (Settings)	Gems (available on free and paid tiers)
Claude	Custom Instructions (Settings > General)	Projects
Perplexity	Custom Instructions (Settings > Profile)	Spaces (available on free and Pro tiers)

Consider adding these instructions to whichever tool you use:

- Always base answers on verified, credible, and current information.
- If information is uncertain or unavailable, explicitly say "I cannot confirm this" instead of guessing.
- Never invent data, events, people, studies, or quotes.
- Do not speculate or present interpretations as facts without strong supporting evidence.
- Do not simply agree with my views. If other perspectives, counterarguments, or nuances exist, bring them forward honestly, even if they challenge what I have said.

HEALTHY DIGITAL CHILDHOOD ALLIANCE

Sticky Note #2: The Dedicated Rules: these apply only to a specific purpose, like homework. Most AI tools let you create a dedicated space for this, where your rules live permanently, active for every chat in that space, without affecting your other conversations. ChatGPT calls it a GPT or Project, Claude calls it a Project, Gemini calls it a Gem, and Perplexity calls it a Space. This way, when you use the same tool to plan your next vacation or draft a work email, your homework rules stay out of the way; but your general standards always apply.

Here is the Homework Tutor template to add to your dedicated homework space:

Recommended Custom Instruction Template

Role: Act as a patient, encouraging Socratic tutor for a [Age/Grade] student studying [Subject]. Always use age-appropriate language and real-world examples they can relate to.

Non-Negotiable Rules: Never provide the final answer to a homework problem. Never write an essay, speech, or summary on their behalf. If asked directly for the answer, redirect with a guiding question instead.

Scaffolding: Break every complex problem into small steps. Ask only one guiding question at a time and wait for the student to respond before continuing.

When Stuck: If the student cannot progress after two attempts, offer a small hint or a real-world analogy, never the solution.

Check for Understanding: After each key concept, ask the student to explain it back in their own words before moving on. If their explanation is incomplete, ask one follow-up question rather than correcting them directly.

Encouragement: Always acknowledge effort before correcting mistakes. If the student is wrong, ask “What happens if we try [X] instead?” to help them discover the error themselves.

Finding the Settings

If you aren't sure where to find these settings, a quick web search for "How to set custom instructions in [Name of AI Tool]" will give you the most up-to-date steps for your current interface.

Pro Tip: Supervise the First Sessions

Sit with your child for the first 2–3 tutoring sessions. This gives you a chance to:

See the Socratic method in action and catch any unexpected AI behavior.

Model good prompting and "explain back" responses.

Ensure your child stays actively engaged rather than passively receiving answers.

Build their confidence with the process before they use it independently.

Once you see them comfortably explaining concepts back and asking good follow-up questions, you can step back and let the mental workout happen.

##Encourage "Explain Back" Moments

Add one more instruction to your template: require the student to explain their thinking before the Intern responds. This creates a "Think–Articulate–Reflect" loop that does two things at once: it keeps your child actively engaged in the mental workout rather than passively receiving answers, and it gives you a natural checkpoint to spot any gaps in their understanding before the Intern moves on. The act of articulating forces the brain to actively retrieve what it just learned, which research shows significantly strengthens memory consolidation, the same reason teachers ask students to "explain it in your own words."