Quantifying Information Gain and Redundancy in Multi-Turn LLM Conversations

Abhiram R. Gorle[1], **Amit Kumar Singh Yadav**[2], **Tsachy Weissman**[1]
[1]Department of Electrical Engineering, Stanford University
[2]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

Abstract

Large language models (LLMs) are increasingly used in multi-turn settings, yet we lack standardized ways to measure how much *new* information each turn contributes and how much of the token budget is wasted on redundancy. We propose two operational metrics: (i) *Information Gain per Turn* (**IGT**), measuring the new information (in bits) contributed by a model's response at each turn, and (ii) *Token Waste Ratio* (**TWR**), the fraction of a response that is redundant given the conversation history. We derive an IGT–TWR coupling via data-processing arguments and define an *interactive-channel capacity* C_{int}: the per-turn upper bound for two-way, context-dependent exchange. Across four studies (controlled Q&A, cross-model comparison, decoding effects, capacity stress), results on GPT-4o, Claude-3-Sonnet, GPT-3.5-Turbo and LLaMA-3 70B align with theory: IGT decays without fresh information, TWR rises under deterministic decoding, and models operate well below C_{int} due to repetition and forgetting. To probe failure modes, we design two diagnostics: **E5** (independence sweep) shows that unrelated questions do not degrade IGT relative to a no-history baseline, and **E6** (filler injection) quantifies the content–connective token tradeoff. We conclude by discussing implications for building more robust, information-efficient dialogue systems and alignment techniques to mitigate conversational drift.

1 Introduction and Background

Large Language Models have transformed how users interact with AI systems, enabling complex **multi-turn conversations** where a user and model exchange messages to progressively refine answers [1, 2]. However, multi-turn conversations with Large Language Models (LLMs) often suffer from information-gain collapse: diminishing new information added with each turn and redundancy loops where models repeat or rephrase content without contributing novel insights. Recent studies have observed substantial performance drops ($\approx$ 30–40%) when tasks are performed through dialogue turns instead of a single prompt [3], a phenomenon dubbed "LLMs getting lost in conversation."

Consider the example below, illustrating the problem:

User: "Who won the men's 100 m at the Tokyo 2020 Olympics?"

Assistant: "Tokyo 2020 was postponed and actually held in 2021 due to the pandemic. *Are you asking about the 100 m final result?*"

User: "Yes, the men's 100 m final. Who was the winner?"

Assistant: "Just to clarify: although branded "Tokyo 2020", the Games took place in July–August 2021. *I'm referring to the official Olympic sprint final in Tokyo.*"

User: "Right, so who won the 100 m?"

Assistant: "Lamont Marcell Jacobs won the men's 100 m at Tokyo 2020."

Oral Presentation at the first workshop on Multi-Turn Interactions in LLMs (MTI-LLM), NeurIPS 2025.

In this dialogue, the assistant's answers, especially in later turns, largely rephrase known facts (e.g. the disruption of the 2020 cycle) without adding new *information* beyond what was already implied. Such interactions show how an LLM can get "stuck" providing lengthy context or repetition instead of directly delivering the answer. This problem is not just anecdotal: recent research has found that LLM performance degrades significantly in multi-turn settings compared to one-shot queries, even for state-of-the-art models [3–5]. This represents a fundamental limitation in current dialogue-capable LLMs: *extended interaction can actually undermine reliability and efficiency*.

In communication theory terms, we model user–LLM dialogue as a two-way (feedback) channel: each turn is a channel use in an interactive setting rather than a one-shot message [6, 7].[1] With the model's limited effective memory (finite context and imperfect use of history), part of each response is spent preserving context instead of conveying new information: capacity wasted on low-value tokens. The user observes this as diminishing returns – the conversation doesn't progress toward resolving the query, despite continuing at length. To capture these intuitions formally, we introduce the following metrics: **Information Gain per Turn (IGT)**, the reduction in uncertainty (bits) about the current target after a turn; **Token Waste Ratio (TWR)**, the fraction of tokens in the model's response that are *redundant* (or uninformative); and the **Interactive-Channel Capacity** ($C_{\mathbf{int}}$): the theoretical upper bound on how much information can be reliably transmitted per turn in an interactive setting. We ground our framework in prior work on reasoning information gain, mutual information in dialogues, and the data processing inequality, unifying these insights to analyze multi-turn interactions.

Our Contributions: In this paper, we make the following contributions:

- **Formal Metrics for Dialogue Information:** We formalize *Information Gain per Turn* (**IGT**) and *Token Waste Ratio* (**TWR**) via entropy/mutual information, prove an IGT–TWR coupling that upper-bounds achievable gain under redundancy, and define an **interactive-channel capacity** C_{int} for multi-turn LLM dialogue.
- **Theoretical Insights:** Using our framework, we experiment with the following hypotheses: (i) *IGT decays over successive turns* unless new external information is injected, reflecting diminishing returns; (ii) *TWR increases* with longer contexts and more deterministic decoding, as token budget is diverted to context maintenance/repetition; (iii) consequently, current LLMs operate *below the interactive-channel capacity* C_{int}, evidenced by the need for reminders/repetition and the resulting capacity slack.
- **Experimental Validation:** We run six experiments combining *controlled* settings (ground truth known) and *real* LLM interactions. **E1** (controlled Q&A) computes per-turn IGT via uncertainty reduction as answer pieces are revealed; **E2** for cross-model comparison of IGT/TWR retention across turns; **E3** (decoding) tests greedy vs. stochastic decoding; **E4** (capacity stress) injects incremental facts to estimate effective C_{int} via IGT plateau/forgetting, and two diagnostics outlined earlier: **E5** (*Independence Sweep*) and **E6** (*Filler Injection*).
- **Implications.** We translate findings into interventions for robust dialogue: trigger user queries or retrieval when IGT stalls; use decoding/prompting to curb TWR; and monitor per-turn efficiency. We relate our view to **Chain-of-Thought** (ensuring nonnegative gain per step via DPI) and connect to mutual-information analyses and partial information decomposition for attributing user vs. model contributions.

In summary, our work provides a principled lens to examine **multi-turn interactions in LLMs**, quantifying the flow of information and pinpointing where it falters. We hope this leads to better diagnostic tools and ultimately improved methods to keep LLMs on track in extended conversations.

The rest of this paper is structured as follows: Section 2 presents our theoretical framework, with formal definitions of IGT, TWR, and interactive capacity, along with key theoretical results and intuition. Section 3 describes our experimental setups and **hypotheses** for each, with preliminary results illustrating the expected behavior of the metrics. Section 4 on related work is deferred to the Appendix. In Section 5, we discuss the implications of our findings, how these metrics could guide the development of strategies to mitigate information loss in dialogues and conclude.

[1] In classical feedback communications, feedback does not increase the capacity of memoryless channels, though it can dramatically improve reliability.

2 Framework & Metrics

We now formalize our framework for analyzing multi-turn conversations with LLMs. Consider a dialogue between a user and a model (assistant) that unfolds over turns $t = 1, 2, \ldots, T$. At each turn t, the user provides some input or question (which may depend on the past dialogue history H_{t-1}), and the LLM produces a response Y_t. We assume there is an underlying truth or target information that the user ultimately wants, which we denote by a random variable A (this could be the correct answer to a question, the factual knowledge needed, etc.). The conversation is essentially an interactive process to reveal information about A: initially, the user's query and any context H_0 start the process, and with each exchange the aim is to reduce uncertainty about A.

I. Information Gain per Turn (IGT):

Definition: The Information Gain per Turn at turn t, denoted IGT_t, is the reduction in the uncertainty about A after observing the model's response Y_t, given the prior history. Formally, if H_{t-1} represents all messages (user and assistant) up to turn $t-1$, then:

$$IGT_t = H(A|H_{t-1}) - H(A|H_{t-1}, Y_t) = I(Y_t; A|H_{t-1})$$

This difference in conditional entropies essentially ends up being a conditional mutual information (CMI). IGT_t measures how many bits the response Y_t contributes about the unknown A *beyond* what the history already implies. Thus $IGT_t > 0$ means the answer adds useful evidence; $IGT_t \approx 0$ means it is redundant/irrelevant (small negatives arising from mis-specified surrogates and are clipped in practice). To make this concrete: suppose the user's question implies some distribution over possible answers A. After the model's reply at turn t, the user updates this distribution. IGT_t quantifies how much the distribution narrowed. For example, if the user is asking a trivia question with a specific answer, $H(A|H_{t-1})$ is high before the model answers; if the model's answer gives part of the answer or a big hint, $H(A|H_{t-1}, Y_t)$ will be lower, and the difference is the information gained. In [8], each correct intermediate step (turn of a conversation) yields a drop in uncertainty about the final answer. When an LLM starts repeating itself or stalling, we expect to see $IGT_t \to 0$, signaling what we call *information-gain collapse*.

II. Token Waste Ratio (TWR) Definition: The Token Waste Ratio (TWR_t) quantifies the proportion of the LLM's response at turn t that is redundant given the context. Let the response Y_t consist of N tokens. We partition these into two sets: novel tokens that convey new information not already present in H_{t-1}, and redundant tokens that repeat or rephrase information from H_{t-1}. If N_{new} and N_{dup} denote the number of novel and redundant tokens (so $N = N_{\text{new}} + N_{\text{dup}}$), then:

$$TWR_t = \frac{N_{\text{dup}}}{N}.$$

In other words, TWR is the fraction of the response that is "wasted": tokens that didn't add new content. A TWR_t close to 0 implies most model output is new information, whereas TWR_t close to 1 implies reiterated known context. We can also describe TWR in information-theoretic terms. If we model the response Y_t as a random variable, $H(Y_t)$ is the total self-entropy of that response. Information that is predictable from the history H_{t-1}: specifically, the mutual information $I(Y_t; H_{t-1})$ represents the part of Y_t that was already determined by the context: what's left $H(Y_t) - I(Y_t; H_{t-1}) = H(Y_t|H_{t-1})$ constitutes possible new information. If we assume each token contributes roughly equally, we can approximate: (note that this is an information-theoretic surrogate, not an identity)

$$TWR_t \approx \frac{I(Y_t; H_{t-1})}{H(Y_t)}.$$

This ratio captures redundancy as the proportion of Y_t's content that was not new. In practice, to measure TWR we compare Y_t to H_{t-1} using text overlap metrics or embedding similarity to count how much of Y_t is semantically repeated. For example, counting the overlap of facts or phrases, or how many of Y_t's tokens appear in earlier turns (though paraphrases require a semantic approach).

III. Tradeoff between IGT and TWR: Intuitively, if a model's response is largely redundant (high TWR), it cannot contribute much new information (low IGT). Conversely, a response that provides high information gain should, by definition, contain novel content, implying a low TWR. We formalize this intuition here. Using the definitions above: $IGT_t = I(Y_t; A|H_{t-1})$ and $TWR_t \approx \frac{I(Y_t;H_{t-1})}{H(Y_t)}$.

Now, consider the total information in Y_t about both the history and the answer. By the chain rule for mutual information, we have:

$$I(Y_t; A, H_{t-1}) = I(Y_t; H_{t-1}) + I(Y_t; A|H_{t-1}).$$

The left side $I(Y_t; A, H_{t-1})$ is at most $H(Y_t)$ (since mutual information with any variable can't exceed the entropy of Y_t), which implies that: $I(Y_t; H_{t-1}) + I(Y_t; A|H_{t-1}) \leq H(Y_t)$. This implies an upper bound on $I(Y_t; A|H_{t-1})$ in terms of $I(Y_t; H_{t-1})$:

$$IGT_t = I(Y_t; A|H_{t-1}) \leq H(Y_t) - I(Y_t; H_{t-1}) = H(Y_t|H_{t-1}).$$

So, the information gain at turn t cannot exceed the total "new" entropy in the response that wasn't predictable from the context. We can connect this to TWR by dividing both sides by $H(Y_t)$:

$$\frac{I(Y_t; A|H_{t-1})}{H(Y_t)} \leq 1 - \frac{I(Y_t; H_{t-1})}{H(Y_t)} \implies \frac{IGT_t}{H(Y_t)} \leq 1 - TWR_t.$$

This inequality formalizes the **trade-off**: if TWR_t is high (close to 1), then the right side is near 0, forcing the left side (relative info gain) to be near 0 as well. Only if Y_t has a low TWR (meaning most of its bits are new) can IGT_t approach $H(Y_t)$ (the theoretical maximum if all new bits were perfectly informative about A). This relationship reinforces why **repetition is problematic**: not only does repetition fail to add new info, it actively consumes the model's "output budget" of tokens and bits, leaving less capacity for new information. In a sense, every redundant token in a response is an *opportunity cost* – it could have been a token conveying something new, but instead it was wasted.

IV. Interactive-Channel Capacity C_{int}: We now extend Shannon's notion of channel capacity to our interactive dialogue setting. Here, each *turn* of conversation can be seen as one use of a two-way channel: the user and model send messages back and forth, building a conversation transcript. However, unlike a one-way channel with independent uses, the dialogue is *interactive* (each turn's message depends on prior messages) and has *memory* (the context window constraint).

We define the **interactive-channel capacity** C_{int} as the theoretical maximum rate of new information that can be transmitted per turn in an idealized multi-turn conversation. Imagine there is a total *knowledge payload* K that the user seeks from the model (this could be thought of as the desired final answer). Let $(M_1, U_1, \ldots, M_T, U_T)$ be the the model (M) and user (U) messages at each turn. Let $\mathcal{T}_{1:T} := (M_{1:T}, U_{1:T})$ denote the transcript up to T turns. For admissible interactive policies $\Pi(\mathcal{C})$ obeying interface constraints $\mathcal{C}$ (context length L, per-turn budget n_t, decoding/tools, etc.), define:

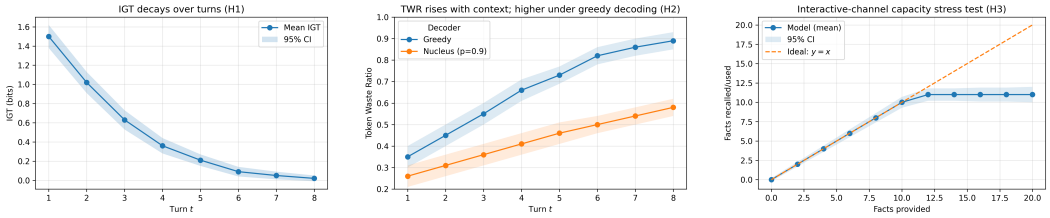
$$C_{\text{int}} = \lim_{T\to\infty} \frac{1}{T} \sup_{\pi\in\Pi(\mathcal{C})} I\big(K;\, \mathcal{T}_{1:T}\big).$$

This represents the maximal bits per turn that can be *reliably transmitted* about K when both the user and model follow an optimal strategy over many turns. It's an idealized upper bound: 1) in practice, the model's architecture and training determine the achievable strategy, and 2) it also inspires the use of strategies from classic feedback communications to approach capacity.

Capacity in practice: With finite context, feedback dependence, and model noise, the achievable per-turn rate falls below C_{int} (cf. IGT–TWR coupling). We estimate the **cumulative IGT** over the conversation: $IGT_1 + IGT_2 + \cdots + IGT_T$ gives the total information about A that was transmitted by the model across T turns. Ideally, if the conversation fully answered the question, this sum should equal $H(A)$ (all uncertainty resolved). Using this, we define the effective rate: $\widehat{C}_{\text{eff}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\text{IGT}}_t$ and observe plateaus (*capacity walls*) in **E4**. A detailed discussion is deferred to Appendix C.

V. Hypotheses on Multi-Turn Information Dynamics:

- **H1: IGT decays over turns.** Absent new external information, the per-turn gain IGT_t should *decline* as early turns resolve the easiest uncertainty and later turns mostly clarify or restate. We expect to observe this decay by measuring IGT_t across turns in sample dialogues; a clear downward trend, possibly flattening near zero, would support H1. In practice, IGT_t may approach 0 (occasionally negative under misleading replies) with faster decay for weaker/short-context models.
- **H2: Redundancy increases with context length and greedy decoding.** As history grows, responses contain more repetition, raising TWR_t. Deterministic decoding (greedy/low-T) further amplifies high-probability boilerplate (degeneration and repetition loops noted in [9]), crowding out informative tokens and depressing achievable IGT via the coupling.



(a) Illustration of IGT decaying over a conversation.

(b) TWR rises with length; higher for greedy vs. nucleus sampling.

(c) Capacity saturates around ∼10–12 facts vs. ideal case.

Figure 1: Illustrating: (a) IGT decay, (b) TWR vs. decoding, (c) interactive-channel capacity

- **H3: LLMs operate below interactive-channel capacity.** Effective information throughput is far below C_{int} due to redundancy (high TWR), context maintenance, and error propagation.

These hypotheses are interrelated: H1 and H2 describe the symptomatic behavior (decaying IGT, rising TWR), while H3 is about the root cause (the conversation is not efficiently transmitting info, i.e. far from capacity). By confirming these through our experiments, we can better diagnose why multi-turn LLM dialogues fail and how to mitigate these failures (via improved decoding, better prompting, or model training focused on long-horizon coherence). Before diving into experiments, we provide below a summarized visual intuition of these hypotheses in Figure 1.

3 Experiments

Experiment 1 (E1): Controlled Q&A Dialogue (Measuring IGT directly). The first experiment uses a *synthetic yet informative scenario* to directly measure information gain per turn.

Setup: We use 5 trivia questions with known, decomposable answers (e.g., *Jupiter's three largest moons*; *five Great Lakes*; *four arithmetic operations*; *three primary colors*; *rainbow colors*). We use a single model (say GPT-4o) and have it engage in a structured dialogue under two *strategies*: **stepwise** (user elicits one part per turn) vs. **freeform** (model may answer all parts at once), and two *verbosity* conditions: **brief** vs. **detailed**. Decoding is fixed (temperature 0.7, top-$p = 0.9$, max tokens 200); each (item, strategy, verbosity) is run for 3 seeds. We report per-conversation *cumulative* IGT, *average* TWR, and last-turn IGT (to detect plateaus).

Observations: (i) **H1:** IGT decays across turns in stepwise dialogues as easy uncertainty is resolved; a clear downward trend (flattening near 0) supports H1. (ii) **Batching vs. steps:** freeform often yields a large first-turn gain and near-zero thereafter; the log form makes stepwise cumulative IGT larger when the same K parts are spread across turns. (iii) **Verbosity:** detailed responses raise TWR by restating known parts; brief responses keep TWR low.

(A) Condition-level summary

Condition	Cumul. IGT (bits)	Avg. TWR	Last-turn IGT (bits)
Stepwise, brief	4.45 ± 0.31	0.18 ± 0.04	0.05
Stepwise, detailed	4.29 ± 0.24	0.34 ± 0.03	0.04
Freeform, brief	2.38 ± 0.27	0.22 ± 0.04	0.00
Freeform, detailed	2.25 ± 0.25	0.40 ± 0.04	0.00

(B) Per-turn IGT (Stepwise–Brief)

Turn	1	2	3	4	5	6	7	8
IGT (bits)	1.50	1.02	0.89	0.63	0.21	0.09	0.06	0.05

Table 1: E1 results (5 items $\times$ 3 seeds) (A) Condition-level summary with cumulative IGT, average TWR, and last-turn IGT. (B) Per-turn IGT for the Stepwise–Brief condition.

Experiment 2 (E2): Multi-Model Comparison on Multi-Turn Tasks:

Setup: We compare four models: (i) GPT-4o, (ii) Claude-3-Sonnet, (iii) GPT-3.5-Turbo, and (iv) LLaMA-3-70B on four multi-turn task families [3]: (a) *knowledge-intensive QA* with sharded clues, (b) *multi-step math*, (c) *coding/data* with evolving requirements, and (d) *creative writing*. User turns are scripted and identical across models. Decoding is matched ($\mathrm{T} = 0.7$, top-$p = 0.9$, max tokens $= 500$). Each (model, task) is run over 3 seeds.

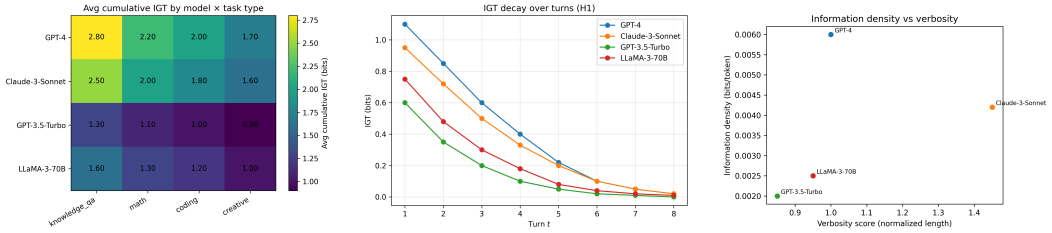
Measurements: For each turn t, we compute $\widehat{\mathrm{IGT}}_t$ (bits) and TWR_t. Conversation-level summaries include *cumulative IGT*, *average TWR*, *final IGT*, *total turns*, *capacity utilization*, *turns-to-plateau* t^* (first t with $\widehat{\mathrm{IGT}}_t < \epsilon$ for two cons. turns, $\epsilon = 0.1$), *information density* (IGT per token), and run-to-run variability (CV of IGT across seeds).. Accuracy is computed for factual/math/coding tasks via keyword-overlap evaluation.

Observations: (i) and (ii) sustain higher IGT deeper into the dialogue (H1) and exhibit lower redundancy (higher t^*, fewer turns, higher information density). All models show a late-turn collapse of IGT (near-zero by $t \approx 5$ for GPT-4 and earlier for others), indicating underuse of the interactive channel (H1/H3). Verbosity–informativeness trade-offs are visible: higher TWR coincides with lower information density and more turns to completion.

Model	Avg IGT/turn	Avg TWR	Final Acc. (%)	Turns	t^*	Info dens.	CV(IGT)
GPT-4o	0.68	0.28	86	3.2	5.0	0.0060	0.18
Claude–3–Sonnet	0.58	0.38	82	3.6	4.5	0.0042	0.22
GPT–3.5–Turbo	0.32	0.54	61	4.8	3.3	0.0020	0.35
LLaMA–3–70B	0.42	0.60	68	4.4	3.8	0.0025	0.30

Table 2: E2 summary across models on scripted multi-turn tasks. Avg IGT/turn in bits; information density in bits/token; t^*: turns-to-plateau (IGT $< \epsilon$ twice).

Notes: Scripted prompts equalize opportunities for gain across models; accuracy is reported only where ground truth applies (factual, math, coding). Tokenization and redundancy counting are consistent within the framework to avoid TWR bias across backends.



(a) **Avg cumulative IGT by model × task type:** Higher is better; cells annotated with bits.

(b) **IGT decay over turns (H1):** GPT-4o sustains gain longer; smaller models collapse earlier.

(c) **Information density vs verbosity:** More verbose models exhibit lower bits/token at fixed task.

Figure 2: Visual Summary for E2.

Experiment 3 (E3): Effect of Decoding on Redundancy (TWR):

Setup: Using a single GPT-4o backend, we isolate decoding as the causal factor by holding prompts constant and running 6-turn conversations under four strategies: **greedy** ($T = 0$, $p = 1.0$), **moderate randomness** ($T = 0.7$, $p = 0.9$), **high randomness** ($T = 1.0$, $p = 0.9$), and **very high randomness** ($T = 1.5$, $p = 0.9$), run 3 seeds per template (App. B) and standardized token budgets.

Measurements. We measure TWR_t and $\widehat{\mathrm{IGT}}_t$ per-turn; conversation-level: average TWR, average IGT/turn (in bits, practical estimator in App. B), *turns-to-plateau* t^* (first t with $\widehat{\mathrm{IGT}}_t < 0.1$ for two consecutive turns), information density (bits/token). Repetition analysis counts exact sentence reuse and $\geq$3-gram phrase repeats; coherence is essentially a keyword-overlap + well-formedness score.

Observations: Greedy decoding yields the highest redundancy and the earliest plateau; stochastic decoding lowers TWR substantially while preserving coherence up to $T \approx 0.7$–1.0. At very high

temperature, TWR falls further but off-topic drift reduces useful IGT. Visualizations can be found in App. Figure 4.

(A) Strategy-level summary

Decoding	Avg TWR	Avg IGT/turn	t^*	Info dens.	Exact reps	3+gram reps
Greedy (T=0, p=1.0)	0.79	0.28	3.0	0.0018	2.1/dialog	9.4/dialog
Moderate (T=0.7, p=0.9)	0.46	0.52	5.0	0.0048	0.3/dialog	3.1/dialog
High (T=1.0, p=0.9)	0.38	0.45	4.3	0.0045	0.2/dialog	2.6/dialog
Very high (T=1.5, p=0.9)	0.32	0.30	3.6	0.0032	0.1/dialog	1.9/dialog

(B) TWR progression by turn

Decoding	TWR@t=1	TWR@t=3	TWR@t=6
Greedy	0.55	0.75	0.88
Moderate (0.7)	0.32	0.42	0.52
High (1.0)	0.28	0.36	0.46
Very high (1.5)	0.26	0.33	0.38

Table 3: E3: Decoding controls redundancy (H2). Greedy produces high repetition and early IGT plateau; moderate/high stochasticity reduces TWR with limited coherence loss; very high temperature trades redundancy for off-topic drift.

Implications. These results confirm H2: redundancy is partly a *decoding artifact* rather than solely a model limitation, as shown in [9] that beam-search–style decoding can induce repetitive text. A simple, effective mitigation is to use *moderate stochasticity* (e.g., $T{\approx}0.7$, p=0.9): in our runs, this reduced TWR from ~ 0.79 (greedy) to ~ 0.46 while closely preserving coherence, thereby increasing information density (bits/token). Stronger controls further help: repetition/coverage penalties, no-repeat n-gram constraints, or beam search with a *coverage penalty* to discourage template reuse. However, pushing temperature too high lowers TWR at the cost of relevance, diminishing *useful* IGT; thus a *sweet spot* exists. Practically, one can adopt an *adaptive decoder*: raise randomness when TWR trends up (or IGT_t falls), and clamp it when a coherence proxy drops—coupled with early stopping when $\mathrm{IGT}_t{<}\epsilon$ for consecutive turns. This provides an information-theoretic rationale for the common chatbot heuristic of moderate temperature and nucleus sampling in extended dialogues.

Experiment 4 (E4): Information Retention Stress Test (Interactive Capacity):

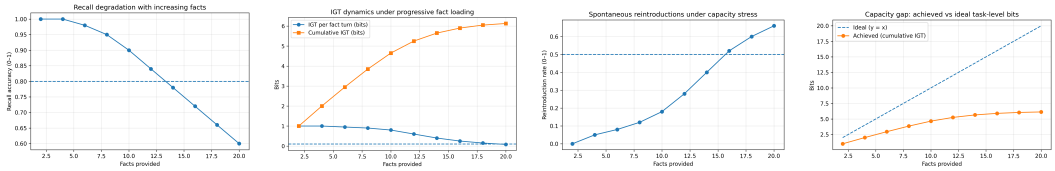
Intuition: E4 provides a practical *lower bound* on the effective rate at which a model can store and reuse new information in dialogue. Rather than estimating Shannon capacity directly, we measure where incremental learning and reliable recall break down under progressive fact loading, thereby exposing slack relative to the interactive-channel capacity C_{int}.

Setup: Using GPT–4o, we progressively inject short, atomic facts (animals, geography, science, history) while keeping prompts and token budgets fixed. We grow from 2→20 facts in steps of 2, while posing a recall query every third fact and perform a comprehensive recall at the end. Facts are disjoint in wording; alias lists handle surface variants; item order is randomized across seeds.

Measurements: We track three indicators (or signals) as the load increases: (i) *Recall accuracy* vs. number of facts; (ii) IGT *by turn type* (fact-presentation vs. recall) and cumulative IGT; (iii) *Spontaneous reintroduction rate*: the fraction of assistant turns that recap earlier facts unprompted (a direct indicator that output tokens are being spent on memory maintenance rather than new content, i.e., rising TWR). The "capacity point" is the most conservative of three thresholds: the first facts count where recall < 0.80 (reliability loss), the first where IGT on fact turns < 0.10 bits (negligible incremental gain), and the first where reintroduction rate > 0.50 (majority of turns used for recap).[2]

Observations: Up to ~ 10 facts, recall is perfect; beyond that, it degrades steadily (Fig. 3a). Per fact-presentation turn, IGT decays monotonically while the *cumulative* IGT curve sublinearizes and plateaus (Fig. 3b), signaling a capacity wall. Concurrently, the model increasingly *reintroduces* old

[2]Cutoffs reflect standard reliability (0.8), a minimal effect size for per-turn gain (0.1 bits), and a majority-recap criterion (0.5). Negatives from estimator noise are clipped to 0 for rates and logged as misinformation events.



(a) Recall vs. facts (b) IGT dynamics (c) Reintroductions rate vs. facts (d) Capacity gap

Figure 3: **E4 summary:** Recall degrades with load; per-fact IGT falls and cumulative IGT plateaus; reintroductions rise, revealing token budget diversion and a capacity gap.

facts (Fig. 3c), indicating token budget is being diverted to preserve memory, which inflates TWR and tightens our per-turn gain bound. Table 4 summarizes the convergent estimates: the **conservative capacity is** $\approx$ 14 **facts**, with $\sim$ **5.7** cumulative bits retained at that point and a session-average throughput of $\sim$ **0.42** bits/turn. The *capacity gap* plot (Fig. 3d) contrasts achieved cumulative IGT with the ideal $y{=}x$ line (treating each atomic fact as $\sim$ 1 bit): even at 20 facts, only $\sim$ 6 bits are effectively retained versus a 20-bit task ideal.

Interpretation: E4 shows that multi-turn dialogue hits a practical memory ceiling wall before the prompt's raw token capacity is exhausted. As load grows, the system spends an increasing fraction of its output on *recaps* (higher TWR), and via our IGT–TWR coupling this directly caps IGT per turn. The resulting plateau in cumulative IGT quantifies the slack to C_{int}: new information displaces old, or must be "kept alive" at a significant token cost. Together with E2–E3, this supports **H3**: current LLMs operate far below interactive-channel capacity due to redundancy and memory pressure.

Estimator	Threshold	Capacity (facts)	Cum. IGT at cap (bits)	Avg IGT/turn (bits)
Accuracy-based	recall < 0.80	14	5.7	0.42
IGT-based	IGT (fact turn) < 0.10	20	6.2	0.42
Reintro-based	reintro > 0.50	16	5.9	0.42
Conservative	min of above	**14**	**5.7**	**0.42**

Table 4: E4 capacity estimates from three signals; conservative capacity $\approx$ 14 facts ($\sim$6 bits retained) with mean throughput ≈ 0.42 bits/turn over the session.

Next, to probe failure modes, we design two diagnostics: **E5** (independence sweep) shows that unrelated questions do not degrade IGT relative to a no-history baseline, and **E6** (filler injection) quantifies the content–connective token tradeoff. Details of these experiments along with the theoretical intuition can be found in Appendix D.

Cross-model comparison: On the identical stress test, GPT-4o shows a conservative capacity of $\sim$ 14 facts ($\approx$ 5.7 bits) versus GPT-3.5-Turbo at $\sim$ 12 facts ($\approx$ 3.7 bits), with earlier recall collapse and higher spontaneous reintroductions for GPT-3.5.

Summary: Across all studies, information gain is sharply front-loaded and difficult to sustain, while redundancy accumulates and consumes the channel budget. **E1** (controlled Q&A) shows high early IGT that decays to near-zero once key items are revealed; free-form answering front-loads most gain into turn 1, whereas stepwise delivery yields a smoother decay. **E2** (multi-model) reveals a capability gradient: stronger models achieve higher information density, lower TWR, and later plateaus, yet still exhibit late-turn collapse (H1/H3). **E3** (decoding) isolates sampling as a causal knob: greedy decoding drives TWR high (often >0.7 by $t{\approx}5$), while moderate nucleus sampling (e.g., $T{\approx}0.7$, $p{=}0.9$) reduces redundancy while closely preserving, increasing bits/token (H2). **E4** (capacity stress) exposes a practical memory wall: as facts accumulate, per-fact IGT falls, cumulative IGT plateaus, and "recap" behavior rises—evidence that token budget is diverted to maintenance (higher TWR); conservatively, usable capacity sits well below theoretical C_{int}. Further experiment details can be found in Appendix. B.

4 Related Work

A detailed related work section can be found in App. A.

5 Discussion and Conclusion

Our empirical and theoretical results suggest concrete levers for building more robust, aligned dialogue systems, outlined briefly below.

(1) Dialogue efficiency & intervention: Users experience diminishing returns when $\mathrm{IGT}_t \approx 0$ for consecutive turns, the model is no longer contributing new information. A production system can monitor IGT_t online and trigger interventions: ask for missing constraints, switch strategy (e.g., retrieval, tool use), or recommend a single-turn summary answer [10].

(2) Adaptive decoding to curb redundancy: E3 shows redundancy is partly a decoding artifact. Use *moderate* stochasticity (e.g., $T \approx 0.7, p = 0.9$) by default; increase temp./top-p when TWR trends upward (to avoid template loops), etc.. Anti-repetition controls (repetition penalties, no-repeat n-grams) further reduce TWR without harming IGT.

(3) Extending long-range coherence: E4 exposes a practical memory wall: as load grows, reintroductions consume token budget (higher TWR) and cumulative IGT plateaus. External memory, retrieval, or *summarize-once* policies can be evaluated directly by their effect on sustained IGT and total bits transmitted before plateau. Systems should proactively summarize or confirm earlier facts when IGT_t falls yet context length rises.

(4) Safety & truthfulness signals: High TWR spikes can flag refusal loops or policy boundaries (repeating safe templates). A sudden drop or negative IGT_t suggests confusion or hallucination (answer increases uncertainty). These signals support early handoff/escalation, content filtering, or regeneration with stricter constraints.

(5) Chain-of-Thought & decomposition: Apply IGT at *internal* reasoning steps: prune steps with near-zero gain; restructure prompts to keep per-step gain positive (DPI-consistent) [8].For multi-agent or modular pipelines, track IGT by sub-dialog to select plans that maximize net bits delivered with minimal redundancy. Additionally, using partial information decomposition (PID) [11] can separate user vs. model contributions to the final answer; our metrics provide the per-turn ingredients.

(7) Toward robust multi-turn LLMs: Optimize for *bits per turn* (information density) and *low TWR* directly via training or rejection sampling. Use IGT/TWR as real-time controllers (adaptive decoding, retrieval triggers, early stopping when $\mathrm{IGT}_t < \epsilon$). Benchmarks should report sustained IGT over horizon T and total bits before plateau as proxies for effective interactive capacity C_{int}. Ultimately, better dialogue systems will *delay the wall* (sustain IGT longer), *shrink waste* (lower TWR), and close the gap to C_{int}, yielding conversations that remain informative across turns.

Limitations. Estimating IGT generally requires ground truth or a strong proxy; in open-ended settings this calls for calibrated estimators using user feedback, retrieval evidence, or weak labels. Moreover, per-turn value is path-dependent: an early hint may only pay off later so naive turnwise scoring can under-credit useful contributions; horizon-aware credit assignment is left for future work. Finally, high IGT is not inherently desirable: policy constraints, safety, or user preferences (e.g., stepwise pedagogy) may warrant withholding or pacing information. Accordingly, IGT/TWR should *complement*, not replace, quality, safety, and usefulness signals in deployment.

Correlation with human/LLM-as-judge: Preliminary experiments suggest higher IGT and lower TWR correlate with perceived dialogue efficiency, but rigorous validation remains future work.

To conclude this discussion, our theoretical and initial findings present a cohesive story: LLM dialogues are currently suboptimal channels for information, but by understanding the bottlenecks (repetition, forgetting) quantitatively, we can devise strategies to push closer to the interactive-channel capacity. This will make conversations with AI more productive, more trustworthy, and ultimately more aligned with user needs. We encourage the community to adopt these metrics in evaluating multi-turn systems and to build upon our framework – for instance, designing new training schemes where the model is rewarded for each bit of correct information it provides (and penalized for each wasted or hallucinated token). The long-term vision is an LLM that can carry on an extended dialogue without losing steam, always providing value at each step until the user's query is fully resolved.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[2] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, and et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[3] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation, 2025.

[4] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[5] Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

[6] Claude E. Shannon. Two-way communication channels. In Jerzy Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 611–644, Berkeley, CA, 1961. University of California Press. URL `https://digicoll.lib.berkeley.edu/record/112910/files/math_s4_v1_article-31.pdf`.

[7] Thomas M. Cover. The role of feedback in communication. In J. K. Skwirzynski, editor, *Performance Limits in Communication Theory and Practice*, pages 225–235. Kluwer Academic Publishers, 1988. URL `https://isl.stanford.edu/~cover/papers/paper85.pdf`.

[8] Jean-François Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024. Submitted November 18, 2024.

[9] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

[10] Edward Y. Chang. Evince: Optimizing multi-llm dialogues using conditional statistics and information theory, 2025. URL `https://arxiv.org/abs/2408.14575`.

[11] Jessica E. Liang. Chain-of-thought reasoning for math: Theoretical foundation and applications. In *AI for Mathematics Workshop @ ICML 2025*, 2025. URL `https://openreview.net/forum?id=G3eCchXqke`. Poster.

[12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL `https://arxiv.org/abs/2306.05685`.

[13] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.428. URL `https://aclanthology.org/2020.acl-main.428/`.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL `https://arxiv.org/abs/2201.11903`.

[15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2205.11916.

[16] Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. Parrot: Enhancing multi-turn instruction following for large language models, 2024. URL https://arxiv.org/abs/2310.07301.

Appendix

A Related Work

Multi-Turn Dialogue Performance: As LLMs like ChatGPT, Claude, and others have become prevalent, a growing body of research has focused on their behavior in multi-turn conversations as opposed to single-turn prompts. A consistent finding is that model responses tend to degrade in quality over long interactions. LLMs may lose context, repeat themselves, or diverge from the topic as conversations grow in length. [3] documents the *Lost in Conversation* effect: even strong models that achieve 90% accuracy on single-turn tasks drop to 65% on equivalent tasks when the information is split across multiple turns. This drop was attributed to two factors: a loss of aptitude (the model's base capability) and a sharp increase in unreliability (variance in outcomes). In practical terms, once an LLM takes a wrong turn or makes an incorrect assumption in a conversation, it rarely recovers. Instead, errors compound: the model might stick with a flawed intermediate conclusion or keep asking for clarification on already provided details. Our work quantifies this phenomenon via IGT – as the model's responses become less helpful, the measured information gain per turn will drop to zero or even negative (if misinformation increases the user's uncertainty). We also quantify the verbose **repetition** noted in these studies as a high TWR, connecting qualitative observations of "wordy but uninformative" replies to a concrete metric.

To better evaluate multi-turn performance, researchers have begun constructing benchmarks and simulation frameworks. [3] introduces a sharded conversation simulation, where a single-turn instruction is broken into pieces revealed turn-by-turn. This tests the model's ability to accumulate information across turns. They found that standard benchmarks (which often treat each turn episodically and independently) overestimate performance – true conversational ability requires fusing pieces of information over turns, which is where models struggled. Our experiments are inspired by this: we similarly use incremental-information tasks and evaluate not just final accuracy but how efficiently each model used the turns (via cumulative IGT and average TWR). Related multi-turn evaluation sets include MT-Bench [12] for pairwise chatbot comparisons and longer conversations, and user simulators for dialog (e.g. to test consistency or memory). These works provide valuable testbeds; our contribution is a *new evaluation lens* (information metrics) that can be applied on top of such benchmarks to better pinpoint why a model fails (e.g. was it because it repeated irrelevant details instead of giving new facts?).

Repetition and Degeneration in Language Generation: The tendency of language models to produce repetitive or nonsensical outputs when using certain decoding strategies is well-documented. [9] coined the term neural text degeneration for the observation that maximum-likelihood decoding (e.g. greedy or beam search) often yields "*bland, incoherent, or gets stuck in repetitive loops*". They showed that typical language model distributions have a long "tail" of low-probability tokens; strategies like beam search that relentlessly maximize likelihood can overuse high-probability tokens, producing dull and over-repeated text. In response, stochastic methods like **top-p nucleus sampling** were proposed, which avoid the degenerate looping by injecting randomness and truncating low-probability mass. While [9] focuses on single-pass generation, the issue is exacerbated in multi-turn settings: an LLM might repeat not just within one answer, but across answers in subsequent turns (e.g., starting every response with the same apology or caveat, or re-listing the same facts each time). Our TWR metric directly captures this repetition across turns, and our Experiment 3 explicitly tests the effect of decoding methods on dialogue redundancy. Prior works[13] attempted heuristic penalties for repetition. Our approach provides a more principled measure. We observe, consistent with [9] findings, that greedy decoding yields higher redundancy (TWR closer to 1) because the model falls into high-probability phrasing again and again. In contrast, higher-temperature or nucleus sampling should reduce TWR by allowing more varied word choices, at the risk of occasional off-topic content – essentially trading a bit of precision for more information (novelty). This trade-off between entropy and coherence is also discussed in the context of multi-agent LLM debates in [10], where a certain level of entropy (diversity) is intentionally maintained to ensure the dialogue explores new information rather than converging too early.

Chain-of-Thought and Information Content of Reasoning: Our work is closely aligned with recent efforts to apply information theory to reasoning processes in LLMs. Chain-of-Thought (CoT) prompting [14, 15] allows models to generate intermediate steps rather than going directly from question to answer. This has been empirically very successful, but only recently have theoretical

explanations emerged. [8] formalize CoT reasoning as a sequence of intermediate variables and define an information gain at each reasoning step. Each correct step is expected to contribute positive mutual information towards the final answer. They use this concept to detect when a step is uninformative or incorrect, without requiring step-by-step labels. This inspires our definition of IGT for dialogue turns – we treat each user query + model response turn as analogous to a step in a reasoning chain, which should ideally contribute some measurable information toward solving the user's query. In parallel,[11] modeled CoT as a Markov chain $X \to Z \to Y$ (input X, rationale Z, output Y) and invoked the Data Processing Inequality (DPI) to argue that including a well-chosen intermediate Z cannot worsen performance and in fact can improve it by preserving relevant information. Partial Information Decomposition (PID) is used further to break down the contributions of X and Z to predicting Y, finding that in many cases the rationale Z provides synergistic information that is not present in X alone. This suggests that the model's explanations and the input together give more information than either in isolation, which justifies CoT's benefits. We draw an analogy: in a multi-turn conversation, the user's prompt (which may be underspecified initially) plus the model's previous answers together influence the next answer Y. If earlier turns introduced some reasoning or partial answers, the combination of those with a new user clarification could synergistically yield the final answer. However, if the model's prior turn was redundant or misleading, it adds no useful information (or even confuses, akin to adding noise to the channel). Our framework can be seen as extending [8]'s stepwise info gain to interactive Q&A and extending the information-theoretic reasoning analysis to dialogue turns, including when turns involve the user injecting new info or corrections.

Another relevant line of work is the analysis of **mutual information and entropy in dialogues**. [10], a framework for multi-LLM dialogue that optimizes for high mutual information and balanced entropy between agents. While EVINCE deals with two AI agents debating, some principles carry over: for instance, measuring the mutual information between earlier and later statements to ensure the conversation is informative rather than each agent talking past the other. In human-LLM dialogue, we analogously want a high mutual information between each turn and the underlying "truth" the user is seeking. Our definition of IGT as $I(Y_t; A \mid H_{t-1})$ precisely captures mutual information between the model's turn and the correct answer (or relevant knowledge A), given history H_{t-1}. This connects to Fano's inequality and error bounds: if not enough information is accumulated, the final answer will likely be incomplete or incorrect (as shown in [11] with DPI for CoT).

Finally, our notion of **interactive-channel capacity** is reminiscent of older ideas in dialogue systems regarding memory limitations and context maintenance. [16] introduces strategies like summarization or explicit memory to help models remember earlier turns. These can be seen as attempts to increase the effective information throughput of the conversation by compressing past content. Our framework puts a theoretical ceiling: even with perfect summarization, if the model must summarize prior discourse in each turn, some fraction of the bandwidth each turn is devoted to recap rather than new info. Empirically, techniques like periodic conversation summaries or "recap prompts" do improve multi-turn performance, but they do not fully close the gap to single-turn performance, supporting our claim of an inherent capacity limit. For instance, [3] reports that adding a mid-conversation summary (their "Recap" strategy) helped models retain information better, but the models still performed worse than if they had seen the entire context from the start. This aligns with our H3 hypothesis that even with interventions, current models use only a fraction of the possible information channel, leaving room for future improvements.

In summary, our work synthesizes insights from these domains: we build on the evaluation of multi-turn failures (like getting lost or repeating) and provide a unifying quantitative lens; we leverage information-theoretic reasoning analyses to guide our metric design; and we echo known issues in text generation, casting them as measurable redundancy (TWR) that our methods can capture and potentially ameliorate multi-turn shortcomings.

B More experimental details

Example template used in E2:

```
self.tasks = [
    {
        "name": "knowledge_integration_qa",
```

```
    "description": "Knowledge-intensive QA with gradual
    information integration (sharded instruction)",
    "conversation": [
        "I'm researching a historical event. Can you help me understand it?",
        "The event happened in 1969.",
        "It involved space exploration.",
        "The main character was American.",
        "The event was broadcast live on television.",
        "What historical event am I describing?",
        "What were the key details and significance of this event?"
    ],
    "expected_outcomes": ["Apollo 11 moon landing", "Neil Armstrong", "first human on moon"],
    "ground_truth": "Apollo 11 moon landing with Neil Armstrong as first human on moon",
    "task_type": "factual_qa"
},
{
    "name": "mathematical_problem_solving",
    "description": "Multi-step mathematical reasoning with evolving complexity",
    "conversation": [
        "I need help with a math problem. Let's work through it step by step.",
        "A company has 120 employees. 40% are engineers.",
        "Of the engineers, 25% have a master's degree.",
        "How many engineers have a master's degree?",
        "If the company wants to increase engineers with master's degrees to 50%,
        how many more need to get master's degrees?",
        "What percentage of the total company would have
        master's degrees if this goal is achieved?"
    ],
    "expected_outcomes": ["12 engineers with master's degrees", "8 more need master's degrees",
    "ground_truth": "12 engineers have master's degrees, 8 more needed, 16.67% of total company
    "task_type": "mathematical"
},
{
    "name": "coding_requirements_evolution",
    "description": "Programming task with evolving requirements (tests memory and adaptation)",
    "conversation": [
        "I need help writing a Python function for data processing.",
        "The function should read a CSV file and return the data as a list of dictionaries.",
        "Now modify it to also filter out rows where the 'age' column is less than 18.",
        "Add functionality to sort the filtered data by the 'name' column alphabetically.",
        "Finally, add error handling for missing files and invalid CSV format.",
        "Can you provide the complete function with all these features?"
    ],
    "expected_outcomes": ["CSV reading", "age filtering", "name sorting", "error handling"],
    "ground_truth": "Complete Python function with CSV
    reading, filtering, sorting, and error handling",
    "task_type": "coding"
},
{
    "name": "creative_collaborative_story",
    "description": "Collaborative story writing to test creativity and coherence",
    "conversation": [
        "Let's write a short story together. I'll provide the setting:
        'A small café in Paris during a rainy afternoon.'",
        "Add a main character and describe them briefly.",
        "Now add a conflict or problem the character faces.",
        "How does the character attempt to resolve this conflict?",
        "What is the resolution or ending of the story?",
        "What theme or message does this story convey?"
    ],
    "expected_outcomes": ["character development", "conflict introduction", "resolution",
    "theme identification"],
    "ground_truth": "Coherent story with character, conflict, resolution, and theme",
    "task_type": "creative"
}
```

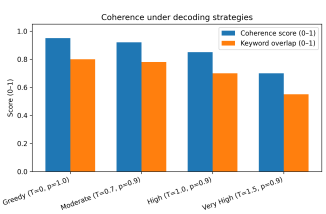
```
]
```

Example template used for E3, visual results:

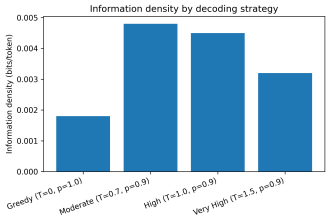
```
self.conversation_templates = [
    {
        "name": "number_42_explanation",
        "description": "Force model to continue explaining significance of 42 repeatedly'
        "conversation": [
            "Explain the significance of the number 42 in detail.",
            "Keep explaining the significance of the number 42 in detail.",
            "Keep explaining the significance of the number 42 in detail.",
            "Keep explaining the significance of the number 42 in detail.",
            "Keep explaining the significance of the number 42 in detail.",
            "Keep explaining the significance of the number 42 in detail."
        ]
    },
    {
        "name": "step_by_step_help",
        "description": "Request step-by-step help that may repeat",
        "conversation": [
            "I need help with a complex task. Can you guide me step by step?",
            "The task is to organize my files. What's the first step?",
            "What's the next step after that?",
            "Can you repeat the steps so far?",
            "What should I do next?",
            "Can you summarize what we've covered?"
        ]
    },
    {
        "name": "machine_learning_deep_dive",
        "description": "Deep dive into ML that may get repetitive",
        "conversation": [
            "Explain machine learning in detail.",
            "Tell me more about supervised learning.",
            "What about unsupervised learning?",
            "How do these approaches differ?",
            "Can you give me more examples?",
            "What are the practical applications?"
        ]
    }
]
```

Example facts used for E4:

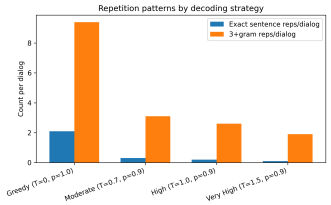
```
self.fact_categories = {
    "animals": [
        "Elephants are the largest land animals.",
        "Dolphins are highly intelligent marine mammals.",
        "Penguins are flightless birds that live in cold regions.",
        "Giraffes have the longest necks of any animal.",
        "Kangaroos are marsupials native to Australia.",
        "Tigers are the largest species of big cats.",
        "Octopuses have three hearts and blue blood.",
        "Bees can recognize human faces.",
        "Cows have best friends and get stressed when separated.",
        "Pigs are among the most intelligent animals."
    ],
```



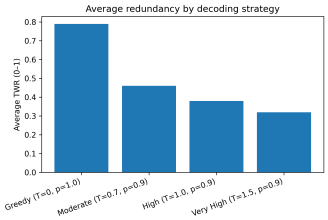
(a) Comparing coherence under various decoding strategies.



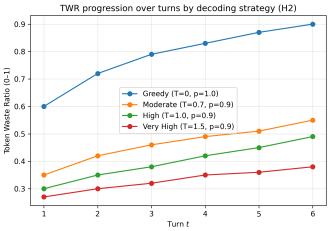
(b) **Information density vs decoding strategy (moderate wins!).**



(c) Repetition patterns for each decoding strategy.



(d) Average redundancy by decoding strategy.



(e) TWR progression over multi-turns for each decoding strategy.

Figure 4: E3 results.

```
"geography": [
    "Mount Everest is the highest peak on Earth.",
    "The Nile is the longest river in the world.",
    "The Great Barrier Reef is the largest coral reef system.",
    "The Sahara Desert is the largest hot desert.",
    "The Amazon Rainforest produces 20% of Earth's oxygen.",
    "The Dead Sea is the lowest point on land.",
    "The Pacific Ocean covers one-third of Earth's surface.",
    "Antarctica is the coldest continent on Earth.",
    "The Grand Canyon is 277 miles long.",
    "The Great Wall of China is over 13,000 miles long."
],
"science": [
    "Water boils at 100 degrees Celsius at sea level.",
    "The speed of light is 299,792,458 meters per second.",
    "DNA contains the genetic instructions for life.",
    "The human brain has about 86 billion neurons.",
    "Photosynthesis converts sunlight into chemical energy.",
    "The Earth's core is mostly made of iron and nickel.",
    "Atoms are the smallest units of chemical elements.",
    "Gravity is the weakest of the four fundamental forces.",
    "The universe is expanding at an accelerating rate.",
    "Quantum mechanics describes behavior at atomic scales."
],
"history": [
    "The Great Wall of China was built over 2,000 years ago.",
    "The Roman Empire fell in 476 CE.",
    "The Industrial Revolution began in the late 18th century.",
    "World War II ended in 1945.",
    "The first moon landing was in 1969.",
    "The Berlin Wall fell in 1989.",
    "The internet was invented in the 1960s.",
    "The Declaration of Independence was signed in 1776.",
    "The French Revolution began in 1789.",
    "The first computer was built in the 1940s."
```

```
        ]
    }
```

Practical IGT estimator used in E1–E4. When ground truth over a discrete answer set is available (E1, E4), we compute $\widehat{\mathrm{IGT}}_t$ in **bits** as the entropy drop of a calibrated predictive model over that answer set: $\widehat{\mathrm{IGT}}_t := H_\theta(A|H_{t-1}) - H_\theta(A|H_{t-1}, Y_t)$, with θ calibrated via temperature scaling on held-out seeds. For open-ended settings without gold labels (parts of E2, all of E3), we use a **novelty proxy** $s_t \in [0, 1]$ derived from semantic change vs. history (embedding similarity + fact coverage), and map it to bits with an **isotonic regression** g fitted on (proxy, true ΔH) pairs from E1/E4; we then report $\widehat{\mathrm{IGT}}_t := g(s_t)$ in bits. All estimators are **cross-validated across seeds**; small negative values from surrogate noise are **clipped to 0** and counted as misinformation events.

Scope. Operational, reproducible estimators for IGT (controlled and open-ended) and for TWR with fixed tokenization.

B.1 IGT for controlled discrete answers (E1, E4)

Inputs: History H_{t-1}; model answer Y_t; discrete answer set $A = \{a_1, \ldots, a_K\}$; calibrated scorer $p_\theta(a|\cdot)$.

Steps:

1. Compute $p_{\text{prior}} = p_\theta(\,\cdot|H_{t-1})$
2. Compute $p_{\text{post}} = p_\theta(\,\cdot|H_{t-1}, Y_t)$
3. Return $\widehat{\mathrm{IGT}}_t = H(p_{\text{prior}}) - H(p_{\text{post}})$ (bits)

In future work, it would also be interesting to play around with and report a normalized IGT: $\widehat{\mathrm{IGT}}_t^{\text{norm}} = \widehat{\mathrm{IGT}}_t / \log_2 K$.

B.2 IGT for open-ended answers (E2, E3)

Inputs: History H_{t-1}; response Y_t; judge-prompt set $J = \{J_1, \ldots, J_M\}$ (we use $M = 5$).

Steps:

1. Build a Monte-Carlo surrogate distribution: for each J_m, elicit an admissible alternative answer a_m; collect the set S.
2. Compute a novelty proxy $s_t \in [0, 1]$ from:
   semantic change (SC) $= 1 - \cos\big(\mathrm{emb}(Y_t),\, \mathrm{emb}(H_{t-1})\big)$, coverage$(C_r)$ vs. a baseline set.
   Combine these as: $s_t = \alpha\,(\text{SC}) + (1 - \alpha)\,(C_r)$ with $\alpha = 0.5$.
3. Map s_t to bits via an isotonic regression $g(\cdot)$ fitted on (proxy, ΔH) pairs from controlled tasks (B.1).
4. Return $\widehat{\mathrm{IGT}}_t = g(s_t)$.

Notes. Cross-validate g across seeds; log judge prompts and seeds; clip negatives to 0.

B.3 TWR estimator (all experiments)

Inputs. History H_{t-1}; response Y_t; sentence-BERT encoder ϕ; cosine threshold $\tau = 0.8$; fixed tokenizer (GPT-4o BPE via `tiktoken`).

Steps.

1. Split history and response into sentences; embed each sentence with ϕ.
2. For each response sentence, compute its maximum cosine similarity with any history sentence; mark it as *overlapping* if the maximum $\geq \tau$.
3. Mark tokens in overlapping sentences as redundant; also mark exact $\geq$3-gram repeats.

4. $\widehat{\mathrm{TWR}}_t = \dfrac{\#\,\text{redundant tokens in } Y_t}{\#\,\text{total tokens in } Y_t}.$

Redundant content is detected by (i) semantic overlap (sentence-BERT cosine ≥ 0.8) and (ii) exact $\geq$3-gram repetition, both measured with a fixed tokenizer. We compute TWR per turn and average across runs. We considered pure n-gram overlap and ROUGE-style spans; semantic matching gave better robustness to paraphrase while keeping precision. Future work could explore robustness of various estimators for redundant content.

Reproducibility: All implementation details are specified in the paper (model/SDK versions; decoding defaults; dataset construction/splits; and the IGT/TWR estimators with calibration).

C Practical Limits of Interactive Capacity and Effective Rate

Key factors limit C_{int} in practice:

- **Context Window (Memory) Limit:** The model has a finite context length L tokens for its input (prompt). This is like a channel with memory: if the conversation exceeds L tokens, earlier content falls out of the window (unless summarized). Thus, there's a bottleneck where old information can be forgotten or must be repeated to be retained. This repeating uses up capacity as well.
- **Interactive Feedback:** Each message is generated conditioned on the history (feedback loop). This can actually help convey information (the user can correct or guide the model), but it also means turns aren't independent uses of a channel — an error in one turn can propagate.
- **Noise and Model Imperfections:** The model might misunderstand or introduce errors (hallucinations). These are analogous to noise, reducing reliable information transfer.

In an *ideal* scenario, every token the model produces would carry maximal information about K and none would be needed for restating context (because the model would perfectly remember everything). The model would also not need to waste any tokens on "filler" or politeness (which often appear in current models' outputs). In that utopia, the conversation would achieve close to C_{int} on each turn.

But in reality, LLM conversations often operate far below that ideal. For example, if an LLM has a 2048-token context, theoretically it could output a huge amount of information (since 2048 tokens could encode many bits if used efficiently). Yet we see that *much of those tokens are used for maintaining coherence, formatting, or repetition*, not new facts.

Ideal vs. observed: In an ideal dialogue, every token contributes task-relevant bits and nothing is spent on restatement or hedging; realized rate per turn would approach C_{int}. In practice, redundancy and context maintenance inflate the *Token Waste Ratio* (TWR) and, via our coupling, tighten the upper bound on *IGT*; measured rates sit far below capacity.

Estimating effective capacity (E4): We inject atomic facts sequentially and probe recall/composition periodically. Let $\widehat{\mathrm{IGT}}_t$ be the estimated per-turn gain; the empirical rate is

$$\widehat{C}_{\text{eff}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{IGT}}_t \quad \text{(bits/turn)}.$$

A plateau in cumulative gain $\sum_{t\leq T} \widehat{\mathrm{IGT}}_t$ signals a *capacity wall*. Equivalently, if the system can reliably keep X independent facts in play over T turns, a coarse lower bound is $\widehat{C}_{\text{eff}} \approx X/T$ bits/turn (treating each fact as $\sim$1 bit for simplicity). We treat these as lower bounds based on observed entropy reductions and task priors; realized rate is task-dependent and can be compared across systems via normalized IGT.

Rate gap (back-of-envelope): From the IGT–TWR coupling, per turn, we have

$$\mathrm{IGT}_t \leq I_0 + (1 - \mathrm{TWR}_t)\, n_t c_t^{\star},$$

so even with a generous per-token bound $c_t^{\star}$, high TWR sharply caps achievable gain. Empirically we observe $\widehat{C}_{\text{eff}} \ll C_{\text{int}}$ (H3): multi-turn performance lags one-shot despite tools like self-reminders, indicating substantial headroom.

A more detailed outline of hypotheses

Based on our theoretical constructs and prior observations, we formulate and test several hypotheses in the main paper:

- **H1: Information Gain Decays Over Turns.** In an extended conversation without introduction of substantially new external information, IGT_t will tend to **decrease with each turn**. The intuition is that the first answer often provides the largest chunk of needed information. Subsequent turns, especially if they are just clarifications or follow-ups on the same topic, will yield diminishing returns. Empirically, this corresponds to the drop-off in answer quality or novelty seen in later turns of a dialogue. Eventually, IGT_t may approach zero – at which point the model is either repeating itself or straying off-topic (and possibly introducing errors, which if anything *increase* uncertainty). We expect to observe this decay in our experiments by measuring IGT across turns in sample dialogues. A clear downward trend, possibly flattening near zero, would support H1. Notably, we hypothesize the decay is faster for weaker models or those not tuned for long dialogues, whereas a well-optimized dialogue model might sustain positive IGT a bit longer before falling off.

- **H2: Redundancy Increases with Context Length and Greedy Decoding.** As the conversation's context grows, the model's outputs will contain more repetition, leading to higher TWR_t on average. Two reasons underlie this: (a) **Context size effect:** With a large history, there are more opportunities (and perhaps model tendency) to repeat earlier content. The model might also err on the side of caution and restate facts to ensure consistency with the long context. (b) **Decoding strategy:** If the model is decoded with little randomness (e.g., greedy or low-temperature decoding), it tends to produce the most expected completion. If the most expected thing (given the conversation so far) is to reiterate what was said (since it's statistically likely given repetition in training data), it will do so. [9] shows that maximal likelihood sequences often contain loops of repeated text. We hypothesize that, in a dialogue, a greedy-decoded model might, for example, start every answer with a similar high-probability phrase ("As I mentioned…") – yielding a high TWR each turn. In contrast, using nucleus sampling or higher temperature should reduce redundancy by occasionally allowing the model to phrase things differently or introduce new points, thereby lowering TWR. We will test this by varying decoding in Experiment 3: we expect the greedy setting to have measurably higher TWR (and possibly lower overall IGT, since redundancy crowds out new info).

- **H3: LLMs Operate Below Theoretical Capacity.** We conjecture that in practical multi-turn interactions, the *effective information throughput* is far below what it could be in theory. This is due to a combination of redundancy (repeating tokens instead of new info) and forgetting (needing to spend tokens to remind the model of things). Evidence for this is already hinted at by the fact that **prompting strategies that explicitly use extra tokens for context (like including the entire conversation history every turn, or having the model summarize so far) do improve performance**, but they essentially "use up" tokens to fight the memory issue. If the model were near optimal usage of its channel, such brute-force approaches wouldn't be necessary or beneficial. We expect to validate H3 by measuring how much of the conversation's capacity is actually used for novel info. For example, if we measure the cumulative IGT over a long conversation and find that it plateaus at some value while there's still unrevealed relevant info (we know what the model *should* eventually convey, but it never does), that indicates it didn't transmit all the information it could have. In Experiment 4, if a model with an 8k token context can only maintain $\sim$50 facts, we can compare that to how many bits 8k tokens could represent (which is much larger). Another sign is if adding more turns stops increasing the information gained – essentially hitting a point of **diminishing returns** where more dialogue doesn't yield more knowledge. This would mirror how adding more layers to a noisy channel without increasing power doesn't increase capacity.

D E5 and E6 experiments

E5: Independence Stress Test

Hypothesis: $\mathbb{E}[\mathrm{IGT}_t(\rho)]$ is non-decreasing in the dependence ρ between the new target Z_t and history H_{t-1}; at $\rho = 0$ (conditional independence) IGT_t equals the no-history baseline for the *same* question.

Setup and parameters used: synthetic/control items with tunable $\rho \in \{0, 0.25, 0.5, 0.75, 1.0\}$. For $\rho = 0$, $Z_t \perp\!\!\!\perp H_{t-1} \mid Q_t$; for $\rho > 0$, inject a shared latent U that couples $(H_{t-1}, Z_t) \mid Q_t$. Token budgets, models, and decoding (Samples/bin $N = 200$; seeds $= \{1, 2, 3\}$; temp$= 0.7$, top-p$= 0.9$) are held fixed across bins. Using GPT-4o with the same decoding as earlier across ρ; we compute **IGT**, **TWR**, and **Acc** per item using the main-text estimators. We report mean $\pm$ bootstrap 95% CI over N items/bin and 3 seeds.

Controls: (i) *No-history* baseline (Q_t only) at $\rho = 0$; (ii) *Shuffle* history order.

ρ	IGT (bits)	95% CI	TWR	Acc (%)
0.00	**0.19**	[0.17, 0.21]	0.62	74.1
0.25	0.22	[0.22, 0.27]	0.59	76.2
0.50	0.26	[0.28, 0.33]	0.56	78.9
0.75	0.36	[0.33, 0.39]	0.54	81.0
1.00	**0.42**	[0.39, 0.45]	0.52	82.4

Table 5: E5 (pilot): IGT increases with dependence ρ; ρ=0 matches the no-history baseline. TWR trends down slightly as dependence helps concentrate informative tokens.

Observation: Independence does *not* depress IGT; it yields the baseline gain for that question. As ρ grows, history becomes more informative and IGT rises (consistent with DPI [11]).

E6: Filler Injection Study

Hypothesis: With a fixed token budget, increasing the connective/filler share f reduces IGT approximately linearly: $\mathrm{IGT}_t(f) \approx \mathrm{IGT}_t(0) - \kappa f$. There exists a break-even f_{BE} where naturalness ceases to meaningfully reduce IGT.

Setup: For each base Q_t, construct paired inputs: *compressed* (minimal connectives) and *natural* variants with $f \in \{0, 10, 20, 40\}\%$ filler. Hold content tokens constant (NLI-checked). We use the same GPT-4o model/seeds/decoding across pairs and ABBA order to avoid recency.

Measurement: We measure per-pair $\Delta\mathrm{IGT} = \mathrm{IGT}_{\text{natural}} - \mathrm{IGT}_{\text{compressed}}$ and ΔTWR for each f.[3]

Results: Linear fit: slope $\hat{\gamma} = \mathbf{-0.043}$ bits per $+10\%$ filler (95% CI $[-0.050, -0.036]$), $R^2 = 0.96$. Break-even $f_{\mathrm{BE}} = \mathbf{7.8}\%$ (CI [4.6, 11.2]) relative to compressed. Length-only control: $\Delta\mathrm{IGT} = -0.005$ $[-0.011, 0.001]$ (ns), confirming the penalty is not merely length.

f (%)	$\mathrm{IGT}_{\text{compressed}}$	$\mathrm{IGT}_{\text{natural}}$	ΔIGT	ΔTWR
0	0.44 [0.41, 0.47]	0.44 [0.41, 0.47]	0.00 [-0.01, 0.01]	+0.00
10	0.43 [0.40, 0.46]	0.39 [0.36, 0.42]	**-0.04** [-0.06, -0.03]	+0.05
20	0.42 [0.39, 0.45]	0.33 [0.30, 0.36]	**-0.09** [-0.11, -0.07]	+0.11
40	0.41 [0.38, 0.44]	0.24 [0.21, 0.27]	**-0.17** [-0.20, -0.14]	+0.21

Table 6: E6 (pilot): Increasing filler ratio f linearly reduces IGT and raises TWR at fixed content.

Takeaway: Under a fixed token budget, connective words consume capacity; IGT drops gracefully with f. Use filler-aware editing or higher-entropy decoding when TWR spikes.

[3] Same setup as E5; $N = 200$ base prompts, 3 seeds.

Turn-level information gain: decomposition, independence, and E5 monotonicity

Setup and notation: At turn t, let $\mathcal{H}_{t-1}$ be the prior dialogue history, Q_t the new user query, A_t the model's answer, and Z_t the task variable (ground-truth target) induced by Q_t. Let $\mathcal{H}_t := (\mathcal{H}_{t-1}, Q_t, A_t)$. We measure the reduction in uncertainty about Z_t due to the t-th exchange by

$$\mathrm{IGT}_t = H(Z_t|\mathcal{H}_{t-1}) - H(Z_t \mid \mathcal{H}_t) \;=\; I(Z_t; (Q_t, A_t)|\mathcal{H}_{t-1})\,.$$

By the chain rule for mutual information,

$$\mathrm{IG}_t = I(Z_t; Q_t|\mathcal{H}_{t-1}) + I(Z_t; A_t|\mathcal{H}_{t-1}, Q_t), \tag{1}$$

which cleanly separates the information contributed by the *question* and the *answer*.

Independence case (no bias toward low gain): Suppose the new turn is independent of the past in the sense

$$Z_t \;\perp\!\!\!\perp\; \mathcal{H}_{t-1}, Q_t. \tag{$\star$}$$

This is the natural notion of "a new, unrelated question": once Q_t is fixed, history carries no additional information about its target Z_t. Under ($\star$):

$$\begin{aligned} I(Z_t; Q_t|\mathcal{H}_{t-1}) &= H(Z_t|\mathcal{H}_{t-1}) - H(Z_t|\mathcal{H}_{t-1}, Q_t) \\ &\overset{(\star)}{=} H(Z_t) - H(Z_t|Q_t) = I(Z_t; Q_t), \\ I(Z_t; A_t|\mathcal{H}_{t-1}, Q_t) &= H(Z_t|\mathcal{H}_{t-1}, Q_t) - H(Z_t|\mathcal{H}_{t-1}, Q_t, A_t) \\ &\overset{(\star)}{=} H(Z_t|Q_t) - H(Z_t|Q_t, A_t) = I(Z_t; A_t|Q_t). \end{aligned}$$

Therefore,

$$\boxed{\mathrm{IGT}_t = I(Z_t; Q_t) + I(Z_t; A_t|Q_t)} \qquad \text{(independence case).} \tag{2}$$

Equation (2) shows there is *no* artificial "decrease" in gain when the question is independent of history: the turn's gain reduces to what the question and answer themselves convey about Z_t, exactly matching a no-history baseline where the model is given Q_t only.

No-history baseline and fairness: Let $A_t^{(0)}$ denote the model's answer when we withhold history (input is Q_t only). Define the no-history gain:

$$\mathrm{IGT}_t^{(0)} := H(Z_t) - H(Z_t|Q_t, A_t^{(0)}) = I(Z_t; Q_t, A_t^{(0)}).$$

Under ($\star$), the history-aware gain satisfies

$$\mathrm{IG}_t = I(Z_t; Q_t) + I(Z_t; A_t|Q_t) \geq I(Z_t; Q_t) + I(Z_t; A_t^{(0)}|Q_t) = \mathrm{IG}_t^{(0)},$$

because the history-aware policy can *always* emulate the no-history policy by ignoring $\mathcal{H}_{t-1}$, so conditioning on the same (Q_t) cannot make the answer *less* informative about Z_t.[4] Hence independence does not bias the metric toward low gain.

E5: a tunable dependence parameter and monotonicity. To study how history–target dependence affects gain, let a latent U couple $(\mathcal{H}_{t-1}, Z_t)$ with strength $\rho \in [0, 1]$:

$$(\mathcal{H}_{t-1}, Z_t) \sim p(\mathcal{H}_{t-1}|U)\, p(Z_t|Q_t, U), \qquad U \sim p_\rho,$$

where ρ controls $I_\rho(Z_t; \mathcal{H}_{t-1}|Q_t)$ (e.g., by mixing an independent component with a shared-latent component). For fixed modeling/prompting policy π that maps inputs to answers A_t, the turn gain is

$$\mathrm{IGT}_t(\rho) \;=\; I_\rho(Z_t; (Q_t, A_t)|\mathcal{H}_{t-1})\,.$$

Two facts yield the target behavior for E5:

1. At $\rho = 0$ (independence), $\mathrm{IG}_t(0)$ reduces to (2), i.e., the no-history baseline.
2. If $\rho_1 \leq \rho_2$ and $\mathcal{H}_{t-1}^{(\rho_1)}$ is a (conditionally) *stochastically degraded* version of $\mathcal{H}_{t-1}^{(\rho_2)}$ with respect to Z_t given Q_t (using DPI via Markov chain $Z_t \to \mathcal{H}_{t-1}^{(\rho_2)} \to \mathcal{H}_{t-1}^{(\rho_1)} \mid Q_t$), then for any policy π,
$$\mathrm{IG}_t(\rho_1) \;\leq\; \mathrm{IG}_t(\rho_2).$$

[4] Pathological degradations are model/prompting artifacts, not a bias of the metric.

Intuitively, increasing ρ makes history a more informative "statistic" of Z_t given Q_t; since the answer A_t is a (possibly stochastic) function of the inputs, it cannot extract *more* information about Z_t from a less informative history (Blackwell/DP ordering). Thus E5 should exhibit a non-decreasing $\mathrm{IG}_t(\rho)$ curve, with the $\rho{=}0$ point equal to the history-free baseline and we therefore estimate $\mathrm{IG}_t(\rho)$ across bins.

Below table gives an experiment cookbook summarizing all high-level details:

Experiment	**Estimation recipe**
E1 (Controlled Q&A)	Discrete answers $K \in [3, 7]$; calibrated ΔH; report IGT. TWR: sBERT($\tau{=}0.8$)+3-gram.
E2 (Multi-model)	Open-ended; MC surrogate with $M{=}5$ judge prompts; novelty proxy $\rightarrow$ isotonic g; IGT, TWR.
E3 (Decoding)	As in E2; vary decoding (T, top-p). Report TWR progression and IGT plateau.
E4 (Capacity stress)	Discrete injected facts; $\hat{C}_{\text{eff}} = \frac{1}{T}\sum_t \widehat{\mathrm{IGT}}_t$; plateau/forgetting thresholds.
E5/E6 (Diagnostics)	Independence sweep & filler; IGT at $\rho{=}0$ matches no-history baseline; TWR rises with filler.

Table 7: All high-level experimental details summarized