A Short Note on Structural Memory, Turnwise Accounting, and Next-Token KL

J. Landers

Draft note – turnwise memory version

Context. The paper *Provable Long-Range Benefits of Next-Token Prediction*, by Xinyuan Cao and Santosh S. Vempala, studies a language model trained by next-token loss. In its formalism, a prefix-only autoregressive model assigns

$$q(x_i \mid x_{:i}), \qquad \bar{q}(x) = \prod_{i=1}^{n} q(x_i \mid x_{:i}),$$

and minimizing next-token log loss is equivalent, up to the entropy of the true distribution, to minimizing

$$\mathrm{D}_{\mathrm{KL}}(\bar{p} \| \bar{q}).$$

The paper's main conceptual move is that if a bounded next-k-token distinguisher can tell model continuations from true continuations, then that distinguisher can be used to improve next-token loss. Thus a sufficiently good loss minimizer becomes indistinguishable from truth to such bounded tests.

The question here is slightly different. Suppose there is another method, not merely prefix-only, that carries an additional structural state. Call that state C_i. It could be a plan, proof state, document-level intent, retrieved memory, conversation summary, scratch state, or any other persistent object. The issue is not just whether this method has lower KL. The structures are fundamentally different, so we want a comparison that says what the extra structure contributes.

Two conditional models. Let

$$S_i = x_{:i}, \qquad Y_i = x_i,$$

where S_i is the visible prefix and Y_i is the next token. The paper's kind of model is

$$q_i(y) = q(y \mid S_i).$$

The structured method is

$$r_i(y) = r(y \mid S_i, C_i).$$

Here C_i is the extra state. If C_i is memory, the point is not that the model is magically smarter. It is simply conditioning on a finer sigma-algebra: it sees (S_i, C_i) instead of only S_i.

There is an immediate relationship between the two. If q is what remains after erasing the structure from r, then

$$q(y \mid S_i) = \mathbb{E}\big[r(y \mid S_i, C_i) \mid S_i\big] = \sum_c r(y \mid S_i, c) \Pr(c \mid S_i).$$

So q is a coarsening, or marginalization, of r. The reverse direction is not determined: many different structured models collapse to the same prefix-only model. Equivalently, whenever the supports agree,

$$r(y \mid S_i, C_i) = q(y \mid S_i) W_i(y, C_i), \qquad \sum_y q(y \mid S_i) W_i(y, C_i) = 1,$$

where W_i is a structure-dependent likelihood ratio. Thus the structured model is a tilt of the prefix-only model toward tokens favored by the additional state.

Truth-relative comparison. Let the true conditional laws be

$$p_i^S(y) = p(y \mid S_i), \qquad p_i^C(y) = p(y \mid S_i, C_i).$$

Define the expected per-token gain of the structured method over the prefix-only method by

$$G_i := \mathbb{E}_p \left[\log \frac{r(Y_i \mid S_i, C_i)}{q(Y_i \mid S_i)} \right].$$

At the sequence level, this is the expected log-likelihood ratio

$$\mathbb{E}_p \left[\log \frac{\prod_i r(Y_i \mid S_i, C_i)}{\prod_i q(Y_i \mid S_i)} \right] = \sum_{i=1}^{n} G_i.$$

Equivalently, it is the KL/log-loss improvement of the structured method over the prefix-only method, evaluated under truth.

Theorem 1 (Structural gain decomposition)**.** *Assume all displayed KL divergences are finite. For each token position i,*

$$G_i = \mathrm{I}_p(Y_i; C_i \mid S_i) + \mathbb{E}_{S_i} \mathrm{D}_{\mathrm{KL}}\big(p_i^S \| q_i\big) - \mathbb{E}_{S_i, C_i} \mathrm{D}_{\mathrm{KL}}\big(p_i^C \| r_i\big).$$

Consequently, if

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{S_i, C_i} \mathrm{D}_{\mathrm{KL}}(p_i^C \| r_i) \le \delta_r, \qquad \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{S_i} \mathrm{D}_{\mathrm{KL}}(p_i^S \| q_i) \le \delta_q,$$

then, with

$$\bar{G} = \frac{1}{n} \sum_i G_i, \qquad \bar{I} = \frac{1}{n} \sum_i \mathrm{I}_p(Y_i; C_i \mid S_i),$$

one has

$$\boxed{\bar{I} - \delta_r \le \bar{G} \le \bar{I} + \delta_q.}$$

In particular, if both methods are Bayes-optimal relative to the information they see,

$$q_i = p(\cdot \mid S_i), \qquad r_i = p(\cdot \mid S_i, C_i),$$

then

$$\boxed{\bar{G} = \bar{I}.}$$

Proof. Start from

$$G_i = \mathbb{E}_p \log r(Y_i \mid S_i, C_i) - \mathbb{E}_p \log q(Y_i \mid S_i).$$

For the structured model,

$$\mathbb{E}_p \log r(Y_i \mid S_i, C_i) = -H_p(Y_i \mid S_i, C_i) - \mathbb{E}_{S_i, C_i} \mathrm{D}_{\mathrm{KL}}(p_i^C \| r_i).$$

For the prefix-only model,

$$\mathbb{E}_p \log q(Y_i \mid S_i) = -H_p(Y_i \mid S_i) - \mathbb{E}_{S_i} \mathrm{D}_{\mathrm{KL}}(p_i^S \| q_i).$$

Subtracting gives

$$G_i = H_p(Y_i \mid S_i) - H_p(Y_i \mid S_i, C_i) + \mathbb{E}_{S_i} \mathrm{D}_{\mathrm{KL}}(p_i^S \| q_i) - \mathbb{E}_{S_i, C_i} \mathrm{D}_{\mathrm{KL}}(p_i^C \| r_i).$$

The entropy difference is $\mathrm{I}_p(Y_i; C_i \mid S_i)$. Averaging over i and using nonnegativity of KL gives the bounds. The Bayes-optimal case sets both KL error terms to zero. □

A turn-level memory invariant. The previous theorem works token by token. For chat-like use, it is more natural to group the text into prompt turns. Let $t = 1, \dots, T$ index turns, let S_t be the visible text prefix at turn t, and let Y_t be the full response at that turn. Let $Z_1, \dots, Z_{m_t}$ denote available previous chunks of text. A memory mechanism can be modeled by similarity scores

$$a_{tj} = \mathrm{sim}(S_t, Z_j), \qquad \pi_{tj} = \frac{\exp(\beta a_{tj})}{\sum_{\ell=1}^{m_t} \exp(\beta a_{t\ell})},$$

and a retrieved state

$$C_t = \mathcal{R}(S_t) = \{(Z_j, \pi_{tj})\}_{j \le m_t}.$$

The scalar β controls how sharply memory locks onto similar prior text. A useful relevance score is

$$\rho_t = 1 - \frac{H(\pi_t)}{\log m_t} \in [0, 1],$$

which is close to 1 when retrieval concentrates on a small number of strongly similar chunks, and close to 0 when memory is diffuse.

Define the local distributional perturbation caused by memory at turn t as

$$\Delta_t(S_t, C_t) := \mathrm{D}_{\mathrm{KL}}\big(r_t(\cdot \mid S_t, C_t) \| q_t(\cdot \mid S_t)\big).$$

This is the amount by which the structured-memory method changes the answer distribution relative to the prefix-only method at that exact turn. The invariant is not that Δ_t is constant. The invariant is the chain-rule accounting: the final KL is the sum of these local increments.

Theorem 2 (Turnwise memory accounting)**.** *Let Q be the prefix-only conversation distribution and R the structured-memory conversation distribution, with conditionals*

$$q_t(Y_t \mid S_t), \qquad r_t(Y_t \mid S_t, C_t).$$

Assume $R \ll Q$ and that C_t is generated from the prior visible history by a fixed retrieval rule. Then

$$\boxed{\mathrm{D}_{\mathrm{KL}}(R\|Q) = \sum_{t=1}^{T} \mathbb{E}_R\, \mathrm{D}_{\mathrm{KL}}\big(r_t(\cdot \mid S_t, C_t) \| q_t(\cdot \mid S_t)\big).}$$

Moreover, for a true conversation distribution P,

$$\boxed{\mathrm{D}_{\mathrm{KL}}(P\|Q) - \mathrm{D}_{\mathrm{KL}}(P\|R) = \sum_{t=1}^{T} \mathbb{E}_P \left[\log \frac{r_t(Y_t \mid S_t, C_t)}{q_t(Y_t \mid S_t)}\right].}$$

Thus the final advantage of memory over prefix-only prediction is the accumulated signed likelihood gain across turns.

Proof. By the chain rule for autoregressive distributions,

$$R(Y_{1:T}) = \prod_{t=1}^{T} r_t(Y_t \mid S_t, C_t), \qquad Q(Y_{1:T}) = \prod_{t=1}^{T} q_t(Y_t \mid S_t).$$

Taking the log ratio gives

$$\log \frac{R(Y_{1:T})}{Q(Y_{1:T})} = \sum_{t=1}^{T} \log \frac{r_t(Y_t \mid S_t, C_t)}{q_t(Y_t \mid S_t)}.$$

Taking expectation under R gives the first identity. Taking expectation under P gives

$$\mathbb{E}_P \log \frac{R}{Q} = \mathrm{D}_{\mathrm{KL}}(P\|Q) - \mathrm{D}_{\mathrm{KL}}(P\|R),$$

which is the second identity. □

The similarity-weighted diagnostic

$$\mathcal{D}_T^{\rho}(R, Q) := \sum_{t=1}^{T} \mathbb{E}_R\big[\rho_t \Delta_t(S_t, C_t)\big]$$

measures the part of the model difference occurring at turns where memory is actually activated by relevant prior text. Since $0 \le \rho_t \le 1$,

$$0 \le \mathcal{D}_T^{\rho}(R, Q) \le \mathrm{D}_{\mathrm{KL}}(R\|Q).$$

If additionally $\Delta_t \le B$ almost surely, then

$$0 \le \mathrm{D}_{\mathrm{KL}}(R\|Q) - \mathcal{D}_T^{\rho}(R, Q) \le B \sum_{t=1}^{T} \mathbb{E}_R(1 - \rho_t).$$

So diffuse or irrelevant retrieval cannot explain much of the structured method's divergence unless the local perturbations are large anyway.

Discussion: what memory is worth. The first theorem says that the value of structure is not mysterious. It is measured by

$$\mathrm{I}_p(Y_i; C_i \mid S_i),$$

the amount of next-token uncertainty removed by the structural state after the visible prefix is already known. In this sense, useful memory is not merely more text. It is a state that remembers something predictive that the prefix alone does not efficiently expose.

The turnwise theorem adds the dynamic picture. Memory is a sequence of local distributional events. At each prompt turn, the current prefix reaches backward into prior text, activates some state C_t, and changes the response law from $q_t(\cdot \mid S_t)$ to $r_t(\cdot \mid S_t, C_t)$. If the remembered structure is relevant and useful, the truth-relative increment is positive. If the remembered structure is irrelevant or misleading, the increment can be negative. The final result is the accumulated signed sum of those moments.

This clarifies a possible failure mode. A structured method can carry a beautiful internal object C_i, but if C_i does not help predict Y_i beyond S_i, then the mutual information term is zero. The structure may be aesthetically or computationally meaningful, but it buys no expected log-loss advantage. Conversely, even a crude memory can be valuable if it preserves the right latent fact: the topic of a conversation, the invariant in a proof, the user's intent, or the unresolved dependency that a local prefix-only view would tend to blur.

So the comparison with next-token prediction should not be phrased only as

$$\mathrm{D}_{\mathrm{KL}}(\mathrm{truth}\|q) \quad \text{versus} \quad \mathrm{D}_{\mathrm{KL}}(\mathrm{truth}\|r).$$

That comparison is correct but too blunt. The sharper statement is that the expected advantage of the structured method decomposes into

$$\text{structural information} + \text{prefix-only fit error} - \text{structured fit error},$$

and, across a conversation, the divergence between the methods decomposes into turnwise memory increments.

Relation back to indistinguishability. The next-token paper shows that minimizing prefix-only loss can remove bounded distinguishers over k-token windows. The present note suggests a complementary refinement: if a class of distinguishers is really detecting information carried by a hidden or persistent state C_i, then the structural gain should show up as conditional mutual information and as accumulated turnwise memory divergence. In that case, memory is not an add-on to next-token prediction; it is a refinement of the conditioning information on which next-token prediction operates.