# LEARNING WHAT TO ATTEND TO: TRAINING DYNAMICS OF MULTI-TOKEN PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-token prediction (MTP) has improved next-token prediction in several large-scale settings, but how its additional supervision changes contextual cue selection remains unclear. In this work, we study a one-layer linear-attention model with two prediction heads sharing a contextual representation. We analyze a two-stage procedure that learns token-to-future associations in the readouts, then freezes the readouts and trains shared attention by gradient flow. Under explicit structural assumptions, we characterize the max-margin solutions governing attention-score growth. We show that MTP assigns leading logarithmic growth only to cues jointly predictive of both future tokens. In contrast, next-token prediction (NTP) assigns leading coefficients according to the occurrence and overlap of next-token-compatible cues, allowing ambiguous cues to receive as much or more leading growth as jointly predictive ones. We derive the resulting prediction rules on recombined contexts. An illustrative example shows that, as training time grows, the correct next-token probability tends to one under MTP and to zero under NTP on the same context, even when inference uses only MTP's first head. Synthetic experiments support the predicted cue preferences, max-margin directions, and prediction separation, while TinyStories experiments illustrate learned token-to-future associations and attention differences between NTP and MTP.

## 1 INTRODUCTION

Multi-token prediction (MTP) extends language-model supervision beyond the immediate next token and can improve both prediction and inference efficiency. Gloeckle et al. (2024) showed that training multiple prediction heads on a shared model trunk improves sample efficiency and code-generation performance, while enabling faster inference through self-speculative decoding. The gains grow with model size (Gloeckle et al., 2024), and curriculum-based MTP training also improves next-token performance and generation quality in smaller models (Aynetdinov & Akbik, 2025). The DeepSeek-V3 report finds gains on most evaluated benchmarks in its MTP ablations, even when the auxiliary MTP module is discarded at inference (DeepSeek-AI, 2025). These findings suggest that MTP's benefits extend beyond faster generation to the quality of next-token prediction.

Despite these empirical advances, the mechanisms underlying MTP's prediction gains remain only partially understood. Adding future-token targets provides more supervision, but does not by itself explain how the model uses contextual information differently. Recent theory has identified mechanisms in planning tasks, including the learning of transitive relations (Zhong et al., 2025) and reverse-reasoning dynamics (Huang et al., 2026). In this work, we study a complementary mechanism, i.e., *contextual cue selection through shared attention*.

Intuitively, a prefix can contain several cues compatible with the next token but associated with different later continuations. Next-token supervision alone need not separate these cues, but supervision on later tokens can favor the cues that also predict those tokens. MTP and next-token prediction (NTP) can both predict the next token correctly on the training data while relying on different cues, so their predictions can differ when familiar cues appear in new combinations. This leads to our central question:

> *How does farther-horizon supervision change contextual cue selection, and when does this improve next-token prediction?*



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study this question in a one-layer linear-attention model with two prediction heads, one for the next token and one for the token after it. Training proceeds in two stages. The readouts learn specified token-to-future associations by gradient descent, then remain fixed while gradient flow trains the shared attention matrix directly. We compare MTP and NTP using the same first-head readout and attention initialization, isolating the effect of second-token supervision on attention learning. Under structural assumptions on the data, we characterize the long-term growth of each cue's attention score and derive the resulting prediction rules on recombined contexts.

**Our contributions.** We establish the following results.

• *Max-margin characterization of attention learning.* With the readouts fixed, we express the margins between correct and competing output logits as linear functions of the shared attention scores. This formulation leads to a max-margin characterization of the scores' asymptotic growth, as in max-margin analyses of softmax attention (Tarzanagh et al., 2023b; Huang et al., 2024). MTP adds constraints from the second prediction head to those imposed by NTP. Under our structural assumptions, these constraints yield an explicit solution supported only on jointly predictive cues, whereas the NTP solution depends on the occurrence and overlap of next-token-compatible cues.

• *Quantitative analysis of coupled dynamics.* With the readouts fixed, attention training is gradient flow on a loss that is convex in the attention scores, so its direction is governed by a hard-margin solution, as in implicit-bias analyses. The two heads can give a cue opposing updates, and cues are shared across overlapping training contexts. Through a Lyapunov-type inequality, we control the deviation of every cue score from the max-margin direction and establish logarithmic growth with an $O(\sqrt{\log t})$ remainder as training time $t \to \infty$. This proves that MTP separates jointly predictive cues from all other cues at logarithmic order, whereas NTP need not.

• *Prediction rules and a next-token advantage.* We derive prediction rules on fixed recombined contexts. At leading order, MTP predictions depend on the number of jointly predictive training cues supporting each output, whereas NTP aggregates training-dependent weighted support. In the controlled setting where the context contains exactly one cue that was jointly predictive at both horizons during training, and the evaluation targets are defined by that cue's learned associations, MTP's correct next-token probability tends to one. NTP's tends to zero when its weighted support favors a competing output.

• *Experimental findings.* Synthetic experiments provide controlled tests of the theoretical predictions, supporting the predicted preferences for different input cues, alignment with the predicted max-margin directions, and separation in model predictions. On TinyStories, we complement these experiments by examining how input tokens are associated with future-token predictions through the learned readouts and how attention patterns differ between NTP and MTP. Together, these experiments connect the theoretical analysis to observable model behavior in a text-generation setting.

A detailed discussion of related work appears in Appendix A.

## 2 Problem Setup

*Notation.* Let $\mathcal{V}$ be a vocabulary of size $V$. A token identity $x$ has one-hot representation $e_x \in \mathbb{R}^V$. All vectors are column vectors; vectors and matrices use ordinary italic symbols, while $\mathbf{1}$ denotes the all-ones vector of the appropriate dimension. We write $\mathbb{1}\{\mathcal{A}\}$ for an indicator and $|\mathcal{S}|$ for the cardinality of a finite set.

### 2.1 Model and Prediction Targets

We consider labeled prefixes $(X_i, y_i^{(1)}, y_i^{(2)})$, where $X_i = [e_{x_{i,1}}, \ldots, e_{x_{i,L_i}}] \in \mathbb{R}^{V \times L_i}$. The query $q_i = x_{i,L_i}$ is the final prefix token and is included in the attention support $\mathcal{X}_i = \{x_{i,1}, \ldots, x_{i,L_i}\}$. The targets are the next two tokens after the full prefix: an observed sequence $(a_{i,1}, \ldots, a_{i,L_i+2})$ gives $x_{i,\ell} = a_{i,\ell}$ and $y_i^{(h)} = a_{i,L_i+h}$ for $h = 1, 2$.

For a generic prefix $X$ with final token $q$, our model has one shared linear-attention trunk and two horizon-specific prediction heads:

$$p_h(\cdot \mid X) = \mathrm{softmax}\Big(W_{uv}^{(h)} X X^\top W_{kq} e_q\Big), \qquad h \in \{1, 2\},$$

where $W_{kq}, W_{uv}^{(h)} \in \mathbb{R}^{V\times V}$. Both heads use the same contextual representation $XX^\top W_{kq}e_q$, but each has its own vocabulary readout $W_{uv}^{(h)}$. Each head predicts one future position from the same attention output; $p_h$ denotes its output distribution.

This parameterization corresponds algebraically to absorbing the key-query product $W_k^\top W_q$ into $W_{kq}$ and the value-output product $W_{\text{out}}^{(h)}W_v$ into $W_{uv}^{(h)}$. We optimize the composite matrices directly, rather than analyzing gradient descent on their factors, which can have a different implicit bias (Tarzanagh et al., 2023a). The unnormalized scores $X^\top W_{kq}e_q$ weight the context tokens directly; only the output distribution uses softmax.

Two horizons are the minimal setting in which farther-future supervision can distinguish cues compatible with the same next token. The head-$h$ cross-entropy loss is $\ell_h(W_{kq}, W_{uv}^{(h)}; X, y^{(h)}) = -\log p_h(y^{(h)} \mid X)$. The two-stage MTP training procedure is given in Section 3; the first-head-only NTP baseline is introduced separately in Section 4.

### 2.2 Collocations and Contextual Roles

**Definition 2.1** (Horizon-wise collocation)**.** For each contextual token $x$ and horizon $h \in \{1, 2\}$, let $n_h(x) \in \mathcal{V}$ denote its fixed associated output token, called the horizon-$h$ collocation of $x$.

This definition extends the collocations of Huang et al. (2024) to two horizons. The horizon is measured from the end of the full prefix, not from the position of $x$ inside it. A cue may occur anywhere in the prefix; the pair $(n_1(x), n_2(x))$ need not follow every occurrence of $x$. These fixed token-level associations are distinct from the observed sentence-level targets $(y_i^{(1)}, y_i^{(2)})$.

This abstraction is motivated by experiments on TinyStories (Eldan & Li, 2023), a dataset of short synthetic stories. In a trained one-layer MTP model, the highest-ranked readout outputs for `Once` are `_upon` in the first head and `_a` in the second. We model such associations using one designated output per token and horizon. Further details on the dataset, training setup, and readout analysis can be found in Appendix G.2.

We condition on these token-to-future associations to study how multi-horizon supervision affects contextual cue selection. The contextual data satisfy the structural role conditions below. For each contextual token $x \in \mathcal{X}_i$, its role is determined by comparing its collocations with the targets of sample $i$:

$$\begin{aligned}
\text{full cue:} \quad & n_1(x) = y_i^{(1)}, \quad n_2(x) = y_i^{(2)}, \\
\text{near cue:} \quad & n_1(x) = y_i^{(1)}, \quad n_2(x) \neq y_i^{(2)}, \\
\text{uninformative cue:} \quad & n_1(x) \neq y_i^{(1)}, \quad n_2(x) \neq y_i^{(2)}.
\end{aligned}$$

Write $\mathcal{F}_i, \mathcal{C}_i, \mathcal{W}_i$ for these three sets. The roles classify prefix tokens, including the query, relative to the sample's targets; they are not intrinsic vocabulary categories. Here *uninformative* means that neither designated association matches the targets. The remaining possibility, matching only the second target, is excluded by the prefix-consistency assumption below.

**Example 2.2** (Ambiguous contextual cues)**.** Consider the constructed prefix and continuation

*Prefix:* A baker with flour will
*Observed targets:* $y^{(1)} = \texttt{bake}, \quad y^{(2)} = \texttt{bread}.$

Assume the associations $\texttt{baker} \mapsto (\texttt{bake}, \texttt{bread})$ and $\texttt{flour} \mapsto (\texttt{bake}, \texttt{today})$. The remaining words `A`, `with`, and `will` each have association $(\texttt{rest}, \texttt{today})$. Thus baker is a full cue, flour is a near cue, and the remaining words are uninformative. Both `baker` and `flour` support the next token `bake`; the second target `bread` distinguishes them.

### 2.3 Structural Data Assumptions

Let $\mathcal{D}$ be the indexed training dataset and $\mathcal{D}_q = \{i : q_i = q\}$ the indices of samples with query $q$. We impose the following conditions on its contextual examples.

**Assumption 1.** The training set satisfies:

- *Distinct tokens.* Token identities within each prefix $X_i$ are distinct.
- *Prefix-consistent predictive roles.* For every sample $i$ and every $x \in \mathcal{X}_i$,
$$n_2(x) = y_i^{(2)} \quad \Longrightarrow \quad n_1(x) = y_i^{(1)}.$$
- *Unique full cue.* Each sample contains exactly one full cue, denoted $x_i^\star$:
$$n_1(x_i^\star) = y_i^{(1)}, \qquad n_2(x_i^\star) = y_i^{(2)}.$$
- *Inactive class.* For every sample $i$ and head $h$, there is an output $\bar{a}_i^{(h)} \neq y_i^{(h)}$ such that
$$n_h(x) \neq \bar{a}_i^{(h)}, \qquad \forall x \in \mathcal{X}_i.$$
- *Role consistency.* For every query $q$, every $i, j \in \mathcal{D}_q$, every $x \in \mathcal{X}_i \cap \mathcal{X}_j$, and $h \in \{1, 2\}$,
$$\mathbb{1}\{n_h(x) = y_i^{(h)}\} = \mathbb{1}\{n_h(x) = y_j^{(h)}\}.$$

*Prefix consistency* excludes far-only cues, which match only the second target. With such cues, the two targets can be supported by different tokens, the heads compete for the same attention scores, and MTP can predict the next token worse than NTP (Remark C.7). The restriction isolates how second-token supervision distinguishes jointly predictive cues from cues compatible only with the next token. The *unique-full-cue condition* ensures that each sample contains exactly one jointly predictive cue. Distinctness lets token sets describe attention contributions without multiplicity weights. An *inactive class* is an output to which no contextual token points at that horizon; it is still in the vocabulary and can have positive softmax probability. Since the full cue supports the target, any unrepresented output is a competitor. Such a class exists whenever $V > |\{n_h(x) : x \in \mathcal{X}_i\}|$, in particular if $V > L_i$. *Role consistency* allows overlapping prefixes, but a shared token cannot change its predictive role across samples with the same query. For example, a full cue cannot become near in another same-query sample. This makes the roles compatible with a single learned score for each token-query pair.[1] Together, these conditions make it possible to characterize how the second target changes the max-margin solution.

## 3 MTP Training and Dynamics

### 3.1 Two-Stage MTP Training

We use a two-stage training procedure with fixed step sizes $\eta_1, \eta_2 > 0$. Stage 1 learns the token-to-future associations in the prediction heads, while Stage 2 freezes these heads and trains the shared attention matrix. We initialize $W_{uv}^{(1)}(0) = W_{uv}^{(2)}(0) = 0,\ W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$.

**Stage 1: learning token-to-future associations.** Stage 1 trains the readouts on prefixes containing a single token.

**Assumption 2** (Stage 1 training data)**.** Let $\mathcal{V}_{\mathrm{c}}$ be the set of token identities appearing in Stage 2 training prefixes. The Stage 1 dataset is $\mathcal{D}_{\mathrm{col}} := \{([e_x], n_1(x), n_2(x)) : x \in \mathcal{V}_{\mathrm{c}}\}$. Each distinct token contributes one training example, regardless of its frequency in Stage 2.

During Stage 1, $W_{kq}$ is fixed at its initialization and only the readout matrices are trained. For head $h$, the objective is

$$\mathcal{L}_{\mathrm{col}}^{(h)}\left(W_{uv}^{(h)}\right) = \frac{1}{|\mathcal{D}_{\mathrm{col}}|} \sum_{([e_x], n_1(x), n_2(x)) \in \mathcal{D}_{\mathrm{col}}} \ell_h\left(W_{kq}(0), W_{uv}^{(h)}; [e_x], n_h(x)\right),$$

where $\ell_h$ is the standard cross-entropy loss of prediction head $h$. Starting from the initialization, we apply the discrete gradient-descent updates

$$W_{uv}^{(h)}(k+1) = W_{uv}^{(h)}(k) - \eta_1 \nabla_{W_{uv}^{(h)}} \mathcal{L}_{\mathrm{col}}^{(h)}\left(W_{uv}^{(h)}(k)\right), \qquad h \in \{1, 2\}. \tag{1}$$

After a fixed number $K_1 \geq 1$ of Stage 1 updates, we define $\widehat{W}_{uv}^{(h)} := W_{uv}^{(h)}(K_1),\ h \in \{1, 2\}$, and freeze these readout matrices throughout Stage 2.

[1] Because $q$ occurs in every prefix in $\mathcal{D}_q$, role consistency also requires $n_h(q)$ to differ from every horizon-$h$ target unless all these prefixes share it. In Example 5.4, the query `will` is therefore uninformative.

**Stage 2: learning contextual attention.** Stage 2 uses the full-sentence training set $\mathcal{D}$. We freeze $W_{uv}^{(h)} = \widehat{W}_{uv}^{(h)}$, $h \in \{1,2\}$, and train only $W_{kq}$. Both prediction heads contribute to the MTP objective:

$$\mathcal{L}_{\mathrm{MTP}}(W_{kq}) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \sum_{h=1}^{2} \ell_h \left( W_{kq}, \widehat{W}_{uv}^{(h)}; X_i, y_i^{(h)} \right). \tag{2}$$

The discrete gradient-descent update is $W_{kq}(k+1) = W_{kq}(k) - \eta_2 \nabla_{W_{kq}} \mathcal{L}_{\mathrm{MTP}}\big(W_{kq}(k)\big)$.

Stage 1 is analyzed directly under the discrete gradient-descent updates above. For Stage 2, our theoretical analysis studies the corresponding continuous-time gradient-flow dynamics. We use $k$ for the discrete iteration index and $t \geq 0$ for continuous training time. Specifically, with a constant rate $\eta_{\mathrm{flow}} > 0$, we let $\dot{W}_{kq}(t) = -\eta_{\mathrm{flow}} \nabla_{W_{kq}} \mathcal{L}_{\mathrm{MTP}}\big(W_{kq}(t)\big)$. In the score dynamics below, we write $\eta := \eta_{\mathrm{flow}}/|\mathcal{D}|$, so that $\eta$ multiplies unaveraged sums of per-sample gradients. Unless stated otherwise, Stage 2 asymptotics are taken as $t \to \infty$ under this gradient flow, with the training data, vocabulary, and Stage 1 checkpoint $K_1$ (hence the frozen readouts) fixed.

### 3.2 Learned Collocation Structure

**Proposition 3.1** (Stage 1 collocation structure and equal readout scales)**.** *After any finite number $K_1 \geq 1$ of Stage 1 gradient-descent iterations, define $\widehat{W}_{uv}^{(h)} := W_{uv}^{(h)}(K_1)$. Then, for every head $h \in \{1,2\}$ and every $x \in \mathcal{V}_{\mathrm{c}}$, $\widehat{W}_{uv}^{(h)} e_x = G(e_{n_h(x)} - \frac{1}{V}\mathbf{1})$, where $G > 0$ is the same for both heads and every covered token identity $x$. Its value depends on the vocabulary size, the Stage 1 learning rate and loss normalization, and the checkpoint $K_1$.*

**Remark 3.2.** In each head, the associated output of every covered token thus has a logit $G$ larger than every other output; the two heads may favor different outputs, but their gaps are matched. The gap is common because each readout column is trained only by its own single-token example, and both heads use the same examples, initialization, learning rate, loss normalization, and checkpoint in Equation (1). Every column, whatever its associated output, therefore follows the same scalar recurrence (Appendix C.1). The shift $-(G/V)\mathbf{1}$ does not affect softmax, so Stage 2 depends on the frozen readouts only through $G$ and $n_h$. The matching matters: with a smaller first-head gap, near cues can also enter MTP's hard-margin solution (see Remark C.8).

### 3.3 MTP Attention Dynamics and Cue Selection

We now study how MTP modifies the shared attention map during Stage 2. For a fixed query token $q$, let $\mathcal{U}_q = \bigcup_{i:q_i=q} \mathcal{X}_i$ be the set of token identities appearing with $q$, and collect their attention scores as $s_q(t) = (s_{x,q}(t))_{x \in \mathcal{U}_q}$, where $s_{x,q}(t) = e_x^\top W_{kq}(t) e_q$. Here $s_q(t) \in \mathbb{R}^{|\mathcal{U}_q|}$ is the score vector, whereas $s_{x,q}(t) \in \mathbb{R}$ is its coordinate for token $x$; equivalently, $[s_q(t)]_x = s_{x,q}(t)$. Let $\mathcal{F}_q = \{x_i^\star : q_i = q\}$ denote the set of full cues associated with $q$. For every sample $i$ with $q_i = q$, head $h \in \{1,2\}$, and competitor $a \neq y_i^{(h)}$, define

$$[\phi_{i,a}^{(h)}]_x = G \mathbb{1}\{x \in \mathcal{X}_i\} \left[ \mathbb{1}\{n_h(x) = y_i^{(h)}\} - \mathbb{1}\{n_h(x) = a\} \right].$$

The inner product $\langle s_q(t), \phi_{i,a}^{(h)} \rangle$ is the logit gap between the correct output $y_i^{(h)}$ and competitor $a$ for sample $i$ and head $h$.

**Lemma 3.3** (MTP score dynamics)**.** *For sample $i$, prediction head $h$, and output token $a \in \mathcal{V}$, let $p_i^{(h)}(a,t) := [p_h(\cdot \mid X_i)]_a$ denote the probability assigned to token $a$ at time $t$, where $p_h$ is evaluated at the Stage 2 parameter $W_{kq}(t)$ and the frozen readout matrix $\widehat{W}_{uv}^{(h)}$. Under Assumptions 1 and 2, with the initialization of Section 3.1,*

$$\dot{s}_q(t) = \eta \sum_{i:q_i=q} \sum_{h=1}^{2} \sum_{a \neq y_i^{(h)}} p_i^{(h)}(a,t) \phi_{i,a}^{(h)}. \tag{3}$$

Lemma 3.3 shows that, within each training sample, a full cue receives positive updates from both heads, a near cue receives opposing updates, and an uninformative cue receives negative updates

from both heads. A near cue's net update can have either sign, and a shared cue accumulates updates from every training prefix with the same query. These coupled updates still form a convex problem. With the readouts frozen, every logit margin is linear in $s_q$, so Stage 2 is gradient flow on a multiclass cross-entropy that is convex in $s_q$, and implicit-bias results imply convergence in direction to a hard-margin solution (Soudry et al., 2018; Lyu & Li, 2020). The second head changes this solution by adding margin constraints.

**Proposition 3.4** (MTP max-margin structure)**.** *Under Assumptions 1 and 2, with the initialization of Section 3.1, for a fixed query $q$, $c_q^\star = \frac{1}{G}\sum_{x\in\mathcal{F}_q} e_x$ uniquely solves*

$$\min_{c\in\mathbb{R}^{|\mathcal{U}_q|}} \frac{1}{2}\|c\|^2 \quad \textit{subject to} \quad \left\langle c, \phi_{i,a}^{(h)} \right\rangle \geq 1,$$

*with constraints for every $i$ with $q_i = q$, $h \in \{1,2\}$, and $a \neq y_i^{(h)}$.*

**Theorem 3.5** (MTP score concentration)**.** *Under the assumptions in Lemma 3.3 and Proposition 3.4, as $t \to \infty$,*

$$s_q(t) = c_q^\star \log t + O\left(\sqrt{\log t}\right).$$

*Consequently,*

$$s_{x,q}^{\mathrm{MTP}}(t) = \frac{1}{G}\log t + O\left(\sqrt{\log t}\right), \qquad x \in \mathcal{F}_q,$$
$$s_{x,q}^{\mathrm{MTP}}(t) = O\left(\sqrt{\log t}\right) = o(\log t), \qquad x \in \mathcal{U}_q \setminus \mathcal{F}_q.$$

*Proof sketch.* The key technical step is to control the cumulative effect of opposing head updates and overlapping training contexts. *Step 1: Bound the deviation from the max-margin direction.* Let $\bar{s}_q(t)$ denote the average full-cue score. Using the margin structure in Proposition 3.4, we derive a Lyapunov-type inequality for the normalized scores that controls all cue coordinates simultaneously. This yields $s_q(t) = G\bar{s}_q(t)c_q^\star + O(\sqrt{\bar{s}_q(t)})$, quantifying both the alignment of full-cue scores and the suppression of other cues relative to them. *Step 2: Establish logarithmic growth with a remainder bound.* We translate the score estimate into bounds on the logit margins and hence on the average-score derivative. Integration gives $\bar{s}_q(t) = G^{-1}\log t + O(\sqrt{\log t})$. Substitution yields the claimed expansion, including its $O(\sqrt{\log t})$ remainder. The complete proof can be found in Appendix C.2.

**Remark 3.6.** For every observed query $q$ with $\mathcal{U}_q \setminus \mathcal{F}_q \neq \varnothing$, Theorem 3.5 implies

$$\min_{x\in\mathcal{F}_q} s_{x,q}^{\mathrm{MTP}}(t) - \max_{x\in\mathcal{U}_q\setminus\mathcal{F}_q} s_{x,q}^{\mathrm{MTP}}(t) = \frac{1}{G}\log t + O\left(\sqrt{\log t}\right).$$

Thus MTP eventually assigns every jointly predictive cue a larger attention score than every near or uninformative cue associated with the same query. In particular, compatibility with the next token alone does not give near cues leading logarithmic growth.

## 4 Comparison with Next-Token Prediction

We now ask how the score allocation changes when training uses only the next-token target. The matched NTP baseline has the same training prefixes, next-token targets, frozen first-head readout $\widehat{W}_{uv}^{(1)}$, and attention initialization $W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$ as MTP. The first-head Stage 1 objective is independent of the second head, so the same examples, initialization, step size, and checkpoint produce this readout using only $n_1(x)$ labels. Stage 2 retains only the $h = 1$ term in Equation (2), with the same gradient-flow rate $\eta_{\mathrm{flow}}$. Its score dynamics therefore follow Equation (3) with only the first-head terms, evaluated at the NTP parameters.

For an observed query $q$, let $A$ have rows indexed by samples with $q_i = q$ and columns by $x \in \mathcal{U}_q$. Its first-horizon incidence entries are $[A]_{i,x} = \mathbb{1}\{x \in \mathcal{X}_i\}\mathbb{1}\{n_1(x) = y_i^{(1)}\}$. Under our structural assumptions, the first-horizon hard-margin problem has the same unique minimizer as

$$\min_{d\in\mathbb{R}^{|\mathcal{U}_q|}} \frac{1}{2}\|d\|^2 \quad \text{subject to} \quad GAd \succeq \mathbf{1}.$$

Let $d_q$ denote its unique solution. Each constraint requires the sum of compatible cue coefficients in a training sample to be at least $1/G$. Full and near cues can jointly satisfy this bound. In contrast, MTP's second-head constraints require each full-cue coefficient individually to be at least $1/G$. Removing these constraints allows NTP to distribute coefficients among compatible cues according to their occurrence and overlap across training samples.

**Theorem 4.1** (NTP score allocation)**.** *Under Assumptions 1 and 2 and the matched two-stage training protocol, Stage 2 gradient flow satisfies, as $t \to \infty$,*

$$s_q^{\mathrm{NTP}}(t) = d_q \log t + O\left(\sqrt{\log t}\right).$$

**Corollary 4.2** (Failure of global full-cue separation under NTP)**.** *If a full cue $x^\star$ and a near cue $x^c$ for the same query satisfy $[A]_{i,x^c} \geq [A]_{i,x^\star}$ for every row $i$, then $d_{x^c,q} \geq d_{x^\star,q}$.*

Thus NTP need not give full cues a leading-order advantage over near cues. A strict coefficient inequality implies that the near cue eventually outranks the full cue; equal coefficients leave their ordering to lower-order terms. In contrast, MTP gives full cues a positive logarithmic-order gap over all non-full cues under the same structural assumptions. Appendix D gives the score dynamics, max-margin reduction, and proofs of the allocation and comparison results.

## 5 Prediction Consequences

The different score coefficients learned by MTP and NTP can lead to different next-token predictions. We first characterize how these coefficients determine predictions on a fixed context, then construct a recombined context on which MTP selects the correct token and NTP selects an incorrect token.

### 5.1 From Attention Scores to Predictions

Fix an observed query $q$ and a finite evaluation context $X_{\mathrm{test}} = [e_{x_1}, \ldots, e_{x_{L_{\mathrm{test}}}}]$ with $x_{L_{\mathrm{test}}} = q$. Its token identities are distinct, and every token belongs to the Stage 1-covered set $\mathcal{V}_{\mathrm{c}}$ and hence has the frozen readout in Proposition 3.1. Write $\mathcal{X}_{\mathrm{test}} = \{x_1, \ldots, x_{L_{\mathrm{test}}}\}$. The context is fixed independently of $t$ and need not have occurred during Stage 2. It may contain tokens previously observed with $q$, as well as covered tokens never paired with $q$ during Stage 2. For the latter, $s^{\mathsf{A}}_{x,q}(t) = 1$ remains at initialization under either objective $\mathsf{A} \in \{\mathrm{MTP}, \mathrm{NTP}\}$.

After dropping a class-independent shift, the first-head logits are

$$Z^{(1)}_{\mathsf{A}}(a,t) = G \sum_{\substack{x \in \mathcal{X}_{\mathrm{test}}: \\ n_1(x) = a}} s^{\mathsf{A}}_{x,q}(t), \qquad a \in \mathcal{V}. \tag{4}$$

Dropping this shift leaves the softmax probabilities unchanged. When the maximizing output is unique, denote it by $\widehat{y}^{(1)}_{\mathsf{A}}(t) = \arg\max_{a \in \mathcal{V}} Z^{(1)}_{\mathsf{A}}(a,t)$.

**Theorem 5.1** (Prediction rule for a fixed context)**.** *Under the training assumptions of Theorems 3.5 and 4.1, define*

$$\mu_a = \left|\{x \in \mathcal{X}_{\mathrm{test}} \cap \mathcal{F}_q : n_1(x) = a\}\right|, \qquad \lambda_a = G \sum_{\substack{x \in \mathcal{X}_{\mathrm{test}} \cap \mathcal{U}_q: \\ n_1(x) = a}} d_{x,q}. \tag{5}$$

*Then, as $t \to \infty$, the effective logits satisfy*

$$Z^{(1)}_{\mathrm{MTP}}(a,t) = \mu_a \log t + O(\sqrt{\log t}), \qquad Z^{(1)}_{\mathrm{NTP}}(a,t) = \lambda_a \log t + O(\sqrt{\log t}).$$

*For either objective, if its leading coefficient has a unique maximizer $a^\star$, the probability of $a^\star$ tends to one. If the largest coefficient is tied, the leading-order result alone does not determine the ordering among the tied outputs.*

Theorem 5.1 gives a direct comparison: under MTP, the coefficient of $\log t$ in each output logit equals the number of full cues supporting that output, whereas NTP sums training-dependent coefficients of next-token-compatible cues. The proof combines Equation (4) with the score asymptotics; see Appendix E.2.

**Remark 5.2.** We specialize Theorem 5.1 to a recombination setting with exactly one training full cue per test context in Section 5.2. The cue's associations define the evaluation targets. This setting admits an explicit prediction separation between MTP and NTP. The theorem also covers other contextual compositions. With no full cue, all MTP coefficients $\mu_a$ vanish, so the leading-order result does not determine the prediction. With multiple full cues, an output supported by strictly more full cues than every competitor is selected. Lower-order terms determine the ordering when the leading coefficients tie. Appendix F extends the fixed-collocation setting of Definition 2.1 by allowing different target pairs for the same Stage 1 input token, and derives the corresponding attention-score updates and sufficient conditions for cue-score ordering.

### 5.2 Recombination with a Single Full Cue

We now restrict $X_{\text{test}}$ to contain exactly one training full cue: $\mathcal{X}_{\text{test}} \cap \mathcal{F}_q = \{x^\star\}$. Set the evaluation targets to $y^{(h)} = n_h(x^\star)$ for $h = 1, 2$. All other tokens are either learned non-full cues for $q$ or Stage 1-covered tokens never paired with $q$ during Stage 2. The coefficients $\lambda_a$ are defined in Equation (5).

**Corollary 5.3** (Prediction under controlled recombination)**.** *For the same fixed test context and target pair, as Stage 2 training time $t \to \infty$,*

$$p_h^{\text{MTP}}(y^{(h)} \mid X_{\text{test}}, q, t) \longrightarrow 1, \qquad h \in \{1, 2\}.$$

*For NTP,*

$$p_1^{\text{NTP}}(y^{(1)} \mid X_{\text{test}}, q, t) \longrightarrow \begin{cases} 1, & \lambda_{y^{(1)}} > \max_{a \neq y^{(1)}} \lambda_a, \\ 0, & \lambda_{y^{(1)}} < \max_{a \neq y^{(1)}} \lambda_a. \end{cases}$$

For MTP's first head, the unique full cue makes $\mu_{y^{(1)}} = 1$ and all competing coefficients zero. The same argument applies to its second head with $n_2$ in place of $n_1$. NTP instead follows the weighted comparison in Theorem 5.1. Appendix E.1 gives the detailed logit-gap argument.

**Example 5.4** (A recombination separating MTP and NTP)**.** We extend the baker/flour example in Example 2.2 with a pianist context, using word-level tokens and the associations below.

*Training contexts.* Use the following Stage 2 prefixes and two-token continuations, with `will` as the common query:

```
A baker with flour will                  ↦ bake bread
A pianist with keyboards backstage will ↦ play music
```

Stage 1 single-token examples cover all eight contextual words and specify these associations:

| Context token(s) | $n_1$ | $n_2$ |
|---|---|---|
| `baker` | `bake` | `bread` |
| `flour` | `bake` | `today` |
| `pianist` | `play` | `music` |
| `keyboards`, `backstage` | `play` | `today` |
| `A`, `with`, `will` | `rest` | `today` |

Thus `baker` and `pianist` are the full cues in their respective training prefixes. The near cues are `flour` in the first and `keyboards`, `backstage` in the second. The shared words `A`, `with`, and `will` are uninformative in both. Include the unused output `walk` as an inactive class for both heads; the data then satisfy Assumption 1.

*Learned allocation.* For these two prefixes, NTP's minimum-norm solution divides the leading score growth equally among the next-token-compatible cues in each prefix. Each of `baker` and `flour` therefore contributes $1/2$ to the coefficient of $\log t$ in the `bake` logit. Each of `pianist`, `keyboards`, and `backstage` contributes $1/3$ to the `play` logit. Under MTP, each full cue contributes 1 to the coefficient of $\log t$ in its associated output's logit, while all other cues contribute only lower-order terms.

*Recombined test.* Replace `pianist` in the second prefix with `baker`, and evaluate both models against the continuation `bake bread`:

```
A baker with keyboards backstage will ↦ bake bread.
```

The leading first-head logit coefficients are

| | `bake` | `play` | Limiting prediction |
|---|---|---|---|
| NTP | $\frac{1}{2}$ | $\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$ | `play` |
| MTP | 1 | 0 | `bake` |

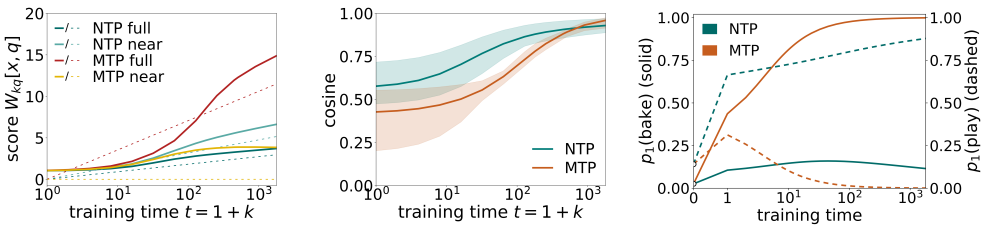
All other output coefficients are zero. NTP adds the contributions of `keyboards` and `backstage`, whose combined support for `play` exceeds the support for `bake`. Under MTP, `baker` is the only full cue present and supplies the only leading contribution.

Hence, as $t \to \infty$, the probability of the correct next token `bake` tends to one under MTP and to zero under NTP. MTP also predicts `bread` at the second horizon with probability tending to one.

## 6 Experiments

*Synthetic experiments.* We construct sequences with full, near, and uninformative cues and use the two-stage training protocol. In Figure 1(a), the MTP full-cue score grows approximately linearly with $\log(t)$ at later checkpoints, while the near-cue score changes little. The coefficients of $\log(t)$ are $1.53$ and $0$ for full-cue and near cue respectively. Under NTP, both scores increase, and near cue has a leading coefficient of $0.69$, which is even larger than full cue's coefficient $0.40$. The score directions also approach their respective max-margin solutions (Figure 1(b)), consistent with Theorems 3.5 and 4.1. Appendix G.1 gives the setup and additional results, including cue-ordering heatmaps and a softmax-attention variant.

An experiment based on the sentence in Example 5.4 is also aligned with our results. We construct a small dataset using all tokens appearing in this example and train the model with the same step size and initialization as in the preceding synthetic experiments. The readout gap entering Stage 2 is $G = 1.70$. We then track the model outputs on the recombined sentence at each checkpoint throughout training. We record the output probability of MTP and NTP on token `bake` and `play` in Figure 1(c). NTP fails slowly: its probability ratio $p_1^{\mathrm{NTP}}(\texttt{bake})/p_1^{\mathrm{NTP}}(\texttt{play})$ is $t^{-1/6}e^{O(\sqrt{\log t})}$, consistent with the NTP curves in the figure.




*(a) Score growth* *(b) Cosine similarity* *(c) Prediction*

Figure 1: (a) Scores of a full cue and a near cue in one selected synthetic sequence, where the solid curves are the observed results and dashed ones are their theoretical leading terms. (b) Cosine alignment with each objective's max-margin solution; lines and bands show the mean and range across queries. (c) First-head prediction separation on the recombined context in Example 5.4. The theoretical limits of the correct-token probability are one for MTP and zero for NTP.

*Real-world experiments.* Additional experiments on TinyStories illustrate learned token-to-future associations and attention differences between NTP and MTP; see Appendix G.2.

## 7 Conclusion

We characterize how multi-token supervision changes cue selection in a two-stage linear-attention model. Under the stated structural assumptions, MTP assigns leading score growth to jointly predictive cues, while NTP's leading coefficients depend on the occurrence and overlap of next-token-compatible cues. These different cue preferences yield a next-token prediction separation on recombined contexts, even when inference uses only MTP's first head. Experimental results support our theoretical findings on cue selection under MTP.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with developing and implementing code for experiments on synthetic and real-world datasets and polishing the manuscript. We have not used generative AI tools to generate research ideas or methodological approaches, or develop or complete the mathematical proofs. All AI-assisted code was reviewed, tested, and verified by the authors, and all AI-assisted language edits were manually reviewed to ensure that they preserved the intended technical meaning and claims. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

We have aimed to make our work reproducible by providing detailed descriptions of the method, experimental setup, implementation choices, and evaluation procedures in the paper and appendix. We also include source code, configuration files, and instructions for reproducing the experiments in the supplementary material.

## REFERENCES

Kwangjun Ahn, Alex Lamb, and John Langford. Efficient joint prediction of multiple future tokens, 2025. URL https://arxiv.org/abs/2503.21801.

Ansar Aynetdinov and Alan Akbik. Pre-training curriculum for multi-token prediction in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25573–25588. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1243. URL https://aclanthology.org/2025.acl-long.1243/.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2296–2318. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bachmann24a.html.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3X2EbBLNsk.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5209–5235. PMLR, 2024. URL https://proceedings.mlr.press/v235/cai24b.html.

Ziqiang Cui, Han Shi, Bowei He, Yu Pan, Peiyang Liu, Shengyin Sun, Yankai Chen, Haoli Bai, Yichun Yin, Xue Liu, and Chen Ma. Adamtp: An adaptive training paradigm for multi-token prediction, 2026. URL https://arxiv.org/abs/2608.00434.

DeepSeek-AI. DeepSeek-V3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=pEWAcejiU2.

Jianhao Huang, Zhanpeng Zhou, Renqiu Xia, Baharan Mirzasoleiman, Weijie J Su, and Wei Huang. How transformers learn to plan via multi-token prediction. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=1uj8eRv3aJ.

Ruiquan Huang, Yingbin Liang, and Jing Yang. Non-asymptotic convergence of training transformers for next-token prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=NfOFbPpYII.

Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25516–25549. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25s.html.

Shawn Im, Changdae Oh, Zhen Fang, and Sharon Li. How do transformers learn to associate tokens: Gradient leading terms bring mechanistic interpretability. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=A4Us8jxVGq.

Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *Journal of Machine Learning Research*, 27(36):1–61, 2026. URL http://jmlr.org/papers/v27/25-0549.html.

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 685–693. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/li24f.html.

Xiaohao Liu, Xiaobo Xia, Weixiang Zhao, Manyi Zhang, Xianzhi Yu, Xiu Su, Shuo Yang, See-Kiong Ng, and Tat-Seng Chua. L-MTP: Leap multi-token prediction beyond adjacent context for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=0VDmWjW456.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeLIgBKPS.

Divyat Mahajan, Sachin Goyal, Badr Youbi Idrissi, Mohammad Pezeshki, Ioannis Mitliagkas, David Lopez-Paz, and Kartik Ahuja. Beyond multi-token prediction: Pretraining LLMs with future summaries. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aeYIFVn4vb.

Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Hi0SyHMmkd.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2401–2410. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.217. URL https://aclanthology.org/2020.findings-emnlp.217/.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=JlWn80mTJi.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/c4127b9194fe8562c64dc0f5bf2c93bc-Abstract.html.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2023a. URL https://arxiv.org/abs/2308.16898.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=WXc8O8ghLH.

Christos Thrampoulidis. Implicit optimization bias of next-token prediction in linear models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xSziO6gQgG.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Shaolei Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=l3HUgVHqGQ.

Diyuan Wu, Aleksandr Shevchenko, Samet Oymak, and Marco Mondelli. Attention with trained embeddings provably selects important tokens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=y5IUGnpDJ8.

Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=qyilOnIRHI.

Qimin Zhong, Hao Liao, Siwei Wang, Mingyang Zhou, Xiaoqun Wu, Rui Mao, and Wei Chen. Understanding and enhancing the planning capability of language models via multi-token prediction, 2025. URL https://arxiv.org/abs/2509.23186.

Qimin Zhong, Hao Liao, Haiming Qin, Mingyang Zhou, Rui Mao, Wei Chen, and Naipeng Chao. Toward consistent world models with multi-token prediction and latent semantic enhancement. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13582–13602, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.618. URL https://aclanthology.org/2026.acl-long.618/.

# A Related Work

**Multi-token prediction and future supervision.** Early work predicted several future tokens for faster decoding (Stern et al., 2018) and for sequence-to-sequence pretraining (Qi et al., 2020); extra decoding heads can also speed up LLM inference (Cai et al., 2024). For language-model pretraining, Gloeckle et al. (2024) attach independent output heads to a shared trunk to predict several future tokens in parallel, improving sample efficiency and downstream performance; our model abstracts this parallel-head design. DeepSeek-V3 instead predicts the additional tokens with sequential modules that keep the causal chain at each depth (DeepSeek-AI, 2025), and curriculum schedules help small models benefit from MTP (Aynetdinov & Akbik, 2025). Variants predict summaries of the long-term future (Mahajan et al., 2026), anchor predictions to latent state trajectories (Zhong et al., 2026), target non-adjacent tokens (Liu et al., 2025), or adapt the prediction depth to each token (Cui et al., 2026). Joint multi-token prediction instead models dependencies among future tokens through teacher forcing and a representation bottleneck (Ahn et al., 2025). Controlled tasks also expose limits of teacher-forced NTP: it can learn shortcut solutions on path-star graphs (Bachmann & Nagarajan, 2024), and it is myopic on open-ended tasks where teacherless and diffusion objectives produce more diverse and original outputs (Nagarajan et al., 2025); these objectives are not parallel-head MTP. These works motivate future supervision, but do not characterize how a second-head supervision changes the attention it shares with the first head.

Existing theory identifies other mechanisms behind MTP. Transfer matrices can propagate future supervision through a fixed attention pattern to learn transitive reachability (Zhong et al., 2025); a farther-future head can supervise an intermediate layer and induce depth-wise gradient decoupling (Huang et al., 2026); and gradient coupling can contract representations toward belief states (Zhong et al., 2026). We instead attach both horizon-specific heads to the same attention output and freeze their readouts. This isolates how the second head changes cue selection, the leading score coefficients, and the resulting first-head prediction.

**NTP training dynamics and token associations.** Several NTP analyses separate stored token associations from contextual selection. Weight matrices can act as associative memories while attention develops context-specific retrieval (Bietti et al., 2023). Other analyses connect corpus statistics to learned associations, either through early-training parameter approximations (Im et al., 2026) or through sparse-plus-low-rank NTP geometry (Zhao et al., 2024). Our question is not whether NTP can learn these mechanisms, but how the selection rule changes when the same representation must also predict a farther token.

The closest modeling lineage analyzes one-layer transformers whose attention is trained together with, or after, a decoder. When the decoder learns faster than attention, attention shifts from common toward distinctive key tokens, ordered by their co-occurrence with the query, and then settles on a nearly fixed combination (Tian et al., 2023). Huang et al. (2024) separate the two time scales completely: they train the feed-forward layer on collocation pairs, freeze it, and then train softmax attention for NTP. Both layers converge in direction to max-margin solutions, and the trained model predicts non-trivially under dataset shift. We adopt this two-stage protocol and the collocation abstraction, but use linear attention, which makes Stage 2 convex in the attention scores, and compare matched NTP and MTP objectives. Under our structural assumptions, MTP's hard-margin solution is explicit, whereas NTP's leading score coefficients depend on the occurrence and overlap of next-token-compatible cues, much as co-occurrence shapes attention in Tian et al. (2023). On recombined contexts, a form of dataset shift, the two sets of coefficients can yield different first-head predictions.

**Implicit bias and max-margin attention selection.** Implicit-bias results connect vanishing classification loss to max-margin directions in binary and multiclass prediction (Soudry et al., 2018; Ravi et al., 2024) and in homogeneous models (Lyu & Li, 2020). With frozen readouts, our Stage 2 is gradient flow on a cross-entropy that is convex in the attention scores, so these results determine its leading direction; we characterize the resulting hard-margin solutions and bound the deviation of the scores from them. For NTP in linear models, when a context can be followed by several tokens, gradient descent fits their log-odds within a data subspace and diverges along an NTP-specific max-margin direction outside it (Thrampoulidis, 2024). Under Assumption 1, identical training contexts have identical targets, so only the max-margin component appears at leading order.

For attention, gradient-based training selects locally optimal tokens through max-margin separation (Tarzanagh et al., 2023b), with extensions to generalized $\ell_p$ solutions and convergence rates (Julistiono et al., 2026). For NTP with self-attention, the weights converge in direction to a max-margin solution defined by a token-priority graph, which retrieves the highest-priority tokens, while a finite component composes them (Li et al., 2024). In a one-layer attention classifier, trained embeddings encode token-label frequencies and subsequently support max-margin selection of predictive tokens (Wu et al., 2025). In one-layer transformers for regular-language recognition, the linear layer grows logarithmically toward a max-margin direction once attention stabilizes (Huang et al., 2025). These analyses, like those of Tian et al. (2023) and Huang et al. (2024), use softmax attention, for which token selection is non-convex and gradient descent can converge to locally optimal tokens. Linear attention keeps our Stage 2 convex, so the global hard-margin solution characterizes the leading score growth; Figure 5 examines a softmax variant empirically. The parameterization also matters: with vanishing regularization, the combined key–query matrix converges in direction to a Frobenius-norm attention SVM, whereas its factors yield a nuclear-norm one (Tarzanagh et al., 2023a). We train the composite $W_{kq}$; with one-hot tokens, its Frobenius norm separates across queries, so the hard-margin problem decomposes into one program per query. These results establish max-margin attention selection, but do not compare supervision horizons. We characterize that comparison directly: the second prediction head adds constraints to the same attention problem, changing the leading score coefficients and, under our structural assumptions, the first-head prediction.

## B Shared Gradient Identities

We derive the gradients with respect to $W_{kq}$ and $W_{uv}^{(h)},\ h = 1, 2$ for the two-stage objectives.

**Proposition B.1** (Gradients of the two-stage objectives)**.** *Suppose that Stage 1 is trained on the single-token examples in $\mathcal{D}_{\mathrm{col}}$, with $W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$ fixed. Then, for each head $h \in \{1, 2\}$,*

$$\nabla_{W_{uv}^{(h)}} \mathcal{L}_{\mathrm{col}}^{(h)} = \frac{1}{|\mathcal{D}_{\mathrm{col}}|} \sum_{([e_x], n_1(x), n_2(x)) \in \mathcal{D}_{\mathrm{col}}} \left(p_h(\cdot \mid [e_x]) - e_{n_h(x)}\right) e_x^\top.$$

*In Stage 2, the learned readout matrices $\widehat{W}_{uv}^{(h)}$ are frozen and only $W_{kq}$ is trained. For MTP,*

$$\nabla_{W_{kq}} \mathcal{L}_{\mathrm{MTP}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \sum_{h=1}^{2} X_i X_i^\top \widehat{W}_{uv}^{(h)\top} \left(p_h(\cdot \mid X_i) - e_{y_i^{(h)}}\right) e_{q_i}^\top.$$

*For NTP,*

$$\nabla_{W_{kq}} \mathcal{L}_{\mathrm{NTP}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} X_i X_i^\top \widehat{W}_{uv}^{(1)\top} \left(p_1(\cdot \mid X_i) - e_{y_i^{(1)}}\right) e_{q_i}^\top.$$

*Proof.* For a softmax cross-entropy loss with logit vector $z_h$ and target token $y$, we have

$$\nabla_{z_h} \ell_h = p_h - e_y,$$

where $p_h = \mathrm{softmax}(z_h)$ denotes the corresponding output probability vector.

For Stage 1, consider a single-token example

$$[e_x] \mapsto \big(n_1(x), n_2(x)\big).$$

For head $h$, the target is therefore $n_h(x)$. Since $X = [e_x]$, $e_q = e_x$, and $W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$, we have

$$X X^\top W_{kq}(0) e_q = e_x e_x^\top \mathbf{1}\mathbf{1}^\top e_x = e_x.$$

Hence the Stage 1 logit vector of head $h$ is $z_h = W_{uv}^{(h)} e_x$, and its output distribution is

$$p_h(\cdot \mid [e_x]) = \mathrm{softmax}\Big(W_{uv}^{(h)} e_x\Big).$$

Applying the chain rule gives

$$\nabla_{W_{uv}^{(h)}} \ell_h \left(W_{kq}(0), W_{uv}^{(h)}; [e_x], n_h(x)\right) = \left(p_h(\cdot \mid [e_x]) - e_{n_h(x)}\right) e_x^\top.$$

Averaging over the single-token examples in $\mathcal{D}_{\mathrm{col}}$ yields

$$\nabla_{W_{uv}^{(h)}} \mathcal{L}_{\mathrm{col}}^{(h)} = \frac{1}{|\mathcal{D}_{\mathrm{col}}|} \sum_{([e_x], n_1(x), n_2(x)) \in \mathcal{D}_{\mathrm{col}}} \left( p_h(\cdot \mid [e_x]) - e_{n_h(x)} \right) e_x^\top .$$

For Stage 2, fix a training sample $i$. With the learned readout matrices frozen, the logit vector of head $h$ is

$$z_{i,h} = \widehat{W}_{uv}^{(h)} X_i X_i^\top W_{kq} e_{q_i}.$$

For an infinitesimal perturbation $dW_{kq}$,

$$dz_{i,h} = \widehat{W}_{uv}^{(h)} X_i X_i^\top (dW_{kq}) e_{q_i}.$$

Writing

$$p_h(\cdot \mid X_i) = \mathrm{softmax}(z_{i,h}),$$

we obtain

$$d\ell_h = \left( p_h(\cdot \mid X_i) - e_{y_i^{(h)}} \right)^\top dz_{i,h} = \left( p_h(\cdot \mid X_i) - e_{y_i^{(h)}} \right)^\top \widehat{W}_{uv}^{(h)} X_i X_i^\top (dW_{kq}) e_{q_i}.$$

Therefore,

$$\nabla_{W_{kq}} \ell_h = X_i X_i^\top \widehat{W}_{uv}^{(h)\top} \left( p_h(\cdot \mid X_i) - e_{y_i^{(h)}} \right) e_{q_i}^\top .$$

For MTP, summing this expression over $h \in \{1, 2\}$ and averaging over $i \in \mathcal{D}$ gives

$$\nabla_{W_{kq}} \mathcal{L}_{\mathrm{MTP}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \sum_{h=1}^{2} X_i X_i^\top \widehat{W}_{uv}^{(h)\top} \left( p_h(\cdot \mid X_i) - e_{y_i^{(h)}} \right) e_{q_i}^\top .$$

For NTP, retaining only the first head gives

$$\nabla_{W_{kq}} \mathcal{L}_{\mathrm{NTP}} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} X_i X_i^\top \widehat{W}_{uv}^{(1)\top} \left( p_1(\cdot \mid X_i) - e_{y_i^{(1)}} \right) e_{q_i}^\top .$$

This proves the three gradient identities. □

# C Proofs for MTP Training and Dynamics

We prove the readout structure, attention-score dynamics, and asymptotic allocation in Section 3.

## C.1 Dynamics of Readout Matrices

We characterize the readout columns learned from the Stage 1 single-token examples. Let $\mathcal{V}_{\mathrm{c}}$ denote the set of token identities covered by the Stage 1 single-token examples. For every $x \in \mathcal{V}_{\mathrm{c}}$, let $e_x$ denote its one-hot representation. By the Stage 1 coverage assumption, the collocation set $\mathcal{D}_{\mathrm{col}}$ contains the single-token example

$$[e_x] \mapsto \big(n_1(x), n_2(x)\big).$$

Thus, for each horizon $h \in \{1, 2\}$, the Stage 1 target associated with $x$ is the deterministic collocation token $n_h(x)$.

During Stage 1, we fix $W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$ and train each readout matrix $W_{uv}^{(h)}$ from the zero initialization $W_{uv}^{(h)}(0) = 0$. Since the query of the single-token prefix is $e_q = e_x$ and $e_x$ is one-hot,

$$X X^\top W_{kq}(0) e_q = e_x e_x^\top \mathbf{1}\mathbf{1}^\top e_x = e_x.$$

Hence the prediction distribution of head $h$ reduces to

$$p_h(\cdot \mid [e_x]) = \mathrm{softmax}\Big( W_{uv}^{(h)} e_x \Big),$$

and its per-sample loss is

$$\ell_h\big(W_{uv}^{(h)} \mid [e_x], n_h(x)\big) = -\log\Big[\mathrm{softmax}\Big(W_{uv}^{(h)} e_x\Big)\Big]_{n_h(x)}.$$

Since every covered token type has exactly one single-token example in $\mathcal{D}_{\mathrm{col}}$, the empirical Stage 1 objective is

$$\mathcal{L}_{\mathrm{col}}^{(h)} = \frac{1}{|\mathcal{V}_{\mathrm{c}}|} \sum_{x \in \mathcal{V}_{\mathrm{c}}} \ell_h\big(W_{uv}^{(h)} \mid e_x, n_h(x)\big).$$

We analyze the discrete gradient-descent iterates used in Stage 1. For head $h$, let

$$W_{uv}^{(h)}(k+1) = W_{uv}^{(h)}(k) - \eta_{\mathrm{col}} \nabla \mathcal{L}_{\mathrm{col}}^{(h)}\big(W_{uv}^{(h)}(k)\big),$$

with $W_{uv}^{(h)}(0) = 0$.

**Proposition C.1** (Stage 1 collocation structure and equal readout scales; formal version of Proposition 3.1)**.** *Suppose that Stage 1 is trained by the discrete gradient-descent updates above and stopped after a finite number $K_1 \geq 1$ of iterations. Then, for every prediction head $h \in \{1, 2\}$ and every $x \in \mathcal{V}_{\mathrm{c}}$,*

$$\widehat{W}_{uv}^{(h)} e_x = G_h \left( e_{n_h(x)} - \frac{1}{V}\mathbf{1} \right),$$

*where*

$$\widehat{W}_{uv}^{(h)} := W_{uv}^{(h)}(K_1),$$

*and $G_h > 0$ depends only on the prediction head and the Stage 1 training checkpoint, but not on $x$.*

*Equivalently,*

$$\widehat{W}_{uv}^{(h)} e_x = M_h e_{n_h(x)} - m_h \sum_{a \neq n_h(x)} e_a,$$

*where*

$$M_h = \frac{V-1}{V} G_h, \qquad m_h = \frac{1}{V} G_h.$$

*If the two heads use the same set of single-token examples, learning rate, zero initialization, loss normalization, and checkpoint, then $G_1 = G_2 =: G > 0$, as in Proposition 3.1.*

*Proof of Proposition C.1.* Fix a prediction head $h$ and a token identity $x \in \mathcal{V}_{\mathrm{c}}$, and write $n := n_h(x)$. Since $e_x$ is one-hot, multiplication by $e_x$ extracts the column of the readout matrix indexed by $x$. Define

$$u_x^{(h)}(k) := W_{uv}^{(h)}(k) e_x \in \mathbb{R}^V.$$

For $a \in \mathcal{V}$, let

$$p_{x,a}^{(h)}(k) := \frac{\exp\big(u_{x,a}^{(h)}(k)\big)}{\sum_{b \in \mathcal{V}} \exp\big(u_{x,b}^{(h)}(k)\big)}.$$

The cross-entropy gradient with respect to this column is

$$\nabla_{u_x^{(h)}} \ell_h = p_x^{(h)} - e_n,$$

where $p_x^{(h)} := \big(p_{x,a}^{(h)}\big)_{a \in \mathcal{V}}$.

Because the Stage 1 input is the one-token prefix $[e_x]$, this sample affects only the $x$-th column of $W_{uv}^{(h)}$. Moreover, every covered token type has exactly one example and hence receives the same empirical weight. Therefore the columnwise discrete update is

$$u_x^{(h)}(k+1) = u_x^{(h)}(k) + \widetilde{\eta}_{\mathrm{col}} \left( e_n - p_x^{(h)}(k) \right),$$

where

$$\widetilde{\eta}_{\mathrm{col}} := \frac{\eta_{\mathrm{col}}}{|\mathcal{V}_{\mathrm{c}}|}.$$

We prove by induction that, for every $k \geq 0$, there exists a scalar $g_h^{(k)}$, independent of $x$, such that

$$u_x^{(h)}(k) = g_h^{(k)} \left( e_n - \frac{1}{V}\mathbf{1} \right).$$

At initialization, $u_x^{(h)}(0) = 0$, so the claim holds with $g_h^{(0)} = 0$.

Suppose the claim holds at iteration $k$. The correct output coordinate then satisfies

$$u_{x,n}^{(h)}(k) = \frac{V-1}{V} g_h^{(k)},$$

whereas every incorrect coordinate $a \neq n$ satisfies

$$u_{x,a}^{(h)}(k) = -\frac{1}{V} g_h^{(k)}.$$

Hence

$$p_{x,n}^{(h)}(k) = \frac{e^{g_h^{(k)}}}{e^{g_h^{(k)}} + V - 1},$$

and, for every $a \neq n$,

$$p_{x,a}^{(h)}(k) = \frac{1}{e^{g_h^{(k)}} + V - 1}.$$

It follows that

$$e_n - p_x^{(h)}(k) = \frac{V}{e^{g_h^{(k)}} + V - 1} \left( e_n - \frac{1}{V}\mathbf{1} \right).$$

Substituting this identity into the discrete update gives

$$u_x^{(h)}(k+1) = \left[ g_h^{(k)} + \widetilde{\eta}_{\mathrm{col}} \frac{V}{e^{g_h^{(k)}} + V - 1} \right] \left( e_n - \frac{1}{V}\mathbf{1} \right).$$

Therefore the same structure is preserved at iteration $k+1$, with the scalar recurrence

$$g_h^{(k+1)} = g_h^{(k)} + \widetilde{\eta}_{\mathrm{col}} \frac{V}{e^{g_h^{(k)}} + V - 1}, \qquad g_h^{(0)} = 0.$$

This recurrence does not depend on the identity of $x$ or on the location of its distinguished output $n_h(x)$. Since every column has the same initialization and empirical weight, the same scalar $g_h^{(k)}$ governs every token column.

Moreover,

$$g_h^{(k+1)} - g_h^{(k)} = \widetilde{\eta}_{\mathrm{col}} \frac{V}{e^{g_h^{(k)}} + V - 1} > 0,$$

so

$$g_h^{(k)} > 0 \qquad \text{for every } k \geq 1.$$

At the finite Stage 1 checkpoint $K_1$, define

$$G_h := g_h^{(K_1)} > 0.$$

Then, for every $x \in \mathcal{V}_{\mathrm{c}}$,

$$\widehat{W}_{uv}^{(h)} e_x = G_h \left( e_{n_h(x)} - \frac{1}{V}\mathbf{1} \right).$$

Finally, writing

$$M_h := \frac{V-1}{V} G_h, \qquad m_h := \frac{1}{V} G_h,$$

we obtain

$$\widehat{W}_{uv}^{(h)} e_x = \frac{V-1}{V} G_h e_{n_h(x)} - \frac{G_h}{V} \sum_{a \neq n_h(x)} e_a = M_h e_{n_h(x)} - m_h \sum_{a \neq n_h(x)} e_a.$$

This proves the readout structure. To show that the two heads have equal scales under symmetric training, use the scalar recurrence:

$$g_h^{(k+1)} = g_h^{(k)} + \widetilde{\eta}\frac{V}{e^{g_h^{(k)}} + V - 1}, \qquad g_h^{(0)} = 0,$$

where the effective learning rate $\widetilde{\eta} = \eta_{\mathrm{col}}/|\mathcal{V}_{\mathrm{c}}|$ is identical for the two heads by assumption.

The recurrence does not depend on the identity of the horizon-specific collocation target $n_h(x)$. Since the two heads have the same initial condition and obey the same scalar recurrence,

$$g_1^{(k)} = g_2^{(k)} \qquad \text{for every } k \geq 0.$$

Evaluating at the common Stage 1 checkpoint $K_1$ yields

$$G_1 = g_1^{(K_1)} = g_2^{(K_1)} = G_2.$$

□

### C.2 Training Dynamics of MTP

We derive the Stage 2 score dynamics and max-margin solution for each query token. Unless stated otherwise, asymptotics are as Stage 2 gradient-flow time $t \to \infty$, with the training data, vocabulary, Stage 1 checkpoint, and frozen readouts fixed. Fix a query token $q$ and let

$$\mathcal{D}_q = \{i : q_i = q\}.$$

Let

$$\mathcal{U}_q = \bigcup_{i \in \mathcal{D}_q} \mathcal{X}_i$$

be the set of distinct tokens appearing in the corresponding attention supports, and collect their scores as

$$s_q(t) = (s_{x,q}(t))_{x \in \mathcal{U}_q}.$$

Let

$$\mathcal{F}_q = \{x_i^\star : i \in \mathcal{D}_q\}$$

denote the set of distinct full-cue tokens in this fixed-query system. By the role consistency assumption, if $x \in \mathcal{F}_q$ appears in a sentence with query token $q$, then it is the full cue of that sentence. Since every sentence has a unique full cue,

$$\mathcal{X}_i \cap \mathcal{F}_q = \{x_i^\star\}, \qquad i \in \mathcal{D}_q. \tag{6}$$

By Proposition C.1, the frozen Stage 1 readouts satisfy

$$\widehat{W}_{uv}^{(h)} e_x = G\left(e_{n_h(x)} - \frac{1}{V}\mathbf{1}\right).$$

For sample $i \in \mathcal{D}_q$, define the head-$h$ logit of output class $a$ by

$$z_{i,h}(a,t) := e_a^\top \widehat{W}_{uv}^{(h)} X_i X_i^\top W_{kq}(t) e_q.$$

Since the tokens are one-hot,

$$X_i X_i^\top W_{kq}(t) e_q = \sum_{x \in \mathcal{X}_i} s_{x,q}(t) e_x,$$

and therefore

$$z_{i,h}(a,t) = G \sum_{\substack{x \in \mathcal{X}_i: \\ n_h(x) = a}} s_{x,q}(t) - \frac{G}{V} \sum_{x \in \mathcal{X}_i} s_{x,q}(t). \tag{7}$$

The second term is independent of $a$ and hence cancels under the softmax. Thus

$$p_i^{(h)}(a,t) = \frac{\exp\left(G \sum_{\substack{x\in\mathcal{X}_i:\\ n_h(x)=a}} s_{x,q}(t)\right)}{\sum_{b\in\mathcal{V}} \exp\left(G \sum_{\substack{x\in\mathcal{X}_i:\\ n_h(x)=b}} s_{x,q}(t)\right)}.$$

**Lemma C.2** (MTP score dynamics; formal version of Lemma 3.3)**.** *For $i \in \mathcal{D}_q$, $h \in \{1,2\}$, and $a \neq y_i^{(h)}$, define*

$$[\phi_{i,a}^{(h)}]_x = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[\mathbb{1}\{n_h(x) = y_i^{(h)}\} - \mathbb{1}\{n_h(x) = a\}\right]. \tag{8}$$

*Then*

$$z_{i,h}(y_i^{(h)}, t) - z_{i,h}(a,t) = \left\langle s_q(t), \phi_{i,a}^{(h)} \right\rangle, \tag{9}$$

*and the Stage 2 score dynamics satisfy*

$$\dot{s}_q(t) = \eta \sum_{i\in\mathcal{D}_q} \sum_{h=1}^{2} \sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)\phi_{i,a}^{(h)}. \tag{10}$$

*Moreover, for the inactive second-head class $\bar{a}_i^{(2)}$,*

$$\phi_{i,\bar{a}_i^{(2)}}^{(2)} = Ge_{x_i^\star}. \tag{11}$$

*Proof.* Equation (9) follows from Equation (7).

It remains to derive the score dynamics. Let

$$\ell_{i,h}(t) = -\log p_i^{(h)}(y_i^{(h)}, t)$$

be the cross-entropy loss of head $h$ on sample $i$. From Equation (7),

$$\frac{\partial z_{i,h}(a,t)}{\partial s_{x,q}(t)} = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[\mathbb{1}\{n_h(x) = a\} - \frac{1}{V}\right].$$

Using

$$\frac{\partial \ell_{i,h}}{\partial z_{i,h}(a)} = p_i^{(h)}(a,t) - \mathbb{1}\{a = y_i^{(h)}\},$$

and

$$\sum_{a\in\mathcal{V}}\left[p_i^{(h)}(a,t) - \mathbb{1}\{a = y_i^{(h)}\}\right] = 0,$$

the class-independent term cancels, giving

$$\frac{\partial \ell_{i,h}}{\partial s_{x,q}} = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[p_i^{(h)}(n_h(x),t) - \mathbb{1}\{n_h(x) = y_i^{(h)}\}\right]. \tag{12}$$

The shared attention matrix follows gradient flow,

$$\dot{W}_{kq}(t) = -\eta \sum_i \sum_{h=1}^{2} \nabla_{W_{kq}} \ell_{i,h}(t).$$

Since

$$s_{u,q_i} = e_u^\top W_{kq} e_{q_i}, \qquad \nabla_{W_{kq}} s_{u,q_i} = e_u e_{q_i}^\top,$$

we have

$$\nabla_{W_{kq}} \ell_{i,h} = \sum_{u \in \mathcal{X}_i} \frac{\partial \ell_{i,h}}{\partial s_{u,q_i}} e_u e_{q_i}^\top.$$

Therefore, for a fixed token-pair coordinate $(x, q)$,

$$\dot{s}_{x,q}(t) = e_x^\top \dot{W}_{kq}(t) e_q = -\eta \sum_i \sum_{h=1}^{2} \sum_{u \in \mathcal{X}_i} \frac{\partial \ell_{i,h}}{\partial s_{u,q_i}} (e_x^\top e_u)(e_{q_i}^\top e_q). \tag{13}$$

Because all tokens are one-hot,

$$e_x^\top e_u = \mathbb{1}\{x = u\}, \qquad e_{q_i}^\top e_q = \mathbb{1}\{q_i = q\}.$$

Substituting Equation (12) into Equation (13) thus yields

$$\dot{s}_{x,q}(t) = \eta G \sum_{i \in \mathcal{D}_q} \mathbb{1}\{x \in \mathcal{X}_i\} \sum_{h=1}^{2} \left[ \mathbb{1}\{n_h(x) = y_i^{(h)}\} - p_i^{(h)}(n_h(x), t) \right]. \tag{14}$$

Finally, for every $i$ and $h$,

$$\sum_{a \neq y_i^{(h)}} p_i^{(h)}(a, t) [\phi_{i,a}^{(h)}]_x = G \mathbb{1}\{x \in \mathcal{X}_i\} \left[ \mathbb{1}\{n_h(x) = y_i^{(h)}\} - p_i^{(h)}(n_h(x), t) \right].$$

Hence Equation (14) is exactly the vector dynamics Equation (10).

By prefix consistency and uniqueness of the full cue, $x_i^\star$ is the only token in $\mathcal{X}_i$ satisfying $n_2(x) = y_i^{(2)}$. Since the inactive class $\bar{a}_i^{(2)}$ is not equal to $n_2(x)$ for any $x \in \mathcal{X}_i$, the definition Equation (8) gives

$$\phi_{i,\bar{a}_i^{(2)}}^{(2)} = G e_{x_i^\star},$$

which proves Equation (11). □

**Proposition C.3** (MTP max-margin structure; formal version of Proposition 3.4)**.** *For a fixed query token $q$, consider the hard-margin problem*

$$\min_{c \in \mathbb{R}^{|\mathcal{U}_q|}} \frac{1}{2} \|c\|^2 \quad \textit{subject to} \quad \left\langle c, \phi_{i,a}^{(h)} \right\rangle \geq 1 \tag{15}$$

*for every $i \in \mathcal{D}_q$, $h \in \{1, 2\}$, and $a \neq y_i^{(h)}$. Define the indicator vector of the full-cue set by $e_{\mathcal{F}_q} := \sum_{x \in \mathcal{F}_q} e_x$. Its unique solution is*

$$c_q^\star = \frac{1}{G} e_{\mathcal{F}_q}.$$

*Proof.* The feasible set of Equation (15) is an intersection of linear half-spaces and is therefore convex. Since the objective $\frac{1}{2}\|c\|^2$ is strictly convex, the problem has at most one minimizer.

We first show that

$$c_q^\star = \frac{1}{G} e_{\mathcal{F}_q}$$

is feasible. For every $i \in \mathcal{D}_q$, Equation (6) gives $\mathcal{X}_i \cap \mathcal{F}_q = \{x_i^\star\}$. Since $x_i^\star$ predicts the correct target at both horizons, the definition of $\phi_{i,a}^{(h)}$ yields

$$\left\langle \frac{1}{G} e_{\mathcal{F}_q}, \phi_{i,a}^{(h)} \right\rangle = \frac{1}{G} [\phi_{i,a}^{(h)}]_{x_i^\star} = \frac{1}{G} G \left[ \mathbb{1}\{n_h(x_i^\star) = y_i^{(h)}\} - \mathbb{1}\{n_h(x_i^\star) = a\} \right] = 1$$

for every $h \in \{1, 2\}$ and $a \neq y_i^{(h)}$. Hence $c_q^\star$ is feasible.

Now let $c$ be any feasible vector. For each $x \in \mathcal{F}_q$, choose $i \in \mathcal{D}_q$ such that $x_i^\star = x$. By Equation (11),

$$\phi_{i,\bar{a}_i^{(2)}}^{(2)} = G e_x.$$

The corresponding margin constraint therefore implies

$$1 \le \left\langle c, \phi^{(2)}_{i,\bar{a}_i^{(2)}} \right\rangle = G c_x,$$

and hence

$$c_x \ge \frac{1}{G}, \qquad x \in \mathcal{F}_q.$$

It follows that every feasible $c$ satisfies

$$\frac{1}{2}\|c\|^2 \ge \frac{1}{2} \sum_{x \in \mathcal{F}_q} c_x^2 \ge \frac{|\mathcal{F}_q|}{2G^2}.$$

On the other hand,

$$\frac{1}{2} \left\| \frac{1}{G} e_{\mathcal{F}_q} \right\|^2 = \frac{|\mathcal{F}_q|}{2G^2}.$$

Thus $c_q^\star$ attains the minimum of Equation (15). Since the objective is strictly convex over a convex feasible set, this minimizer is unique. □

**Lemma C.4** (Fixed-query loss comparison)**.** *For a fixed query $q$ and $\mathcal{H} \in \{\{1,2\},\{1\}\}$, define*

$$\mathcal{L}_{q,\mathcal{H}}(s) = \sum_{i \in \mathcal{D}_q} \sum_{h \in \mathcal{H}} -\log p_i^{(h)}(y_i^{(h)}; s), \qquad s \in \mathbb{R}^{|\mathcal{U}_q|}, \tag{16}$$

*where $p_i^{(h)}(a; s)$ is obtained by replacing $s_q(t)$ with $s$ in Equation* (7) *and applying softmax. Write*

$$\mathcal{L}_q := \mathcal{L}_{q,\{1,2\}}, \qquad \mathcal{L}_q^{\mathrm{NTP}} := \mathcal{L}_{q,\{1\}}.$$

*Both losses are convex in $s$. Moreover, for either pair*

$$(\mathcal{L}, s(t)) \in \left\{ (\mathcal{L}_q, s_q(t)), (\mathcal{L}_q^{\mathrm{NTP}}, s_q^{\mathrm{NTP}}(t)) \right\},$$

*if $\dot{s}(t) = -\eta \nabla \mathcal{L}(s(t))$, then every fixed $\widetilde{s} \in \mathbb{R}^{|\mathcal{U}_q|}$ and $t > 0$ satisfy*

$$\mathcal{L}(s(t)) - \mathcal{L}(\widetilde{s}) \le \frac{\|s(0) - \widetilde{s}\|^2}{2\eta t}. \tag{17}$$

*Proof.* By Equation (7), each logit is linear in $s$. Each summand in Equation (16) is therefore a log-sum-exp of linear functions minus a linear function, and hence is convex in $s$. Summation preserves convexity.

For a fixed $\widetilde{s}$, convexity and the gradient-flow equation give

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\|s(t) - \widetilde{s}\|^2 &= -\eta \left\langle \nabla \mathcal{L}(s(t)), s(t) - \widetilde{s} \right\rangle \\ &\le -\eta \left[ \mathcal{L}(s(t)) - \mathcal{L}(\widetilde{s}) \right]. \end{aligned}$$

The loss is nonincreasing because $\frac{d}{dt}\mathcal{L}(s(t)) = -\eta \|\nabla \mathcal{L}(s(t))\|^2 \le 0$. Integrating the preceding inequality from 0 to $t$ and using $\mathcal{L}(s(\tau)) \ge \mathcal{L}(s(t))$ for $0 \le \tau \le t$ proves Equation (17). □

**Lemma C.5** (Normalized concentration of MTP scores)**.** *Define the average full-cue score*

$$\bar{s}_q(t) = \frac{1}{|\mathcal{F}_q|} \sum_{x \in \mathcal{F}_q} s_{x,q}(t).$$

*As $t \to \infty$,*

$$\left\| \frac{s_q(t)}{\bar{s}_q(t)} - e_{\mathcal{F}_q} \right\| = O\left( \frac{1}{\sqrt{\bar{s}_q(t)}} \right). \tag{18}$$

*Equivalently,*

$$s_q(t) = G \bar{s}_q(t) c_q^\star + O\left( \sqrt{\bar{s}_q(t)} \right), \tag{19}$$

*where $c_q^\star$ is the solution of Equation* (15)*.*

*Proof.* For any $R > 0$, consider the fixed score vector $Re_{\mathcal{F}_q}$. Then every correct-to-competing margin under this vector equals $GR$, and therefore

$$\mathcal{L}_q\left(Re_{\mathcal{F}_q}\right) = 2|\mathcal{D}_q| \log\left(1 + (V-1)e^{-GR}\right) \longrightarrow 0$$

as $R \to \infty$.

By Lemma C.2, the fixed-query score vector follows the gradient flow of $\mathcal{L}_q$. Applying Lemma C.4 with $\widetilde{s} = Re_{\mathcal{F}_q}$ gives

$$\mathcal{L}_q(s_q(t)) - \mathcal{L}_q\left(Re_{\mathcal{F}_q}\right) \le \frac{\left\|s_q(0) - Re_{\mathcal{F}_q}\right\|^2}{2\eta t}.$$

Sending first $t \to \infty$ and then $R \to \infty$ yields

$$\mathcal{L}_q(s_q(t)) \longrightarrow 0. \tag{20}$$

For each $x \in \mathcal{F}_q$, choose a sentence whose full cue is $x$. Its inactive second-head margin is $Gs_{x,q}(t)$. Hence Equation (20) implies

$$s_{x,q}(t) \longrightarrow \infty, \qquad x \in \mathcal{F}_q,$$

and therefore

$$\bar{s}_q(t) \longrightarrow \infty.$$

For sufficiently large $t$, define

$$r_q(t) = \frac{s_q(t)}{\bar{s}_q(t)}$$

and

$$\gamma_{i,a}^{(h)}(t) = \frac{1}{G}\left\langle r_q(t), \phi_{i,a}^{(h)} \right\rangle.$$

By Equation (6),

$$\left\langle e_{\mathcal{F}_q}, \phi_{i,a}^{(h)} \right\rangle = G \tag{21}$$

for every $i \in \mathcal{D}_q$, $h \in \{1,2\}$, and $a \neq y_i^{(h)}$.

Moreover,

$$\frac{1}{|\mathcal{F}_q|} \sum_{x \in \mathcal{F}_q} [r_q(t)]_x = 1.$$

For each $x \in \mathcal{F}_q$ there is an inactive second-head constraint whose normalized margin is exactly $[r_q(t)]_x$. Consequently,

$$\min_{\substack{i \in \mathcal{D}_q,\ h \in \{1,2\},\\ a \neq y_i^{(h)}}} \gamma_{i,a}^{(h)}(t) \le 1. \tag{22}$$

By the softmax ratio,

$$p_i^{(h)}(a,t) = p_i^{(h)}(y_i^{(h)},t) \exp\left(-G\bar{s}_q(t)\gamma_{i,a}^{(h)}(t)\right). \tag{23}$$

Equation (20) and the finiteness of the fixed-query training system imply

$$p_i^{(h)}(y_i^{(h)},t) \longrightarrow 1$$

uniformly over $i \in \mathcal{D}_q$ and $h \in \{1,2\}$. Hence, for all sufficiently large $t$, there exists a constant $P_q^{\mathrm{MTP}} > 0$ such that

$$p_i^{(h)}(y_i^{(h)},t) \ge P_q^{\mathrm{MTP}}$$

for every $i$ and $h$.

Moreover, by Equation (23), for every $a \neq y_i^{(h)}$,

$$\frac{p_i^{(h)}(a,t)}{p_i^{(h)}(y_i^{(h)},t)} = \exp\left(-G\bar{s}_q(t)\gamma_{i,a}^{(h)}(t)\right).$$

Since the left-hand side converges to zero,

$$G\bar{s}_q(t)\gamma_{i,a}^{(h)}(t) \longrightarrow +\infty.$$

Therefore all normalized margins $\gamma_{i,a}^{(h)}(t)$ are positive for sufficiently large $t$.

By Equation (22), there exists an inactive second-head constraint with normalized margin $\gamma_0(t) \le 1$. For sufficiently large $t$, $0 < \gamma_0(t) \le 1$, and hence

$$\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t) \ge P_q^{\mathrm{MTP}} \exp\left(-G\bar{s}_q(t)\gamma_0(t)\right) \ge P_q^{\mathrm{MTP}} e^{-G\bar{s}_q(t)}.$$

We next split the weighted sum according to whether $0 < \gamma_{i,a}^{(h)} \le 1$ or $\gamma_{i,a}^{(h)} > 1$. The former terms satisfy $p_i^{(h)}(a,t)\gamma_{i,a}^{(h)}(t) \le p_i^{(h)}(a,t)$. For a term with $\gamma_{i,a}^{(h)} = 1 + d$, $d > 0$,

$$p_i^{(h)}(a,t)d = p_i^{(h)}(y_i^{(h)},t)e^{-G\bar{s}_q(t)}de^{-G\bar{s}_q(t)d} \le \frac{e^{-G\bar{s}_q(t)}}{eG\bar{s}_q(t)}.$$

Since the fixed-query training system contains finitely many constraints, there exists a constant $J_q^{\mathrm{MTP}} > 0$ such that

$$\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)\gamma_{i,a}^{(h)}(t) \le \sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t) + \frac{J_q^{\mathrm{MTP}} e^{-G\bar{s}_q(t)}}{\bar{s}_q(t)}.$$

Combining the last two displays and defining $K_q^{\mathrm{MTP}} := J_q^{\mathrm{MTP}}/P_q^{\mathrm{MTP}}$ yields

$$\frac{\displaystyle\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)\gamma_{i,a}^{(h)}(t)}{\displaystyle\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)} \le 1 + \frac{K_q^{\mathrm{MTP}}}{\bar{s}_q(t)}. \tag{24}$$

On the other hand, Equation (21) and Equation (10) give

$$\dot{\bar{s}}_q(t) = \frac{\eta G}{|\mathcal{F}_q|}\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t). \tag{25}$$

Set

$$\mathcal{V}_q(t) = \frac{1}{2}\left\|r_q(t) - e_{\mathcal{F}_q}\right\|^2.$$

Since

$$\left\langle r_q(t) - e_{\mathcal{F}_q}, e_{\mathcal{F}_q}\right\rangle = 0,$$

differentiating $r_q = s_q/\bar{s}_q$ with respect to $\log \bar{s}_q(t)$ gives

$$\frac{dr_q}{d\log\bar{s}_q} = \frac{\dot{s}_q}{\dot{\bar{s}}_q} - r_q.$$

Consequently,

$$\frac{d\mathcal{V}_q}{d\log\bar{s}_q} = \left\langle r_q - e_{\mathcal{F}_q}, \frac{\dot{s}_q}{\dot{\bar{s}}_q}\right\rangle - \left\langle r_q - e_{\mathcal{F}_q}, r_q\right\rangle.$$

The orthogonality above implies $\langle r_q - e_{\mathcal{F}_q}, r_q\rangle = 2\mathcal{V}_q$. Using Equations (10) and (25) in the first inner product therefore yields

$$\frac{d\mathcal{V}_q}{d\log\bar{s}_q} = |\mathcal{F}_q|\left[\frac{\displaystyle\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)\gamma_{i,a}^{(h)}(t)}{\displaystyle\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)} - 1\right] - 2\mathcal{V}_q.$$

Applying Equation (24) and defining $B_q^{\mathrm{MTP}} := |\mathcal{F}_q| K_q^{\mathrm{MTP}} > 0$ gives

$$\frac{d\mathcal{V}_q}{d\log \bar{s}_q} \le -2\mathcal{V}_q + \frac{B_q^{\mathrm{MTP}}}{\bar{s}_q}.$$

By Equation (25), $\dot{\bar{s}}_q(t) > 0$. We may therefore regard $\mathcal{V}_q$ as a function of $\bar{s}_q$ for all sufficiently large $t$. Since $d/d\log \bar{s}_q = \bar{s}_q\, d/d\bar{s}_q$, the preceding inequality is equivalent to

$$\frac{d\mathcal{V}_q}{d\bar{s}_q} + \frac{2}{\bar{s}_q}\mathcal{V}_q \le \frac{B_q^{\mathrm{MTP}}}{\bar{s}_q^2}.$$

Multiplication by the integrating factor $\bar{s}_q^2$ yields

$$\frac{d}{d\bar{s}_q}\left(\bar{s}_q^2 \mathcal{V}_q\right) \le B_q^{\mathrm{MTP}}.$$

Fix a sufficiently large $t_0$ and write $s_0 = \bar{s}_q(t_0) > 0$. Integrating from $s_0$ to $\bar{s}_q(t)$ gives

$$\bar{s}_q(t)^2 \mathcal{V}_q(t) \le s_0^2 \mathcal{V}_q(t_0) + B_q^{\mathrm{MTP}}\big(\bar{s}_q(t) - s_0\big).$$

Because $\bar{s}_q(t) \to \infty$, division by $\bar{s}_q(t)^2$ gives

$$\mathcal{V}_q(t) = O\left(\frac{1}{\bar{s}_q(t)}\right).$$

This proves Equation (18).

Equation (19) follows from Proposition C.3. □

**Theorem C.6** (MTP score concentration; formal version of Theorem 3.5)**.** *For every fixed query token $q$, as $t \to \infty$,*

$$\frac{s_q(t)}{\log t} \longrightarrow c_q^\star = \frac{1}{G} e_{\mathcal{F}_q}. \tag{26}$$

*More precisely,*

$$s_{x,q}(t) = \frac{1}{G}\log t + O\left(\sqrt{\log t}\right), \qquad x \in \mathcal{F}_q, \tag{27}$$

*whereas*

$$s_{x,q}(t) = O\left(\sqrt{\log t}\right) = o(\log t), \qquad x \in \mathcal{U}_q \setminus \mathcal{F}_q. \tag{28}$$

*Consequently, for every training sentence $i$ and every non-full token $x \in \mathcal{X}_i \setminus \{x_i^\star\}$,*

$$s_{x_i^\star, q_i}(t) - s_{x,q_i}(t) = \frac{1}{G}\log t + O\left(\sqrt{\log t}\right). \tag{29}$$

*Proof.* By Lemma C.5,

$$s_q(t) = \bar{s}_q(t) e_{\mathcal{F}_q} + O\left(\sqrt{\bar{s}_q(t)}\right).$$

Together with Equation (21), this implies, uniformly over all training margins,

$$\left\langle s_q(t), \phi_{i,a}^{(h)} \right\rangle = G\bar{s}_q(t) + O\left(\sqrt{\bar{s}_q(t)}\right).$$

The softmax ratios therefore give constants $\kappa_q^-, \kappa_q^+, \beta_q > 0$ such that

$$\kappa_q^- e^{-G\bar{s}_q(t) - \beta_q\sqrt{\bar{s}_q(t)}} \le \sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)$$

and

$$\sum_{i\in\mathcal{D}_q}\sum_{h=1}^{2}\sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t) \le \kappa_q^+ e^{-G\bar{s}_q(t) + \beta_q\sqrt{\bar{s}_q(t)}}$$

for all sufficiently large $t$.

Combining these estimates with Equation (25), there exist constants $c_1, c_2, C > 0$ and $t_0$ such that, for $t \geq t_0$,

$$c_1 e^{-G\bar{s}_q(t)-C\sqrt{\bar{s}_q(t)}} \leq \dot{\bar{s}}_q(t) \leq c_2 e^{-G\bar{s}_q(t)+C\sqrt{\bar{s}_q(t)}}.$$

In particular, $\dot{\bar{s}}_q(t) > 0$, so we may use $u = \bar{s}_q(t)$ as the time variable. Writing $u_0 = \bar{s}_q(t_0)$ and letting $t(u)$ denote its inverse, the preceding bounds give

$$\frac{1}{c_2} e^{Gu-C\sqrt{u}} \leq \frac{dt}{du} \leq \frac{1}{c_1} e^{Gu+C\sqrt{u}}.$$

For all sufficiently large $u$, both exponents $Gu - C\sqrt{u}$ and $Gu + C\sqrt{u}$ are increasing. Integrating the lower bound over $[u-1, u]$ and the upper bound over $[u_0, u]$ therefore yields

$$c_3 e^{Gu-C\sqrt{u}} \leq t(u) \leq c_4 u e^{Gu+C\sqrt{u}}$$

for constants $c_3, c_4 > 0$. Taking logarithms and using $\log u = O(\sqrt{u})$ gives

$$\log t = G\bar{s}_q(t) + O\left(\sqrt{\bar{s}_q(t)}\right).$$

This relation first implies $\bar{s}_q(t) = O(\log t)$: for sufficiently large $\bar{s}_q(t)$, its square-root error is at most $(G/2)\bar{s}_q(t)$. Substituting this bound back into the error term then gives

$$\bar{s}_q(t) = \frac{1}{G}\log t + O\left(\sqrt{\log t}\right). \tag{30}$$

Substituting Equation (30) into the normalized concentration estimate proves Equations (26) to (28).

Equation (29) then follows. □

### C.3 Role of the Structural Assumptions

The following remarks show how the conclusions change when a structural condition fails. Both reuse the margin identity of Lemma C.2; neither requires new asymptotic analysis.

**Remark C.7** (Competing heads without a full cue)**.** Prefix consistency and the unique-full-cue condition ensure that one score vector satisfies the margin constraints of both heads. Without them, the heads can require opposite orderings of the same scores. Consider a single training prefix $X = [e_{x^c}, e_{x^f}, e_q]$ with targets $(A, B)$ and associations

$$n(x^c) = (A, R_2), \qquad n(x^f) = (R_1, B), \qquad n(q) = (Q_1, Q_2),$$

where $A, R_1, Q_1$ are distinct, $B, R_2, Q_2$ are distinct, and $V > 3$. Here $x^c$ is a near cue, $x^f$ is a far-only cue, and $q$ is uninformative. Only the unique-full-cue and prefix-consistency conditions of Assumption 1 fail.

Under NTP, $e_{x^c}/G$ gives every first-head margin the value one. Applying Lemma C.4 with $\widetilde{s} = Re_{x^c}/G$ and letting $R \to \infty$, as in the proof of Theorem D.3, shows that the NTP loss tends to zero; hence $p_1^{\mathrm{NTP}}(A \mid X) \to 1$.

Under MTP, the first-head margin against $R_1$ is $G(s_{x^c,q} - s_{x^f,q})$, whereas the second-head margin against $R_2$ is $G(s_{x^f,q} - s_{x^c,q})$. Exchanging $s_{x^c,q}$ and $s_{x^f,q}$ exchanges the two head losses, so the MTP objective is invariant under this exchange. Both scores start at one, so uniqueness of the gradient-flow solution gives $s_{x^c,q}(t) = s_{x^f,q}(t)$ for all $t$. Hence

$$p_1^{\mathrm{MTP}}(A \mid X) = \frac{e^{Gs_{x^c,q}(t)}}{2e^{Gs_{x^c,q}(t)} + e^{Gs_{q,q}(t)} + V - 3} < \frac{1}{2} \qquad \text{for all } t \geq 0.$$

When no cue supports both targets, second-token supervision competes with next-token supervision for the same attention scores, and MTP can predict the next token worse than NTP, even on its training prefix.

**Remark C.8** (Unequal readout gaps)**.** Proposition C.3 uses the common gap $G_1 = G_2 = G$ from Proposition C.1. Suppose instead that the frozen readouts have gaps $G_1$ and $G_2$, so that $\phi_{i,a}^{(h)}$ carries the factor $G_h$. If $G_1 \geq G_2$, the proof of Proposition C.3 still applies: the second-head inactive-class constraints force $c_x \geq 1/G_2$ for every $x \in \mathcal{F}_q$, and $e_{\mathcal{F}_q}/G_2$ is feasible because its first-head margins equal $G_1/G_2 \geq 1$. The hard-margin solution is therefore still supported only on full cues.

Thus, the full-cue-only conclusion allows unequal gaps whenever $G_1 \geq G_2$. When $G_1 < G_2$, near cues can enter the solution, although the following construction requires an additional separation among the second-head associations. Consider a single prefix $[e_{x^\star}, e_{x^c}, e_q]$ with full cue $x^\star$, near cue $x^c$, and uninformative query $q$. Let the two targets be $(A, B)$ and take five distinct output tokens $A, B, C, D, E$ such that

$$\big(n_1(x^\star), n_2(x^\star)\big) = (A, B), \qquad \big(n_1(x^c), n_2(x^c)\big) = (A, C), \qquad \big(n_1(q), n_2(q)\big) = (D, E).$$

In particular, $n_2(q) = E \neq C = n_2(x^c)$. The hard-margin solution then has coordinates

$$\big(c_{x^\star}, c_{x^c}, c_q\big) = \Big(\tfrac{1}{2}\big(G_1^{-1} + G_2^{-1}\big),\ \tfrac{1}{2}\big(G_1^{-1} - G_2^{-1}\big),\ 0\Big).$$

The first-head inactive-class constraint and the second-head constraint against $C = n_2(x^c)$ are tight. Their normal vectors are $G_1(1, 1, 0)$ and $G_2(1, -1, 0)$, respectively, and multipliers $1/(2G_1^2)$ and $1/(2G_2^2)$ recover the displayed solution. All remaining constraints are feasible; those involving the query may be assigned zero multipliers. Hence the KKT conditions certify optimality. The full cue has the larger coefficient, while the near cue has a strictly positive coefficient. This construction shows that, once $G_1 < G_2$, MTP need not select only full cues. The matched Stage 1 schedules, for which $G_1 = G_2$, lie at the boundary between this regime and the full-cue-only guarantee above.

# D Proofs for NTP Score Dynamics and Allocation

We prove the NTP score dynamics, max-margin allocation, and score asymptotics in Section 4.

Fix a query token $q$ and use the fixed-query notation introduced in the MTP analysis. Under NTP, only the first-horizon margin vectors are present:

$$[\phi_{i,a}^{(1)}]_x = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[\mathbb{1}\{n_1(x) = y_i^{(1)}\} - \mathbb{1}\{n_1(x) = a\}\right],$$

for $i \in \mathcal{D}_q$ and $a \neq y_i^{(1)}$. The corresponding correct-to-competing logit margin is

$$z_{i,1}^{\mathrm{NTP}}(y_i^{(1)}, t) - z_{i,1}^{\mathrm{NTP}}(a, t) = \left\langle s_q^{\mathrm{NTP}}(t), \phi_{i,a}^{(1)} \right\rangle. \tag{31}$$

**Lemma D.1** (Coupled NTP dynamics)**.** *For every query token $q$, the Stage 2 gradient flow satisfies*

$$\dot{s}_q^{\mathrm{NTP}}(t) = \eta \sum_{i \in \mathcal{D}_q} \sum_{a \neq y_i^{(1)}} p_{i,1}^{\mathrm{NTP}}(a, t)\phi_{i,a}^{(1)}. \tag{32}$$

*Proof.* Since NTP uses the same frozen first prediction head as MTP, Proposition C.1 gives

$$\widehat{W}_{uv}^{(1)} e_x = G\left(e_{n_1(x)} - \frac{1}{V}\mathbf{1}\right).$$

Hence, for $i \in \mathcal{D}_q$,

$$z_{i,1}^{\mathrm{NTP}}(a, t) = G \sum_{\substack{x \in \mathcal{X}_i: \\ n_1(x) = a}} s_{x,q}^{\mathrm{NTP}}(t) - \frac{G}{V} \sum_{x \in \mathcal{X}_i} s_{x,q}^{\mathrm{NTP}}(t). \tag{33}$$

Subtracting the competing logit from the correct logit and using the definition of $\phi_{i,a}^{(1)}$ yields Equation (31).

Let

$$\ell_{i,1}^{\mathrm{NTP}}(t) = -\log p_{i,1}^{\mathrm{NTP}}\left(y_i^{(1)}, t\right).$$

From Equation (33),

$$\frac{\partial z_{i,1}^{\mathrm{NTP}}(a,t)}{\partial s_{x,q}^{\mathrm{NTP}}(t)} = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[\mathbb{1}\{n_1(x) = a\} - \frac{1}{V}\right].$$

Together with

$$\frac{\partial \ell_{i,1}^{\mathrm{NTP}}}{\partial z_{i,1}^{\mathrm{NTP}}(a)} = p_{i,1}^{\mathrm{NTP}}(a,t) - \mathbb{1}\{a = y_i^{(1)}\},$$

the chain rule gives

$$\frac{\partial \ell_{i,1}^{\mathrm{NTP}}}{\partial s_{x,q}^{\mathrm{NTP}}} = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[p_{i,1}^{\mathrm{NTP}}\left(n_1(x),t\right) - \mathbb{1}\{n_1(x) = y_i^{(1)}\}\right], \tag{34}$$

where the class-independent terms cancel because

$$\sum_{a\in\mathcal{V}}\left[p_{i,1}^{\mathrm{NTP}}(a,t) - \mathbb{1}\{a = y_i^{(1)}\}\right] = 0.$$

The Stage 2 gradient flow of the shared attention matrix is

$$\dot{W}_{kq}^{\mathrm{NTP}}(t) = -\eta\sum_i \nabla_{W_{kq}}\ell_{i,1}^{\mathrm{NTP}}(t).$$

Since

$$\nabla_{W_{kq}} s_{u,q_i} = e_u e_{q_i}^\top,$$

for any $x \in \mathcal{U}_q$,

$$\dot{s}_{x,q}^{\mathrm{NTP}}(t) = e_x^\top \dot{W}_{kq}^{\mathrm{NTP}}(t) e_q = -\eta\sum_i\sum_{u\in\mathcal{X}_i}\frac{\partial \ell_{i,1}^{\mathrm{NTP}}}{\partial s_{u,q_i}^{\mathrm{NTP}}}(e_x^\top e_u)(e_{q_i}^\top e_q). \tag{35}$$

Because the token representations are one-hot,

$$e_x^\top e_u = \mathbb{1}\{x = u\}, \qquad e_{q_i}^\top e_q = \mathbb{1}\{q_i = q\}.$$

Substituting Equation (34) into Equation (35) therefore gives

$$\dot{s}_{x,q}^{\mathrm{NTP}}(t) = \eta G\sum_{i\in\mathcal{D}_q}\mathbb{1}\{x \in \mathcal{X}_i\}\left[\mathbb{1}\{n_1(x) = y_i^{(1)}\} - p_{i,1}^{\mathrm{NTP}}\left(n_1(x),t\right)\right]. \tag{36}$$

Finally, from the definition of $\phi_{i,a}^{(1)}$,

$$\sum_{a\neq y_i^{(1)}} p_{i,1}^{\mathrm{NTP}}(a,t)[\phi_{i,a}^{(1)}]_x = G\mathbb{1}\{x \in \mathcal{X}_i\}\left[\mathbb{1}\{n_1(x) = y_i^{(1)}\} - p_{i,1}^{\mathrm{NTP}}\left(n_1(x),t\right)\right].$$

Thus Equation (36) is equivalent, coordinate by coordinate, to Equation (32). □

Consider the hard-margin problem generated by the NTP margin vectors,

$$\min_{d\in\mathbb{R}^{|\mathcal{U}_q|}}\frac{1}{2}\|d\|^2 \quad \text{subject to} \quad \left\langle d, \phi_{i,a}^{(1)}\right\rangle \geq 1 \tag{37}$$

for every $i \in \mathcal{D}_q$ and $a \neq y_i^{(1)}$.

To characterize this problem, define the binary incidence matrix

$$[A]_{i,x} = \mathbb{1}\{x \in \mathcal{X}_i\}\mathbb{1}\{n_1(x) = y_i^{(1)}\}, \qquad i \in \mathcal{D}_q, \quad x \in \mathcal{U}_q.$$

Thus, each row of $A$ records the next-token-compatible cues in one training sentence.

**Lemma D.2** (Equivalent NTP max-margin problem)**.** *The hard-margin problem Equation* (37) *and*

$$\min_{d\in\mathbb{R}^{|\mathcal{U}_q|}}\frac{1}{2}\|d\|^2 \quad \textit{subject to} \quad GAd \succeq \mathbf{1} \tag{38}$$

*have the same unique minimizer, denoted by* $d_q$.

*Proof.* We first characterize the feasible set of Equation (38). For each sentence $i$, let $\bar{a}_i^{(1)} \neq y_i^{(1)}$ be an inactive first-head output, so that

$$n_1(x) \neq \bar{a}_i^{(1)} \qquad \forall x \in \mathcal{X}_i.$$

By the definition of $\phi_{i,a}^{(1)}$,

$$\begin{aligned}\left\langle d, \phi_{i,\bar{a}_i^{(1)}}^{(1)} \right\rangle &= G \sum_{x \in \mathcal{U}_q} d_{x,q} \mathbb{1}\{x \in \mathcal{X}_i\} \left[ \mathbb{1}\{n_1(x) = y_i^{(1)}\} - \mathbb{1}\{n_1(x) = \bar{a}_i^{(1)}\} \right] \\ &= G \sum_{x \in \mathcal{U}_q} d_{x,q} \mathbb{1}\{x \in \mathcal{X}_i\} \mathbb{1}\{n_1(x) = y_i^{(1)}\} \\ &= G[Ad]_i.\end{aligned}$$

Hence every vector feasible for Equation (37) is feasible for Equation (38).

Next, we show that Equation (38) has a unique minimizer. The feasible set

$$\{d : GAd \succeq \mathbf{1}\}$$

is closed and convex. It is also nonempty: since every sentence contains a full cue, $\widetilde{d} = \frac{1}{G} e_{\mathcal{F}_q}$ satisfies $G[A\widetilde{d}]_i = 1$ for every $i \in \mathcal{D}_q$. Since $\frac{1}{2}\|d\|^2$ is coercive, a minimizer exists; since it is strictly convex on the convex feasible set, this minimizer is unique. Denote it by $d_q$.

It remains to show that $d_q$ is also feasible for Equation (37). If the column of $A$ corresponding to a token $x$ is zero, then replacing $d_{x,q}$ by zero leaves every constraint in Equation (38) unchanged. By uniqueness and minimum norm,

$$d_{x,q} = 0 \qquad \text{whenever the } x\text{-column of } A \text{ is zero.} \tag{39}$$

Fix $i \in \mathcal{D}_q$ and $a \neq y_i^{(1)}$. If $x \in \mathcal{X}_i$ satisfies $n_1(x) = a$, then $x$ is first-horizon incorrect in sample $i$. By role consistency, it cannot be first-horizon correct in any other sample with query token $q$. Its column in $A$ is therefore zero, and Equation (39) gives $d_{x,q} = 0$. Consequently,

$$\left\langle d_q, \phi_{i,a}^{(1)} \right\rangle = G \sum_{\substack{x \in \mathcal{X}_i: \\ n_1(x) = y_i^{(1)}}} d_{x,q} - G \sum_{\substack{x \in \mathcal{X}_i: \\ n_1(x) = a}} d_{x,q} = G[Ad_q]_i \geq 1.$$

Thus $d_q$ is feasible for Equation (37).

Let $\mathcal{C}_\phi$ and $\mathcal{C}_A$ denote the feasible sets of Equation (37) and Equation (38), respectively. We have shown

$$\mathcal{C}_\phi \subseteq \mathcal{C}_A, \qquad d_q \in \mathcal{C}_\phi.$$

Therefore

$$\min_{d \in \mathcal{C}_A} \frac{1}{2}\|d\|^2 \leq \min_{d \in \mathcal{C}_\phi} \frac{1}{2}\|d\|^2 \leq \frac{1}{2}\|d_q\|^2.$$

Since $d_q$ is the minimizer over $\mathcal{C}_A$, both inequalities are equalities. Hence $d_q$ also minimizes Equation (37).

Finally, any other minimizer of Equation (37) would belong to $\mathcal{C}_A$ and attain the same objective value as $d_q$, contradicting the uniqueness of the minimizer of Equation (38). Thus the two problems have the same unique minimizer. □

**Theorem D.3** (NTP score allocation; formal version of Theorem 4.1)**.** *For every fixed query token $q$, let $d_q$ be the unique solution of Equation* (38)*. Then, as $t \to \infty$,*

$$s_q^{\mathrm{NTP}}(t) = d_q \log t + O\left(\sqrt{\log t}\right). \tag{40}$$

*Equivalently,*

$$\frac{s_q^{\mathrm{NTP}}(t)}{\log t} \longrightarrow d_q. \tag{41}$$

*In particular, every token that is never first-horizon correct in the fixed-query system satisfies*

$$s_{x,q}^{\mathrm{NTP}}(t) = O\left(\sqrt{\log t}\right) = o(\log t). \tag{42}$$

*Proof.* By Lemma D.2, $d_q$ satisfies every NTP margin constraint. Hence, for any $R > 0$, every correct-to-competing margin at $Rd_q$ is at least $R$, and therefore

$$\mathcal{L}_q^{\mathrm{NTP}}(Rd_q) \longrightarrow 0 \qquad \text{as } R \to \infty.$$

By Lemma D.1, the NTP score vector follows the gradient flow of $\mathcal{L}_q^{\mathrm{NTP}}$. Applying Lemma C.4 with $\widetilde{s} = Rd_q$ gives

$$\mathcal{L}_q^{\mathrm{NTP}}(s_q^{\mathrm{NTP}}(t)) - \mathcal{L}_q^{\mathrm{NTP}}(Rd_q) \le \frac{\left\|s_q^{\mathrm{NTP}}(0) - Rd_q\right\|^2}{2\eta t}.$$

Sending first $t \to \infty$ and then $R \to \infty$ shows that

$$\mathcal{L}_q^{\mathrm{NTP}}(s_q^{\mathrm{NTP}}(t)) \longrightarrow 0. \tag{43}$$

The optimality conditions of Equation (37) express $d_q$ as a nonnegative linear combination of margin vectors whose constraints are active at the optimum. Since Equation (43) implies that every training margin diverges to $+\infty$,

$$\left\langle s_q^{\mathrm{NTP}}(t), d_q \right\rangle \longrightarrow \infty.$$

For the remainder of the proof, use the temporary normalization

$$r(t) = \frac{\left\langle s_q^{\mathrm{NTP}}(t), d_q \right\rangle}{\|d_q\|^2}.$$

Then $r(t) \to \infty$, and

$$\left\langle \frac{s_q^{\mathrm{NTP}}(t)}{r(t)} - d_q, d_q \right\rangle = 0. \tag{44}$$

Moreover, by Equation (32),

$$\dot{r}(t) = \frac{\eta}{\|d_q\|^2} \sum_{i \in \mathcal{D}_q} \sum_{a \neq y_i^{(1)}} p_{i,1}^{\mathrm{NTP}}(a,t) \left\langle d_q, \phi_{i,a}^{(1)} \right\rangle > 0. \tag{45}$$

To make the consequence of the optimality conditions explicit, let $\mathcal{J}_q^\star$ denote the active max-margin constraints. There exist multipliers $\alpha_{i,a} \ge 0$, supported on $\mathcal{J}_q^\star$, such that

$$d_q = \sum_{(i,a) \in \mathcal{J}_q^\star} \alpha_{i,a} \phi_{i,a}^{(1)}.$$

Because every active constraint satisfies $\langle d_q, \phi_{i,a}^{(1)} \rangle = 1$, taking the inner product with $d_q$ gives

$$\|d_q\|^2 = \sum_{(i,a) \in \mathcal{J}_q^\star} \alpha_{i,a}.$$

On the other hand, Equation (44) gives

$$\sum_{(i,a) \in \mathcal{J}_q^\star} \alpha_{i,a} \left\langle \frac{s_q^{\mathrm{NTP}}(t)}{r(t)}, \phi_{i,a}^{(1)} \right\rangle = \|d_q\|^2.$$

Thus the nonnegative weighted average of the current normalized margins over the active constraints equals one. At least one active constraint therefore satisfies

$$\left\langle \frac{s_q^{\mathrm{NTP}}(t)}{r(t)}, \phi_{i,a}^{(1)} \right\rangle \le 1. \tag{46}$$

For sufficiently large $t$, Equation (43) gives

$$p_{i,1}^{\mathrm{NTP}}(y_i^{(1)}, t) \ge \frac{1}{2}.$$

Together with Equation (46), the softmax identity

$$p_{i,1}^{\mathrm{NTP}}(a,t) = p_{i,1}^{\mathrm{NTP}}(y_i^{(1)},t)\exp\left(-\left\langle s_q^{\mathrm{NTP}}(t), \phi_{i,a}^{(1)}\right\rangle\right)$$

and the equality $\langle d_q, \phi_{i,a}^{(1)}\rangle = 1$ for an active constraint show that the following bound holds with the time-independent constant $P_q^{\mathrm{NTP}} := 1/2$:

$$\sum_{i\in\mathcal{D}_q}\sum_{a\neq y_i^{(1)}} p_{i,1}^{\mathrm{NTP}}(a,t)\left\langle d_q, \phi_{i,a}^{(1)}\right\rangle \geq P_q^{\mathrm{NTP}} e^{-r(t)}. \tag{47}$$

For every constraint, define

$$\Delta_{i,a}(t) := \left\langle \frac{s_q^{\mathrm{NTP}}(t)}{r(t)}, \phi_{i,a}^{(1)}\right\rangle - \left\langle d_q, \phi_{i,a}^{(1)}\right\rangle.$$

Writing $m_{i,a} := \langle d_q, \phi_{i,a}^{(1)}\rangle \geq 1$, consider first a constraint with $\Delta_{i,a}(t) > 0$. The softmax identity and $p_{i,1}^{\mathrm{NTP}}(y_i^{(1)},t) \leq 1$ give

$$\begin{aligned} p_{i,1}^{\mathrm{NTP}}(a,t)\Delta_{i,a}(t) &\leq e^{-r(t)m_{i,a}}\Delta_{i,a}(t)e^{-r(t)\Delta_{i,a}(t)} \\ &\leq \frac{e^{-r(t)}}{e\, r(t)}, \end{aligned}$$

where the last step uses $\sup_{u>0} ue^{-r(t)u} = 1/(e\, r(t))$. Terms with $\Delta_{i,a}(t) \leq 0$ do not increase the difference between the numerator and denominator below. Since the fixed-query system has finitely many constraints, summing the last estimate and using Equation (47) gives a constant $K_q^{\mathrm{NTP}} > 0$ such that

$$\frac{\displaystyle\sum_{i\in\mathcal{D}_q}\sum_{a\neq y_i^{(1)}} p_{i,1}^{\mathrm{NTP}}(a,t)\left\langle \frac{s_q^{\mathrm{NTP}}(t)}{r(t)}, \phi_{i,a}^{(1)}\right\rangle}{\displaystyle\sum_{i\in\mathcal{D}_q}\sum_{a\neq y_i^{(1)}} p_{i,1}^{\mathrm{NTP}}(a,t)\left\langle d_q, \phi_{i,a}^{(1)}\right\rangle} \leq 1 + \frac{K_q^{\mathrm{NTP}}}{r(t)}. \tag{48}$$

Set

$$\mathcal{V}_q^{\mathrm{NTP}}(t) = \frac{1}{2}\left\| \frac{s_q^{\mathrm{NTP}}(t)}{r(t)} - d_q \right\|^2.$$

Write $u_q(t) := s_q^{\mathrm{NTP}}(t)/r(t)$. Differentiating with respect to $\log r(t)$ gives

$$\frac{du_q}{d\log r} = \frac{\dot{s}_q^{\mathrm{NTP}}}{\dot{r}} - u_q.$$

Using Equation (44), the derivative of the Lyapunov function can therefore be written as

$$\frac{d\mathcal{V}_q^{\mathrm{NTP}}}{d\log r} = \left\langle u_q - d_q, \frac{\dot{s}_q^{\mathrm{NTP}}}{\dot{r}}\right\rangle - 2\mathcal{V}_q^{\mathrm{NTP}}.$$

Substituting Equations (32) and (45) into the first inner product and applying Equation (48) gives, with $B_q^{\mathrm{NTP}} := \|d_q\|^2 K_q^{\mathrm{NTP}} > 0$,

$$\frac{d\mathcal{V}_q^{\mathrm{NTP}}}{d\log r} \leq -2\mathcal{V}_q^{\mathrm{NTP}} + \frac{B_q^{\mathrm{NTP}}}{r}.$$

Since $\dot{r}(t) > 0$ by Equation (45), we may regard $\mathcal{V}_q^{\mathrm{NTP}}$ as a function of $r$. Since $d/d\log r = r\, d/dr$, the preceding inequality is equivalent to

$$\frac{d\mathcal{V}_q^{\mathrm{NTP}}}{dr} + \frac{2}{r}\mathcal{V}_q^{\mathrm{NTP}} \leq \frac{B_q^{\mathrm{NTP}}}{r^2}.$$

Multiplication by the integrating factor $r^2$ gives

$$\frac{d}{dr}\left(r^2\mathcal{V}_q^{\mathrm{NTP}}(r)\right) \le B_q^{\mathrm{NTP}}.$$

Fix a sufficiently large $t_0$ and write $r_0 = r(t_0) > 0$. Integrating from $r_0$ to $r(t)$ yields

$$r(t)^2\mathcal{V}_q^{\mathrm{NTP}}(t) \le r_0^2\mathcal{V}_q^{\mathrm{NTP}}(t_0) + B_q^{\mathrm{NTP}}\big(r(t) - r_0\big).$$

Since $r(t) \to \infty$, division by $r(t)^2$ gives $\mathcal{V}_q^{\mathrm{NTP}}(t) = O(r(t)^{-1})$, and hence

$$s_q^{\mathrm{NTP}}(t) = r(t)d_q + O\left(\sqrt{r(t)}\right). \tag{49}$$

Recall that $m_{i,a} = \langle d_q, \phi_{i,a}^{(1)}\rangle$. Feasibility of $d_q$ gives $m_{i,a} \ge 1$ for every constraint, and $m_{i,a} = 1$ for at least one active constraint. Since the number of training constraints is finite, Equation (49) gives uniformly

$$\left\langle s_q^{\mathrm{NTP}}(t), \phi_{i,a}^{(1)}\right\rangle = r(t)m_{i,a} + O\left(\sqrt{r(t)}\right).$$

Together with the softmax ratio and the eventual bound $p_{i,1}^{\mathrm{NTP}}(y_i^{(1)}, t) \ge 1/2$ from Equation (43), the active constraint supplies the lower bound below, while $m_{i,a} \ge 1$ for all constraints supplies the upper bound. Thus Equation (45) gives constants $\kappa_q^-, \kappa_q^+, \beta_q > 0$ such that

$$\kappa_q^- e^{-r(t)-\beta_q\sqrt{r(t)}} \le \dot{r}(t) \le \kappa_q^+ e^{-r(t)+\beta_q\sqrt{r(t)}}$$

for all sufficiently large $t$.

Because $r(t)$ is increasing, we may view $t$ as a function of $r$. Inverting the last display gives

$$\frac{1}{\kappa_q^+}e^{r-\beta_q\sqrt{r}} \le \frac{dt}{dr} \le \frac{1}{\kappa_q^-}e^{r+\beta_q\sqrt{r}}.$$

For sufficiently large $r$, both exponents are increasing. Integrating the lower bound over $[r-1, r]$ and the upper bound from a fixed value to $r$ therefore yields constants $c_1, c_2 > 0$ such that

$$c_1 e^{r-\beta_q\sqrt{r}} \le t \le c_2 r e^{r+\beta_q\sqrt{r}}.$$

Taking logarithms and using $\log r = O(\sqrt{r})$ gives

$$\log t = r(t) + O\left(\sqrt{r(t)}\right).$$

This relation first implies $r(t) = O(\log t)$; substituting this bound back into the remainder gives

$$r(t) = \log t + O\left(\sqrt{\log t}\right).$$

Substituting this estimate into Equation (49) proves Equation (40) and Equation (41).

Finally, every token that is never first-horizon correct has a zero column in $A$, and hence a zero coordinate in $d_q$ by Lemma D.2. Equation (42) then follows from Equation (40). □

**Corollary D.4** (Failure of global full-cue separation under NTP; formal version of Corollary 4.2)**.** *Suppose there exist a full cue $x^\star$ and a near cue $x^c$ associated with the same query token $q$ such that*

$$[A]_{i,x^c} \ge [A]_{i,x^\star}, \qquad i \in \mathcal{D}_q.$$

*Then*

$$d_{x^c,q} \ge d_{x^\star,q},$$

*and consequently, as $t \to \infty$,*

$$\min_{x \text{ full}} s_{x,q}^{\mathrm{NTP}}(t) - \max_{x \text{ near}} s_{x,q}^{\mathrm{NTP}}(t) = \left(\min_{x \text{ full}} d_{x,q} - \max_{x \text{ near}} d_{x,q}\right)\log t + O\left(\sqrt{\log t}\right), \tag{50}$$

*where*

$$\min_{x \text{ full}} d_{x,q} - \max_{x \text{ near}} d_{x,q} \le 0.$$

*Thus NTP does not achieve a positive logarithmic-order separation between all full cues and all near cues.*

*Proof.* Suppose, to the contrary, that $d_{x^\star,q} > d_{x^c,q}$. For sufficiently small $\epsilon > 0$, decrease $d_{x^\star,q}$ by $\epsilon$ and increase $d_{x^c,q}$ by the same amount.

For every $i \in \mathcal{D}_q$, the change in the $i$-th component of $Ad_q$ is

$$\epsilon\left([A]_{i,x^c} - [A]_{i,x^\star}\right) \geq 0.$$

Hence the modified vector remains feasible for Equation (38). Its squared norm changes by

$$-2\epsilon\left(d_{x^\star,q} - d_{x^c,q}\right) + 2\epsilon^2 < 0$$

for sufficiently small $\epsilon$, contradicting the minimum-norm optimality of $d_q$. Therefore

$$d_{x^c,q} \geq d_{x^\star,q}.$$

It follows that

$$\min_{x \text{ full}} d_{x,q} - \max_{x \text{ near}} d_{x,q} \leq d_{x^\star,q} - d_{x^c,q} \leq 0.$$

Finally, Theorem D.3 gives Equation (50). □

**Remark D.5.** Since $A$ is binary,

$$[AA^\top]_{ij} = \left|\{x : x \text{ is first-horizon correct in both } X_i \text{ and } X_j\}\right|.$$

Thus $d_q$ depends entirely on how often next-token-compatible cues occur and overlap across training sentences.

When all constraints in Equation (38) are active and $AA^\top$ is invertible,

$$d_q = \frac{1}{G} A^\top (AA^\top)^{-1}\mathbf{1}.$$

# E Proofs for Prediction Consequences

We derive the prediction rules in Section 5 from the learned score asymptotics.

Here $t$ denotes Stage 2 gradient-flow time. All asymptotic statements are taken as $t \to \infty$, with the training data, vocabulary, Stage 1 checkpoint and frozen readouts, and evaluation context fixed. At each $t$, predictions use the parameters reached at that time; no parameter update is performed on the evaluation context.

In each evaluation result, $q$ is an observed query and $X$ is a fixed finite sequence ending in $q$, with distinct token identities. Every token in its token set $\mathcal{X}$ is covered by the Stage 1 examples and has the frozen readout in Proposition 3.1. Token–query pairs need not have appeared during Stage 2. These are the evaluation conditions of Section 5.1. After dropping the class-independent shift, the effective head-$h$ logit of a frozen MTP model is

$$Z^{(h)}_{\mathrm{MTP}}(a \mid X, q, t) = G \sum_{\substack{x \in \mathcal{X}: \\ n_h(x) = a}} s^{\mathrm{MTP}}_{x,q}(t), \qquad h \in \{1, 2\}.$$

Similarly, the effective NTP logit is

$$Z^{(1)}_{\mathrm{NTP}}(a \mid X, q, t) = G \sum_{\substack{x \in \mathcal{X}: \\ n_1(x) = a}} s^{\mathrm{NTP}}_{x,q}(t).$$

## E.1 Controlled Recombination Testing

Fix a query token $q$ and a full cue $x^\star \in \mathcal{F}_q$, and define

$$y^{(h)} = n_h(x^\star), \qquad h \in \{1, 2\}.$$

Let $X_{\mathrm{test}}$ be a fixed finite test context whose token set $\mathcal{X}_{\mathrm{test}}$ contains $q$ and $x^\star$, where $x^\star$ is the unique full cue. The remaining tokens may include non-full tokens previously observed with $q$ as well as token–query pairs that were not observed during Stage 2.

**Corollary E.1** (Prediction under controlled recombination, NTP part; formal version of Corollary 5.3)**.** *For each output class $a \in \mathcal{V}$, define*

$$\lambda_a := G \sum_{\substack{x \in \mathcal{X}_{\text{test}} \cap \mathcal{U}_q: \\ n_1(x)=a}} d_{x,q}.$$

*The following limits hold as $t \to \infty$:*

1. *If $\lambda_{y^{(1)}} > \max_{a \neq y^{(1)}} \lambda_a$, then*

$$p_1^{\text{NTP}}\left(y^{(1)} \mid X_{\text{test}}, q, t\right) \longrightarrow 1.$$

2. *If there exists $a \neq y^{(1)}$ such that $\lambda_a > \lambda_{y^{(1)}}$, then*

$$p_1^{\text{NTP}}\left(y^{(1)} \mid X_{\text{test}}, q, t\right) \longrightarrow 0.$$

*Proof.* By Theorem D.3, every token $x \in \mathcal{U}_q$ satisfies

$$s_{x,q}^{\text{NTP}}(t) = d_{x,q} \log t + O(\sqrt{\log t}). \tag{51}$$

If the token pair $(x, q)$ was not observed during Stage 2, then the corresponding coordinate of $W_{kq}$ is never updated. Since $W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$,

$$s_{x,q}^{\text{NTP}}(t) = 1. \tag{52}$$

Using the frozen Stage 1 readout,

$$\widehat{W}_{uv}^{(1)} e_x = G\left(e_{n_1(x)} - \frac{1}{V}\mathbf{1}\right),$$

the first-head logit of class $a$ on the test context is

$$z_{\text{NTP}}^{(1)}(a,t) = G \sum_{\substack{x \in \mathcal{X}_{\text{test}}: \\ n_1(x)=a}} s_{x,q}^{\text{NTP}}(t) - \frac{G}{V} \sum_{x \in \mathcal{X}_{\text{test}}} s_{x,q}^{\text{NTP}}(t).$$

The second term is independent of $a$. Because $X_{\text{test}}$ has fixed finite length, Equation (51) and Equation (52) imply

$$z_{\text{NTP}}^{(1)}(a,t) = \lambda_a \log t + O(\sqrt{\log t}) + C(t),$$

where $C(t)$ is independent of $a$.

Suppose first that $\lambda_{y^{(1)}} > \max_{a \neq y^{(1)}} \lambda_a$ holds. For every $a \neq y^{(1)}$,

$$z_{\text{NTP}}^{(1)}(y^{(1)},t) - z_{\text{NTP}}^{(1)}(a,t) = \left(\lambda_{y^{(1)}} - \lambda_a\right) \log t + O(\sqrt{\log t}) \longrightarrow +\infty.$$

Hence

$$\frac{p_1^{\text{NTP}}(a \mid X_{\text{test}}, q, t)}{p_1^{\text{NTP}}(y^{(1)} \mid X_{\text{test}}, q, t)} = \exp\left(z_{\text{NTP}}^{(1)}(a,t) - z_{\text{NTP}}^{(1)}(y^{(1)},t)\right) \longrightarrow 0$$

for every $a \neq y^{(1)}$. Since the output vocabulary is finite,

$$p_1^{\text{NTP}}\left(y^{(1)} \mid X_{\text{test}}, q, t\right) = \frac{1}{1 + \sum_{a \neq y^{(1)}} \exp\left(z_{\text{NTP}}^{(1)}(a,t) - z_{\text{NTP}}^{(1)}(y^{(1)},t)\right)} \longrightarrow 1.$$

Now suppose that there exists $a \neq y^{(1)}$, such that $\lambda_a > \lambda_{y^{(1)}}$ holds. Then

$$z_{\text{NTP}}^{(1)}(a,t) - z_{\text{NTP}}^{(1)}(y^{(1)},t) = \left(\lambda_a - \lambda_{y^{(1)}}\right) \log t + O(\sqrt{\log t}) \longrightarrow +\infty.$$

Therefore

$$p_1^{\text{NTP}}\left(y^{(1)} \mid X_{\text{test}}, q, t\right) \leq \frac{1}{1 + \exp\left(z_{\text{NTP}}^{(1)}(a,t) - z_{\text{NTP}}^{(1)}(y^{(1)},t)\right)} \longrightarrow 0.$$

□

**Corollary E.2** (Prediction under controlled recombination, MTP part; formal version of Corollary 5.3)**.** *As $t \to \infty$, for every fixed test context above with at least one distractor, i.e., $\mathcal{X}_{\text{test}} \setminus \{x^\star\} \neq \varnothing$,*

$$s^{\text{MTP}}_{x^\star,q}(t) - \max_{x \in \mathcal{X}_{\text{test}} \setminus \{x^\star\}} s^{\text{MTP}}_{x,q}(t) = \frac{1}{G} \log t + O(\sqrt{\log t}).$$

*For every fixed test context above, including a singleton context,*

$$p^{\text{MTP}}_h \left( y^{(h)} \mid X_{\text{test}}, q, t \right) \longrightarrow 1, \qquad h \in \{1, 2\}. \tag{53}$$

*Proof.* By Theorem C.6,

$$s^{\text{MTP}}_{x^\star,q}(t) = \frac{1}{G} \log t + O(\sqrt{\log t}).$$

For every Stage 2-observed non-full token $x \in \mathcal{U}_q \setminus \mathcal{F}_q$,

$$s^{\text{MTP}}_{x,q}(t) = O(\sqrt{\log t}).$$

As in Equation (52), a token–query coordinate unseen during Stage 2 remains equal to its initialization and hence has score $O(1)$.

Because the test context is fixed and finite, the maximum score among its distractors is $O(\sqrt{\log t})$. Subtracting this quantity from the full-cue expansion above proves the score-gap claim in the corollary.

Fix $h \in \{1, 2\}$ and a competing output $a \neq y^{(h)}$. Using the frozen Stage 1 readout, the correct-to-competing logit gap is

$$z^{(h)}_{\text{MTP}}(y^{(h)}, t) - z^{(h)}_{\text{MTP}}(a, t) = G \sum_{x \in \mathcal{X}_{\text{test}}} s^{\text{MTP}}_{x,q}(t) \left[ \mathbb{1}\{n_h(x) = y^{(h)}\} - \mathbb{1}\{n_h(x) = a\} \right]. \tag{54}$$

Since $n_h(x^\star) = y^{(h)}$, the contribution of $x^\star$ to Equation (54) is

$$G s^{\text{MTP}}_{x^\star,q}(t) = \log t + O(\sqrt{\log t}).$$

Every other token in the test context contributes at most $O(\sqrt{\log t})$ if its token–query pair was observed during Stage 2, and $O(1)$ otherwise. Because $X_{\text{test}}$ is fixed and finite, their total contribution is $O(\sqrt{\log t})$. Therefore

$$z^{(h)}_{\text{MTP}}(y^{(h)}, t) - z^{(h)}_{\text{MTP}}(a, t) = \log t + O(\sqrt{\log t}) \longrightarrow +\infty.$$

Hence, for every $a \neq y^{(h)}$,

$$\frac{p^{\text{MTP}}_h(a \mid X_{\text{test}}, q, t)}{p^{\text{MTP}}_h(y^{(h)} \mid X_{\text{test}}, q, t)} = \exp\left( z^{(h)}_{\text{MTP}}(a, t) - z^{(h)}_{\text{MTP}}(y^{(h)}, t) \right) \longrightarrow 0.$$

Since $\mathcal{V}$ is finite,

$$p^{\text{MTP}}_h \left( y^{(h)} \mid X_{\text{test}}, q, t \right) = \frac{1}{1 + \sum_{a \neq y^{(h)}} \exp\left( z^{(h)}_{\text{MTP}}(a, t) - z^{(h)}_{\text{MTP}}(y^{(h)}, t) \right)} \longrightarrow 1.$$

Since $h \in \{1, 2\}$ was arbitrary, Equation (53) follows. □

### E.2 Prediction on Fixed Contexts and Compositional Extrapolation

We now consider an arbitrary fixed context $X_{\text{test}}$ satisfying the evaluation conditions in Section 5.1, with zero, one, or multiple training full cues. This context may be unseen during Stage 2. Unlike controlled recombination, this setting does not prescribe a ground-truth next token; the prediction follows from the learned score allocation.

**Theorem E.3** (Prediction rule for a fixed context, NTP part; formal version of Theorem 5.1)**.** *For every output class $a \in \mathcal{V}$, define*

$$\lambda_a = G \sum_{\substack{x \in \mathcal{X}_{\text{test}} \cap \mathcal{U}_q: \\ n_1(x) = a}} d_{x,q}.$$

*Then, as $t \to \infty$,*

$$Z^{(1)}_{\text{NTP}}(a,t) = \lambda_a \log t + O(\sqrt{\log t}). \tag{55}$$

*Define*

$$\mathcal{A}^\star_{\text{NTP}} = \arg\max_{a \in \mathcal{V}} \lambda_a.$$

*If $\mathcal{A}^\star_{\text{NTP}} = \{a^\star\}$ is a singleton, then $\widehat{y}^{(1)}_{\text{NTP}}(t) = a^\star$ for all sufficiently large $t$, and*

$$p^{(1)}_{\text{NTP}}\left(a^\star \mid X_{\text{test}}, q, t\right) \longrightarrow 1. \tag{56}$$

*If $|\mathcal{A}^\star_{\text{NTP}}| > 1$, the leading logarithmic scale only guarantees that outputs outside $\mathcal{A}^\star_{\text{NTP}}$ are asymptotically dominated; lower-order terms determine the preference within $\mathcal{A}^\star_{\text{NTP}}$.*

*Proof.* By Theorem D.3, every token observed with query token $q$ during Stage 2 satisfies

$$s^{\text{NTP}}_{x,q}(t) = d_{x,q} \log t + O(\sqrt{\log t}).$$

A token that is never first-horizon correct has $d_{x,q} = 0$. Moreover, if the token pair $(x, q)$ was never observed during Stage 2, then its score remains equal to its initialization and hence contributes only $O(1)$.

Substituting these score estimates into Equation (4), and using the fixed finite length of $X_{\text{test}}$, gives

$$Z^{(1)}_{\text{NTP}}(a,t) = \lambda_a \log t + O(\sqrt{\log t}),$$

which proves Equation (55).

If $a^\star$ uniquely maximizes $\lambda_a$, then for every $a \neq a^\star$,

$$Z^{(1)}_{\text{NTP}}(a^\star,t) - Z^{(1)}_{\text{NTP}}(a,t) = (\lambda_{a^\star} - \lambda_a) \log t + O(\sqrt{\log t}) \longrightarrow \infty.$$

The argmax statement and Equation (56) then follow from the softmax formula.

If several classes attain the same maximal coefficient, every class with a strictly smaller coefficient is asymptotically dominated, while the leading logarithmic scale does not distinguish among the maximizers. □

**Theorem E.4** (Prediction rule for a fixed context, MTP part; formal version of Theorem 5.1)**.** *For every output class $a \in \mathcal{V}$, define*

$$\mu_a = \left|\{x \in \mathcal{X}_{\text{test}} \cap \mathcal{F}_q : n_1(x) = a\}\right|.$$

*Then, as $t \to \infty$,*

$$Z^{(1)}_{\text{MTP}}(a,t) = \mu_a \log t + O(\sqrt{\log t}). \tag{57}$$

*Define*

$$\mathcal{A}^\star_{\text{MTP}} = \arg\max_{a \in \mathcal{V}} \mu_a.$$

*If $\mathcal{A}^\star_{\text{MTP}} = \{a^\star\}$ and $\mu_{a^\star} > 0$, then $\widehat{y}^{(1)}_{\text{MTP}}(t) = a^\star$ for all sufficiently large $t$, and*

$$p^{(1)}_{\text{MTP}}\left(a^\star \mid X_{\text{test}}, q, t\right) \longrightarrow 1.$$

*If several output classes attain the largest positive value of $\mu_a$, then the leading-order prediction is restricted to those classes, while lower-order terms determine the preference among them.*

*If $\mathcal{X}_{\text{test}} \cap \mathcal{F}_q = \varnothing$, then $\mu_a = 0$ for every $a \in \mathcal{V}$, so no output class has a nonzero leading logarithmic coefficient.*

*Proof.* By Theorem C.6, every full cue satisfies

$$Gs^{\mathrm{MTP}}_{x,q}(t) = \log t + O(\sqrt{\log t}),$$

whereas every trained non-full token satisfies

$$Gs^{\mathrm{MTP}}_{x,q}(t) = O(\sqrt{\log t}).$$

If the token pair $(x, q)$ was never observed during Stage 2, its score remains equal to its initialization and hence contributes only $O(1)$.

Therefore,

$$Z^{(1)}_{\mathrm{MTP}}(a,t) = G \sum_{\substack{x \in \mathcal{X}_{\mathrm{test}} \cap \mathcal{F}_q: \\ n_1(x)=a}} s^{\mathrm{MTP}}_{x,q}(t) + O(\sqrt{\log t}) = \mu_a \log t + O(\sqrt{\log t}),$$

which proves Equation (57).

If $a^\star$ uniquely maximizes $\mu_a$, then for every $a \neq a^\star$,

$$Z^{(1)}_{\mathrm{MTP}}(a^\star, t) - Z^{(1)}_{\mathrm{MTP}}(a,t) = (\mu_{a^\star} - \mu_a) \log t + O(\sqrt{\log t}) \longrightarrow \infty.$$

The claimed prediction and probability follow from the softmax formula.

Finally, if $\mathcal{X}_{\mathrm{test}} \cap \mathcal{F}_q = \varnothing$, every contextual contribution is of lower order than $\log t$. Thus no output receives a nonzero leading logarithmic coefficient. □

**Corollary E.5** (One full cue)**.** *Suppose the generalized context contains exactly one full cue $x^\star$. Then, as $t \to \infty$,*

$$\widehat{y}^{(1)}_{\mathrm{MTP}}(t) = n_1(x^\star)$$

*for all sufficiently large $t$, and*

$$p^{(1)}_{\mathrm{MTP}}\left(n_1(x^\star) \mid X_{\mathrm{test}}, q, t\right) \longrightarrow 1.$$

*Thus, in the presence of a unique full cue, the leading MTP prediction is unaffected by any fixed finite collection of non-full or query-unseen tokens in the same context.*

*Proof.* In this case,

$$\mu_{n_1(x^\star)} = 1, \qquad \mu_a = 0 \quad \forall a \neq n_1(x^\star).$$

The result follows from Theorem E.4. □

**Corollary E.6** (Multiple full cues)**.** *Suppose the generalized context contains $m \geq 2$ full cues. Then, at the leading logarithmic scale, MTP predicts according to the plurality of their first-horizon collocations.*

*In particular, if some $a^\star$ uniquely maximizes*

$$\left|\{x \in \mathcal{X}_{\mathrm{test}} \cap \mathcal{F}_q : n_1(x) = a\}\right|,$$

*then*

$$\widehat{y}^{(1)}_{\mathrm{MTP}}(t) = a^\star$$

*for all sufficiently large $t$. If several classes tie for the largest count, lower-order terms determine the prediction among the tied classes.*

*Proof.* This follows from Theorem E.4. □

**Corollary E.7** (Prediction without full-cue evidence)**.** *Suppose*

$$\mathcal{X}_{\mathrm{test}} \cap \mathcal{F}_q = \varnothing.$$

*Then, for every $a \in \mathcal{V}$, as $t \to \infty$,*

$$Z^{(1)}_{\mathrm{MTP}}(a,t) = O(\sqrt{\log t}). \tag{58}$$

*Hence no output class is distinguished at the leading logarithmic scale.*

*Proof.* Every trained token in the generalized context is non-full and therefore has score

$$O(\sqrt{\log t}),$$

while every token never paired with $q$ has constant score. Since $X_{\mathrm{test}}$ is fixed and finite, Equation (58) follows directly from Equation (4). □

# F Extension to General Readouts

In the main analysis, deterministic single-token examples and symmetric Stage 1 training yield the single-peaked frozen readout

$$\widehat{W}_{uv}^{(h)} e_x = G\big(e_{n_h(x)} - \frac{1}{V}\mathbf{1}\big).$$

This structure gives the sharp score separation in the main results. Here we relax the conditions on the Stage 1 collocation data. The same token may have multiple single-token examples, and its realized future token at a given horizon need not be identical across these examples. The Stage 1 training procedure itself is unchanged, and the learned prediction heads are still frozen before Stage 2.

For every token $x$ and prediction head $h$, denote its frozen Stage 1 readout by

$$w_h(x) := \widehat{W}_{uv}^{(h)} e_x \in \mathbb{R}^V.$$

We no longer impose the role-based assumptions defined through the deterministic maps $n_h$, such as prefix consistency, unique full cues, or role consistency. We retain distinct token identities within each prefix, one-hot token representations, linear attention, frozen readouts, and Stage 2 gradient flow. Under these conditions, the derivation below depends on the Stage 2 samples, their queries and targets, and the binary incidence of token identities in each contextual sequence. With repeated token identities, the binary incidence factors below would instead have to be replaced by their multiplicities.

For a fixed query $q$, define $\mathcal{D}_q := \{i : q_i = q\}$ as the index set of training sentences with query $q$, and still let $\mathcal{U}_q := \bigcup_{i\in\mathcal{D}_q} \mathcal{X}_i$ denote the token identities appearing in Stage 2 samples with query $q$. For a training sentence $i$, define the target-relative margin

$$\rho_{i,h,a}(x) := [w_h(x)]_{y_i^{(h)}} - [w_h(x)]_a, \qquad a \neq y_i^{(h)}.$$

The margin $\rho_{i,h,a}(x)$ measures how strongly the frozen readout for $x$ favors the realized horizon-$h$ target over a competing output $a$. It is invariant to adding an arbitrary multiple of $\mathbf{1}$ to $w_h(x)$ and therefore depends only on the softmax-relevant part of the readout.

**Theorem F.1** (MTP attention with general Stage 1 readouts)**.** *Fix a query $q$. For every token $x \in \mathcal{U}_q$, its MTP attention score satisfies*

$$\frac{d}{dt} s_{x,q}(t) = \eta \sum_{i\in\mathcal{D}_q} \mathbb{1}\{x \in \mathcal{X}_i\} \sum_{h=1}^{2} \sum_{a\neq y_i^{(h)}} p_i^{(h)}(a,t)\rho_{i,h,a}(x). \tag{59}$$

*Consequently, for any two tokens $x, x' \in \mathcal{U}_q$, if*

$$\mathbb{1}\{x \in \mathcal{X}_i\}\rho_{i,h,a}(x) \geq \mathbb{1}\{x' \in \mathcal{X}_i\}\rho_{i,h,a}(x') \tag{60}$$

*for every $i \in \mathcal{D}_q$, $h \in \{1,2\}$, and $a \neq y_i^{(h)}$, then, under equal initialization,*

$$s_{x,q}(t) \geq s_{x',q}(t), \qquad t \geq 0. \tag{61}$$

The main collocation analysis is a structured special case of this update rule. For general frozen readouts, the second horizon shapes attention through the readout margins relative to each target and competitor. Comparing these margins extends the full-versus-near cue distinction to readouts that are not single-peaked. NTP uses only the first-horizon terms.

## F.1 Proof of the General-Readout Result

*Proof of Theorem F.1.* For a training sentence $i$ and prediction head $h$, the Stage 2 logit vector is

$$z_i^{(h)}(t) = \widehat{W}_{uv}^{(h)} X_i X_i^\top W_{kq}(t) e_{q_i} = \sum_{u\in\mathcal{X}_i} s_{u,q_i}(t) w_h(u). \tag{62}$$

Hence

$$p_i^{(h)}(t) = \mathrm{softmax}\Big(z_i^{(h)}(t)\Big).$$

Let

$$\ell_i^{(h)}(t) = -\log p_i^{(h)}(y_i^{(h)}, t).$$

Since

$$\nabla_{z_i^{(h)}} \ell_i^{(h)} = p_i^{(h)} - e_{y_i^{(h)}},$$

Equation (62) gives, for $u \in \mathcal{X}_i$,

$$\frac{\partial z_i^{(h)}}{\partial s_{u,q_i}} = w_h(u).$$

Therefore

$$\begin{aligned} -\frac{\partial \ell_i^{(h)}}{\partial s_{u,q_i}} &= [w_h(u)]_{y_i^{(h)}} - \sum_{a \in \mathcal{V}} p_i^{(h)}(a,t)[w_h(u)]_a \\ &= \sum_{a \neq y_i^{(h)}} p_i^{(h)}(a,t) \left( [w_h(u)]_{y_i^{(h)}} - [w_h(u)]_a \right) \\ &= \sum_{a \neq y_i^{(h)}} p_i^{(h)}(a,t) \rho_{i,h,a}(u). \end{aligned} \tag{63}$$

It remains to account for the fact that the same score coordinate may contribute to several training sentences. Since

$$s_{u,q_i} = e_u^\top W_{kq} e_{q_i},$$

we have

$$\nabla_{W_{kq}} s_{u,q_i} = e_u e_{q_i}^\top.$$

Under gradient flow,

$$\dot{W}_{kq}(t) = -\eta \sum_i \sum_{h=1}^{2} \nabla_{W_{kq}} \ell_i^{(h)}(t).$$

Thus, for a fixed token pair $(x, q)$,

$$\frac{d}{dt} s_{x,q}(t) = e_x^\top \dot{W}_{kq}(t) e_q = -\eta \sum_i \sum_{h=1}^{2} \sum_{u \in \mathcal{X}_i} \frac{\partial \ell_i^{(h)}}{\partial s_{u,q_i}} e_x^\top e_u \, e_{q_i}^\top e_q.$$

Because the token representations are one-hot,

$$e_x^\top e_u = \mathbb{1}\{x = u\}, \qquad e_{q_i}^\top e_q = \mathbb{1}\{q_i = q\}.$$

Using Equation (63), we obtain

$$\frac{d}{dt} s_{x,q}(t) = \eta \sum_{i \in \mathcal{D}_q} \mathbb{1}\{x \in \mathcal{X}_i\} \sum_{h=1}^{2} \sum_{a \neq y_i^{(h)}} p_i^{(h)}(a,t) \rho_{i,h,a}(x),$$

which proves Equation (59).

For two tokens $x, x' \in \mathcal{U}_q$, subtracting their score dynamics gives

$$\begin{aligned} &\frac{d}{dt} \left( s_{x,q}(t) - s_{x',q}(t) \right) \\ &= \eta \sum_{i \in \mathcal{D}_q} \sum_{h=1}^{2} \sum_{a \neq y_i^{(h)}} p_i^{(h)}(a,t) \Big[ \mathbb{1}\{x \in \mathcal{X}_i\} \rho_{i,h,a}(x) \\ &\qquad - \mathbb{1}\{x' \in \mathcal{X}_i\} \rho_{i,h,a}(x') \Big]. \end{aligned} \tag{64}$$

If Equation (60) holds, every term on the right-hand side of Equation (64) is nonnegative. Hence

$$\frac{d}{dt} \left( s_{x,q}(t) - s_{x',q}(t) \right) \geq 0.$$

Since all score coordinates have the same initialization,

$$s_{x,q}(0) = s_{x',q}(0),$$

which proves Equation (61).

If Equation (60) is strict for at least one triple $(i, h, a)$, then the corresponding softmax probability $p_i^{(h)}(a, t)$ is strictly positive for every finite $t$. All remaining terms in Equation (64) are nonnegative, so

$$\frac{d}{dt}\left(s_{x,q}(t) - s_{x',q}(t)\right) > 0$$

for every finite $t$. Equal initialization therefore gives

$$s_{x,q}(t) > s_{x',q}(t), \qquad t > 0.$$

□

# G Experimental Validation

We examine how the prediction objective affects cue selection in two settings. Synthetic experiments use the two-stage training protocol to study cue ordering, score growth, and alignment with the max-margin solutions. Experiments on TinyStories examine learned readout associations and their weighting by attention in jointly trained models.

## G.1 Experiments on Synthetic Data

The synthetic setting isolates the effect of the prediction objective on shared attention. We generate sequences satisfying Assumption 1: each sequence contains a unique full cue, together with near-cue and uninformative tokens whose horizon-wise roles are preserved across samples with the same query. We represent each token identity by a one-hot vector and impose the idealized collocation structure from Proposition 3.1. The frozen readout associated with token $x$ is concentrated on its designated future token $n_h(x)$ for each horizon $h$.

**Synthetic model parameterization.** We optimize the composite matrices $W_{kq}$ and $W_{uv}^{(h)}$ directly, as in the theoretical analysis. In linear attention, the unnormalized scores $X^\top W_{kq} e_q$ weight the context tokens directly. The softmax variant retains the same composite matrices and applies softmax to these scores instead.

For each attention mechanism, we compare NTP and MTP using the same synthetic data, initialization, and training protocol. We train the composite readouts in Stage 1, freeze them at a finite checkpoint, and update only $W_{kq}$ in Stage 2. The models differ only in whether shared attention is trained with the first prediction head or with both heads. We compare cue ordering and score growth within each sequence, and alignment with the corresponding max-margin solution.

**Synthetic corpus.** We generate the corpus with a vocabulary of $V = 404$ coordinates. Stage 1 contains 383 singleton anchors, one for every evidence or query token. Stage 2 contains 504 training sequences and a 96-sequence recombination test set. Each sequence ends in one of 16 query tokens. The corpus is generated with seed 42. Among the queries, three are query-full, six are query-near, and seven are query-uninformative; they generate 18, 108, and 378 training sequences, respectively.

Each input token $x$ has fixed latent collocations $(n_1(x), n_2(x))$. For each query $q$, the generator samples three active first-horizon branches $A_q \subset \mathcal{Y}_1$. For each $a \in A_q$, it samples three active second-horizon branches $B_{q,a} \subset \mathcal{Y}_2$, yielding a $3 \times 3$ target-pair pool. For every active pair $(a, b)$, it creates four distinct evidence tokens with collocations $(a, b)$, with seeded log-normal reuse weights.

The query regime determines the allowed targets. For a query-full query, the query has fixed pair $(n_1(q), n_2(q)) = (a_0, b_0)$, and only $(a_0, b_0)$ is used as a target. For a query-near query, $n_1(q) = a_0$, while $n_2(q) \notin \bigcup_{a \in A_q} B_{q,a}$; hence targets are restricted to the three pairs $(a_0, b)$, $b \in B_{q,a_0}$. For a query-uninformative query, both collocations lie outside the active branch pool, and all nine pairs $(a, b)$, $a \in A_q,\ b \in B_{q,a}$, are used. Each allowed target pair occurs in six training sequences. Thus,

each query-full, query-near, and query-uninformative query contributes 6, 18, and 54 sequences, respectively.

The generator first samples and randomly permutes the preceding evidence tokens, then appends the query at the final position. In the evidence portion, a query-full sequence contains no evidence full cue, whereas query-near and query-uninformative sequences contain one; all regimes contain 2–4 near evidence cues and 3–6 uninformative evidence cues. The appended query is respectively the unique full cue, an additional near cue, or an additional uninformative cue in the three regimes. Consequently, the total sequence lengths are 6–11 for query-full and 7–12 for query-near and query-uninformative sequences. Equivalently, the whole-sequence role counts are

| regime | (#full, #near, #uninformative) |
|---|---|
| query-full | $(1, 2\text{–}4, 3\text{–}6)$ |
| query-near | $(1, 3\text{–}5, 3\text{–}6)$ |
| query-uninformative | $(1, 2\text{–}4, 4\text{–}7)$. |

Here full, near, and uninformative mean $(n_1, n_2) = (y_1, y_2)$, $n_1 = y_1,\ n_2 \neq y_2$, and $n_1 \neq y_1,\ n_2 \neq y_2$, respectively; far-only cues are prohibited.

The generator enforces these constraints automatically. It aborts if a target lies outside its query-specific branch pool, if an allowed target has an incorrect multiplicity, if the appended query has an incorrect role, or if any role-count, duplicate-token, inactive-class, Stage 1-coverage, unique-full-cue, or repeated token–query role-consistency check fails.

**Optimization.** All runs use deterministic full-batch gradient descent, with both Python and PyTorch; there is no minibatching, momentum, weight decay, gradient clipping, or stochastic sampling. Python and PyTorch both use seed 42. Stage 1 initializes both readout matrices at zero and performs 250 updates with fixed step size. The Stage 1 step size is 1.0. For the $N_1 = 383$ anchors, each head minimizes its own sample-averaged cross-entropy,

$$\mathcal{L}_h^{(1)} = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathrm{CE}\Big(z_i^{(h)}, y_i^{(h)}\Big) .$$

At the saved checkpoint, both heads attain 100% singleton top-1 recovery . More directly, the two heads share the one-vs-rest gap $G \approx 0.65$, and their mean single-peak residuals are $9.68 \times 10^{-9}$ and $3.55 \times 10^{-9}$, respectively.

Stage 2 freezes these readouts, initializes $W_{kq}(0) = \mathbf{1}\mathbf{1}^\top$, and performs 1,800 full-batch updates with fixed step size. The Stage 2 step size is 5.0. For $N_2 = 504$ sequences, NTP uses

$$\mathcal{L}_{\mathrm{NTP}} = \frac{1}{N_2} \sum_{i=1}^{N_2} \mathrm{CE}\Big(z_i^{(1)}, y_i^{(1)}\Big) ,$$

whereas MTP sums the two independently sample-normalized head losses,

$$\mathcal{L}_{\mathrm{MTP}} = \frac{1}{N_2} \sum_{i=1}^{N_2} \mathrm{CE}\Big(z_i^{(1)}, y_i^{(1)}\Big) + \frac{1}{N_2} \sum_{i=1}^{N_2} \mathrm{CE}\Big(z_i^{(2)}, y_i^{(2)}\Big) .$$

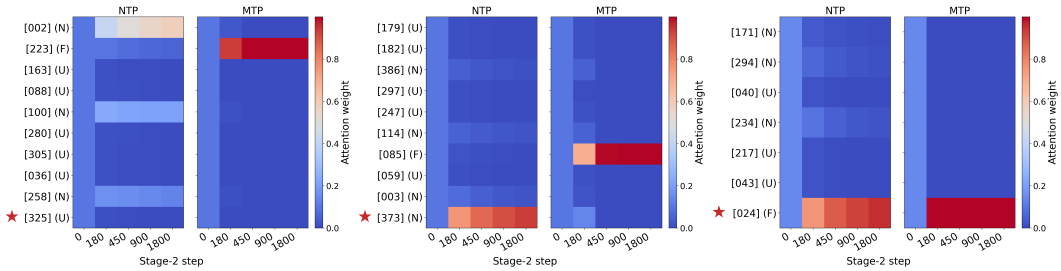
**Alignment targets and plotted time.** The incidence programs are post-hoc theoretical targets, not training objectives.

For MTP, we construct $c_q^\star = G^{-1} \sum_{x \in \mathcal{F}_q} e_x$ and check every stored constraint $\langle c_q^\star, \phi_{i,a}^{(h)} \rangle \geq 1$; the minimum margin over the actual corpus is exactly 1.

For NTP, we form the first-horizon sentence-by-token incidence matrix $A_q$ over $\mathcal{U}_q$ , and set $\boldsymbol{B}_q = GA_q$. We compute the minimum-norm feasible solution of $\boldsymbol{B}_q d \succeq \mathbf{1}$ by solving its nonnegative dual with projected gradient descent from zero, using $\lambda_{\max}(\boldsymbol{B}_q \boldsymbol{B}_q^\top)^{-1}$ as the step size. The solver stops when the largest dual-coordinate update falls below $10^{-8}$, with a 30,000-step safeguard. All 16 query-wise programs converge; the smallest recorded primal margin is $1 - 3.58 \times 10^{-6}$, while the explicit MTP construction has minimum margin exactly 1.

For the manuscript plots, the discrete time is $t_k = 1 + k$, where $k$ is the Stage 2 update index. The $k = 0$ point is omitted from diagnostics involving $\log t_k$, since $\log(1 + k) = 0$ there.

**Linear attention.** We first examine cue ordering in the linear-attention model analyzed in the main text. Figure 2 tracks the softmax-normalized attention of full, near, and uninformative cues within three selected sequences, which covers different roles of query tokens. The comparison tests whether MTP assigns higher scores to the full cue, while NTP can also assign high scores to near cues compatible with the first target.




*(a) uninformative query* *(b) near-cue query* *(c) full-cue query*

Figure 2: We select three example sentences from the dataset in which the query token serves as an uninformative, near-cue, or full-cue token, respectively. We plot the evolution of the attention distribution during training and normalize the attention scores using softmax. The index preceding each row denotes the token ID in the dataset. The symbol (★) marks the query token, while (F), (N), and (U) denote full-cue, near-cue, and uninformative tokens, respectively.

We next examine score growth in a selected sequence. Figure 3 shows that, at later checkpoints, the MTP full-cue score grows approximately linearly with $\log(1+k)$, while the near-cue score changes little. Under NTP, both scores increase and the near cue even attain a larger score than the full cue. These trajectories are consistent with the selective growth of full-cue scores characterized by Theorem 3.5.

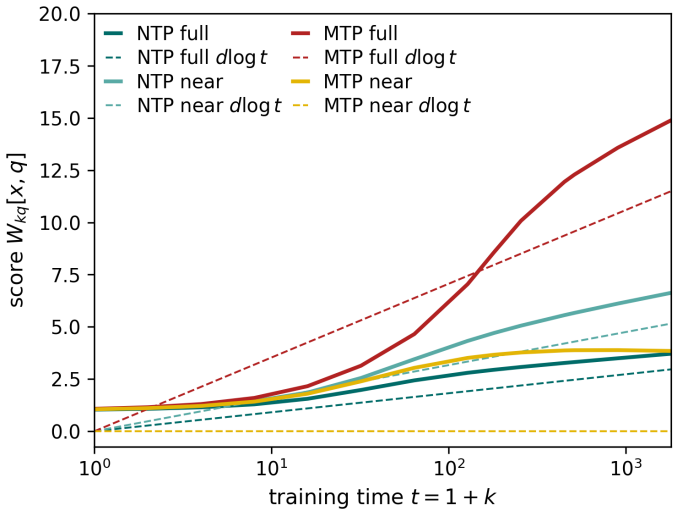



Figure 3: Raw attention-score growth for a full cue and a near cue in one selected synthetic sequence under NTP and MTP. The dashed lines are the growth of theoretical leading terms corresponding to each cue and model. The horizontal axis shows $1 + k$ on a logarithmic scale, where $k$ is the Stage 2 update index.

We further examine alignment with the max-margin solutions. For every query $q$ in the dataset, we compute the cosine similarity between $s_q(t)/\log t$ and the corresponding solution: $c_q^\star$ for MTP

and the computed first-horizon incidence-program solution $d_q$ for NTP. Figure 4 shows how this alignment evolves during Stage 2. For $t > 1$, dividing by $\log t$ leaves cosine similarity unchanged, so the comparison measures the direction of $s_q(t)$.

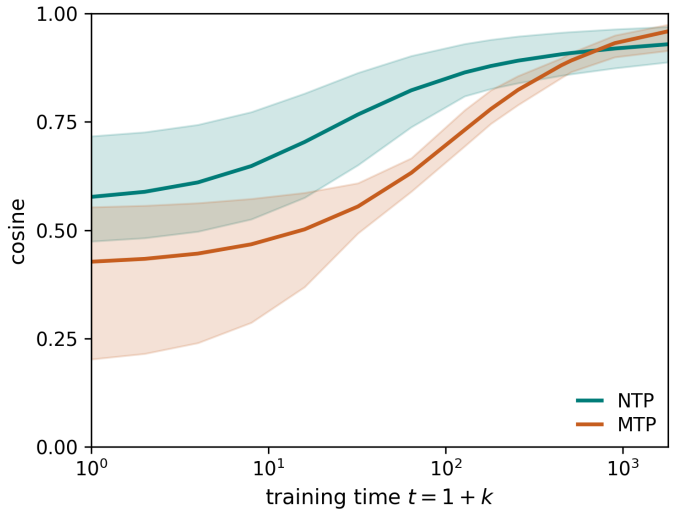



Figure 4: Synthetic linear attention: cosine alignment with each objective's max-margin solution. Solid curves show the unweighted mean over queries at each checkpoint, and bands show the range across queries; the horizontal coordinate is $t_k = 1 + k$, where $k$ is the Stage 2 update index.

**Unseen recombinations.** We compare first-head predictions on 96 recombined contexts that did not occur in Stage 2. Each context combines two Stage 2 prefixes $i \neq j$ with the same query $q$. From prefix $i$, we keep $q$ and its full cue $f = x_i^\star$, and use $(y^{(1)}, y^{(2)}) = (n_1(f), n_2(f))$ as the evaluation targets. From prefix $j$, we add non-full cues other than $q$, sampled at random, and drop any token whose associations equal both targets. By role consistency (Assumption 1), a non-full cue of $j$ is not a training full cue for $q$, so every context satisfies $\mathcal{X}_{\text{test}} \cap \mathcal{F}_q = \{f\}$, the setting of Corollary 5.3. Because the model depends on a context only through its token set, we keep a candidate only if its token set differs from that of every Stage 2 prefix. We shuffle the eligible candidates with seed 42 and keep the first 96.

Role consistency also shapes the NTP comparison. If $q$ is a near cue, every prefix with query $q$ has first target $n_1(q)$, so $\lambda_a = 0$ for every $a \neq y^{(1)}$; competing first-target support can arise only when $q$ is uninformative. The 96 contexts contain 10 near-cue and 86 uninformative queries; no eligible context has a full-cue query.

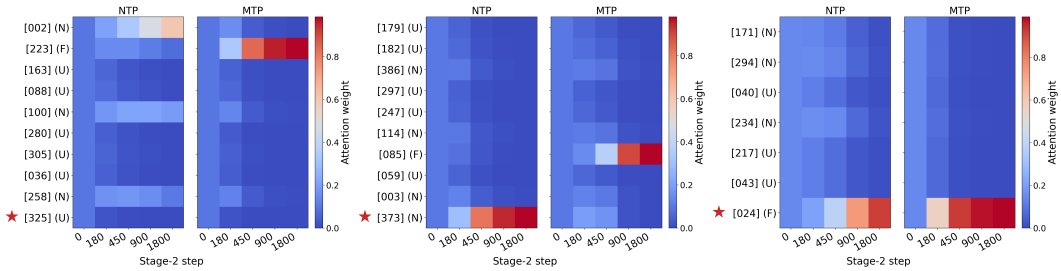
At the last Stage 2 checkpoint ($k = 1800$), MTP predicts $y^{(1)}$ on all 96 contexts, whereas NTP does so on 36 (Table 1). These finite-time results agree with the limits in Corollary 5.3. The contexts, construction code, and predictions are included in the supplementary material.

Table 1: First-head accuracy of linear attention on the 96 unseen recombination sequences.

| ATTENTION | NTP ACCURACY | MTP ACCURACY |
|---|---|---|
| Linear | 37.5% | 100.0% |

**Softmax attention.** We also examine cue selection under standard softmax attention, keeping the synthetic data, composite-matrix parameterization, two-stage training protocol, and NTP–MTP comparison unchanged. Figure 5 compares normalized attention on selected sequences to assess

whether MTP places more weight than NTP on the cue compatible with both future targets. With fixed finite readouts, the output is a convex combination of readout columns, so its margins remain bounded.




*(a) uninformative query* *(b) near-cue query* *(c) full-cue query*

Figure 5: We select three same example sentences in Figure 2. We plot the evolution of the softmax attention distribution during training. The index preceding each row denotes the token ID in the dataset. The symbol (★) marks the query token, while (F), (N), and (U) denote full-cue, near-cue, and uninformative tokens, respectively.

The softmax-attention models use the same serialized set of 96 recombinations. Their results are listed at the Table 2.

Table 2: First-head accuracy of softmax attention on the unseen recombination sequences.

| ATTENTION | NTP ACCURACY | MTP ACCURACY |
|---|---|---|
| Softmax | 38.5% | 100.0% |

### G.2 Experiments on Real-World Data

**Data and architecture.** We train an attention-only, one-layer GPT-style model on a fixed TinyStoriesV2-GPT4 subset containing approximately five million training tokens. The tokenizer is ByteLevel BPE with vocabulary size $V = 8192$; its entries include subwords, whitespace, and punctuation, rather than only complete words. The saved training split contains 25,374 complete stories and 5,000,001 tokens including EOS. Evaluation uses a disjoint validation split of 1,025 stories and 200,379 tokens. The tokenizer is learned only from training-source stories. Both objectives use the same 28,237 story-contained training windows and 1,145 validation windows. Each span contains 128 input positions and the two following targets, uses stride 128, and neither pads nor crosses story boundaries. Every input position is supervised, not only the final query. The model has one causal softmax-attention head, with embedding, key, and value dimensions all equal to 128. The causal mask includes the query position itself. There are no positional encodings, LayerNorm, feed-forward blocks, residual connections, biases, or dropout. Without positional encodings, the prediction at a query position depends on the query identity and the multiset of tokens in its causal prefix, not on their order.

**Optimization and comparison budgets.** All active parameters are trained jointly by AdamW (Loshchilov & Hutter, 2019). We train for 10,000 updates with batches of 32 windows. The learning rate warms up linearly for 500 updates to $3 \times 10^{-4}$ and then follows cosine decay to a floor of $3 \times 10^{-5}$; weight decay is 0.1, and the global gradient norm is clipped at 1.0. NTP uses the first-horizon loss, whereas parallel MTP uses the average of the two horizon losses, each averaged over batch and input positions. Query and key projections are initialized to the identity; embedding, value, output, and readout parameters are initialized from $\mathcal{N}(0, 0.02^2)$ and all remain trainable. Both runs use the same seed, the same windows, and identical shared-parameter and first-head initialization. The comparison uses one training seed and matches input exposure and optimizer updates. MTP additionally supervises the second horizon. The common training seed is 42.

**Effective matrices.** Let $E \in \mathbb{R}^{V\times d}$ denote the embedding table, $W_Q, W_K, W_V, W_O \in \mathbb{R}^{d\times d}$ the query, key, value, and attention-output projections of the single attention head, and $U_h \in \mathbb{R}^{V\times d}$ the horizon-$h$ prediction head. Using column-vector notation, the trained model gives the following exact effective value–output–readout map for each active head ($h = 1$ for NTP and $h \in \{1, 2\}$ for MTP):

$$W_{uv}^{(h)} := U_h W_O W_V E^\top \in \mathbb{R}^{V\times V}.$$

This matrix is derived from the jointly trained parameters. Its rows index output tokens and its columns index input tokens. The exact effective vocabulary-space key–query matrix is

$$W_{kq} := \frac{1}{\sqrt{d}} E W_K^\top W_Q E^\top \in \mathbb{R}^{V\times V}.$$

This includes the attention scaling; there is no normalization before the projections. At query position $t$, the raw score of position $\ell \le t$ is $e_{x_\ell}^\top W_{kq} e_{x_t}$, and the weights $a_{t\ell}$ are its softmax over the causal support, including the query itself. The full horizon-$h$ logit vector is $\sum_{\ell\le t} a_{t\ell} W_{uv}^{(h)} e_{x_\ell}$. These identities follow from the specified linear value–output path.

### G.2.1 Learned Associations and Attention Selection

Readout associations and attention determine how each context token contributes to the output. A token's readout column specifies its contribution, and attention assigns its weight in the context. We first inspect these columns, then examine the columns selected by the largest attention weights on training fragments.

*Readout profiles.* The trained readouts provide a token-to-future map analogous to the frozen readout $\widehat{W}_{uv}^{(h)} e_x$ in the main analysis. We examine this map as qualitative motivation for the collocation abstraction. For a token identity $x$, define

$$r_h(x) := W_{uv}^{(h)} e_x.$$

The full-prefix logits are an attention-weighted combination of these columns. For the single-token prefix $[e_x]$, the sole attention weight is one, so $r_h(x)$ is exactly its logit vector and $\mathrm{softmax}(r_h(x))$ is its predictive distribution.

We inspect non-special input tokens occurring at least 100 times in the training token stream, retaining the full output vocabulary. Figure 6 pairs the two heads' readout profiles for six manually selected input tokens in the trained two-horizon MTP model ($K = 2$). Bars show $[\mathrm{softmax}(r_h(x))]_a$, the probability that head $h$ assigns to output $a$ when the prefix is the single token $x$.

Following the figure notation, a leading _ marks a BPE token whose decoded spelling begins with a space; for example, `_boy` and `boy` denote distinct token identities.

The paired profiles reveal interpretable associations at both horizons. Examples include the following pairs.

1. For `_boy`, the readouts rank `_named` first at $h = 1$ and `_Tim` second at $h = 2$, compatible with a character-naming phrase such as *boy named Tim*.
2. `_wanted` and `_liked` both rank `_to` first at $h = 1$, while their second-horizon lists include `_play` and `_help`, compatible with phrases such as *wanted to play* and *liked to help*.
3. For `_little`, the first-horizon list includes `_boy`, `_girl`, and `_bird`, while the second-horizon list includes `_named`, compatible with a descriptive phrase such as *little boy named*.

These phrases provide possible interpretations of the separate rankings at each horizon. The figure retains five highest-ranked outputs of these examples.

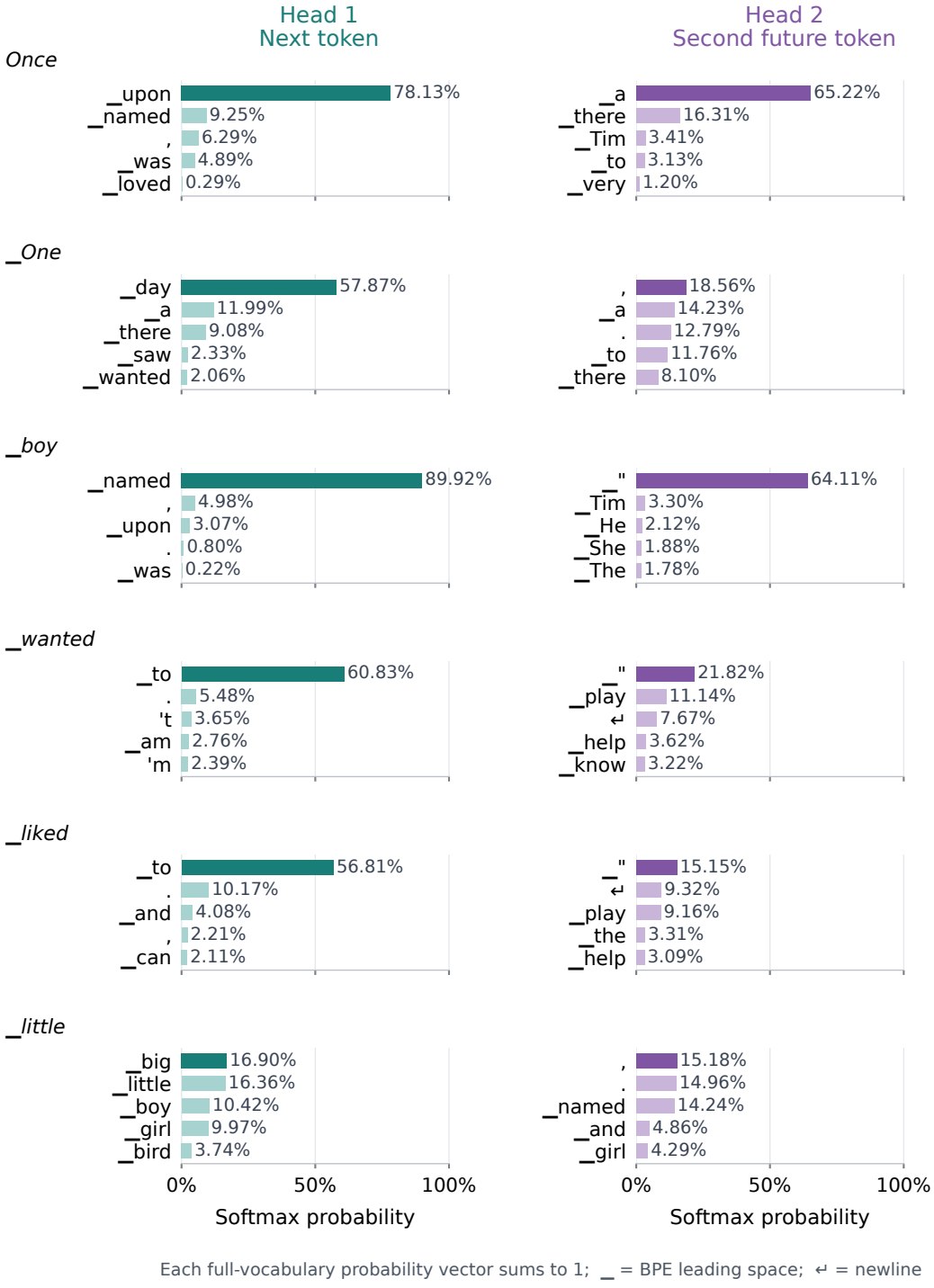



Figure 6: Top-five readout outputs for six selected input tokens in the trained two-horizon MTP model. Rows correspond to input tokens; columns show head 1 (left) and head 2 (right). Bars show softmax probability, ordered from highest to lowest, with the top output in a darker shade. Numeric labels are rounded to two decimal places. A leading underscore marks a BPE leading space, and ↵ denotes a newline token. Each profile is an isolated-token readout column.

*Attention-selected readouts on training prefixes.* We next examine which readout columns receive the most attention in selected training prefixes. The illustration set contains manually reviewed prefixes. Candidates were inspected and ranked by the total-variation distance between the two models' last-query attention distributions. We select four examples to show different forms of attention redistribution and readout alignment.

Each example is a contiguous fragment of a training-window prefix. We evaluate only the displayed fragment, retaining punctuation, BPE whitespace, and the query itself. The raw scores are $s_{x_\ell,q} = e_{x_\ell}^\top W_{kq} e_q$; the reported attention weights are their softmax over the displayed positions. Truncating earlier context leaves these scores unchanged in this architecture, but changes the softmax denominator. The labels $y^{(1)}, y^{(2)}$ are the two tokens following the final query in the original training window.

For each model, we select the position with maximum attention weight, breaking ties by the earliest position. Figure 7 reports this weight and the current targets' ranks in the selected token's readout.

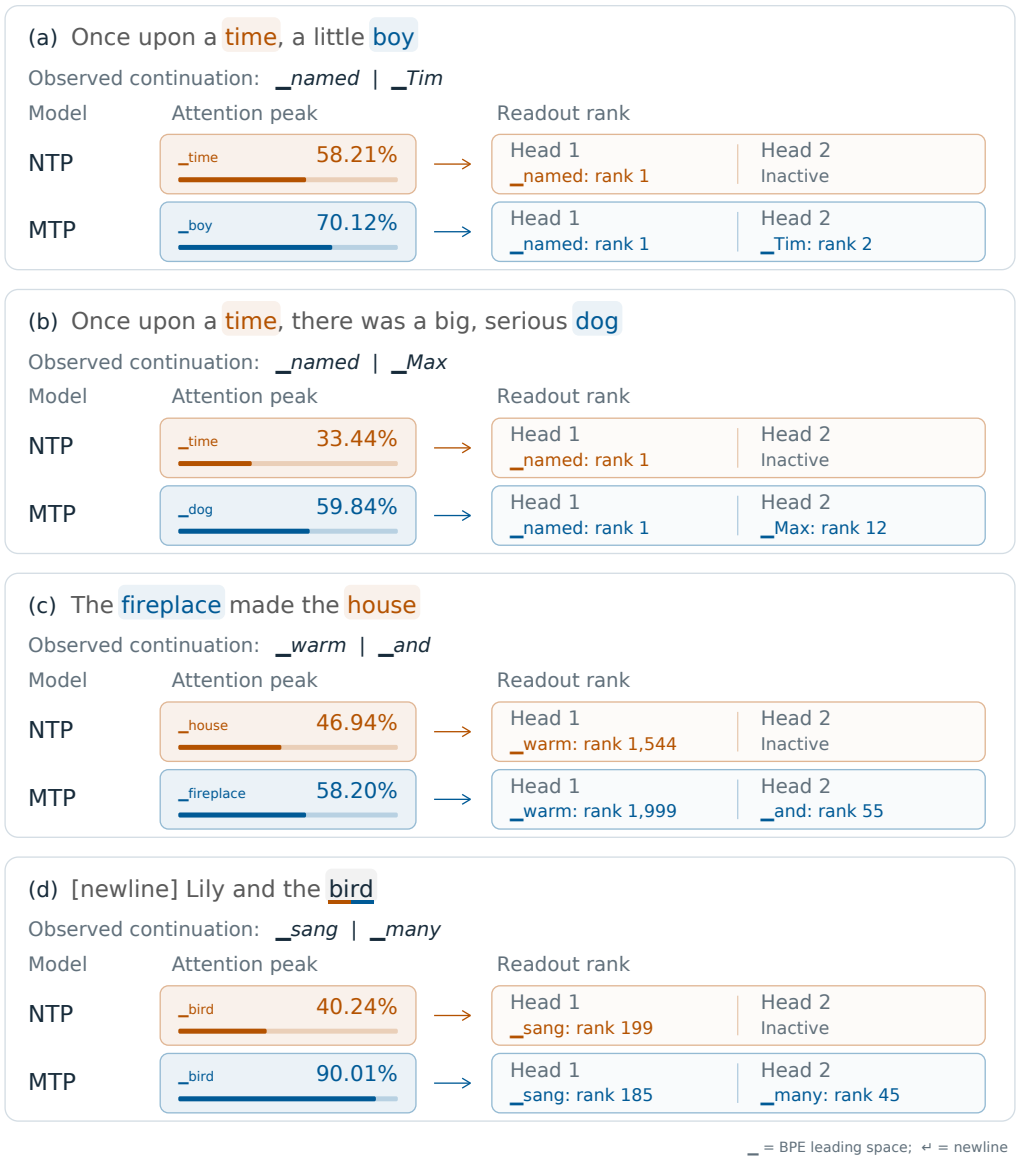



Figure 7: Attention peaks and readout ranks for four selected TinyStories training fragments. Orange and blue identify NTP and MTP, respectively; the split underline in (d) marks their shared peak. Percentages and bars show attention weights normalized over the displayed fragment. Ranks are evaluated over the full vocabulary using each model's own readout column for its peak token, with the two observed continuation tokens as the head 1 and head 2 targets. NTP has no active second head. A leading underscore marks a BPE leading space. In (a,b), the peak shifts from `_time` to `_boy` or `_dog`, while both selected columns rank `_named` first. In (c,d), MTP's first-head ranks for the next target are 1,999 and 185, respectively. The initial newline is shown explicitly.

*Different peaks with a shared first-head top token.* The boy and dog fragments show that the two objectives can select different readout columns whose first-head top output is the same. NTP places its largest attention weight on _time, whereas MTP places it on the query token _boy or _dog. Both models' selected columns rank the first target _named first. The MTP second head ranks _Tim second and _Max twelfth, connecting the attention shift to character-naming associations. The NTP association from _time to _named is not an ordinary lexical collocation. Because the columns are learned jointly with the rest of this position-free model, a high-ranked output need not admit a token-to-future interpretation based on the input token alone. The comparison illustrates a change in which token supplies a target-associated column.

*Redistribution without target dominance.* The fireplace and bird fragments show attention redistribution even when the selected readout columns rank many outputs above the observed targets. In the fireplace fragment of Figure 7, the peak shifts from the query _house to the earlier token _fireplace. The observed target _warm ranks only 1,999th in the MTP _fireplace column, so the redistributed attention does not select a target-associated column in this example. Attention on _fireplace increases while attention on _house decreases.

In the bird fragment, the peak remains on _bird, whose attention weight is 0.40 under NTP and 0.90 under MTP. The first-head column ranks the observed target _sang 199th under NTP and 185th under MTP, while the MTP second-head column ranks _many 45th. These quantities are reported jointly; without an intervention, they do not identify which readout association caused the attention change. These examples distinguish attention redistribution from target dominance in the selected column.

Together, the readout profiles and selected-prefix examples describe learned token-to-future associations and their weighting by attention.