

## A Detailed MAVEN-FACT Analysis

**Data extraction.** We use the *train* split of MAVEN-FACT v1.0 (73,939 event-mentions drawn from 2,913 news articles).<sup>4</sup> Each mention carries a FactBank-style factuality code (CT++, CT+, CT-, CT-, PS±, PR±, CF±, U, NA, ...). We retain only **certain** judgments:

$$\text{certain-true} = \{\text{CT++}, \text{CT+}\}, \quad \text{certain-false} = \{\text{CT-}, \text{CT-}\}.$$

All other codes are discarded, leaving  $N = 71,274$  labelled mentions.

**Grouping key.** Mentions are grouped by their originating article ID (`doc_id`), giving  $M = 2,913$  documents with at least two certain mentions ( $n_i > 1$ ). Let  $Z_{ij} \in \{0, 1\}$  indicate whether mention  $j$  in document  $i$  is *certain-false*.

**Statistics reported in the main text.**

- **Corpus certain-false rate.**  $p = \frac{1}{N} \sum_{i,j} Z_{ij} = 0.0209$ .
- **Pairwise certain-false probability.**  $\Pr(Z_j = Z_k = 1 \mid \text{same doc}) = \frac{\sum_i \binom{f_i}{2}}{\sum_i \binom{n_i}{2}} = 0.00090$ , where  $f_i = \sum_j Z_{ij}$ .
- **Independence baseline.**  $p^2 = 0.00044$ .
- **Clustering ratio.**  $\frac{\text{Var}_{\text{obs}}(\hat{p}_i)}{\text{Var}_{\text{binom}}} = \frac{\frac{1}{M} \sum_i (\hat{p}_i - p)^2}{\frac{1}{M} \sum_i p(1-p)/n_i} = 1.23$ , with  $\hat{p}_i = f_i/n_i$ .
- **$\chi^2$  test.** The  $2 \times M$  contingency table of  $\{f_i, n_i - f_i\}$  yields  $\chi^2 = 4174$  ( $p \approx 9 \times 10^{-49}$ ).

These figures show that *certain-false* events, though rare (2.1%), occur about twice as often as chance would predict when two events come from the same article, and the distribution of false rates across articles is 23 % more heterogeneous than a binomial model would permit—confirming the co-occurrence signal predicted by TCH.

The MAVEN-ED dataset is released with CC BY-SA 4.0 license. The MAVEN-ARG and MAVEN-ERE are published with GPLv3 license.

## B Experimental Setup

**Model.** We experiment with an attention-only transformer with a single attention head with a post-attention LN:

$$X^0 = E + P \quad // E, P \in \mathbb{R}^{V \times d} \text{ (token + positional embeddings)} \quad (10)$$

$$Q^{(i)} = X^{(i-1)} W_Q^{(i)}, K^{(i)} = X^{(i-1)} W_K^{(i)}, V^{(i)} = X^{(i-1)} W_V^{(i)} \quad (11)$$

$$A^{(i)} = \text{softmax}\left(\frac{Q^{(i)} K^{(i)\top}}{\sqrt{d}}\right) V^{(i)} \quad // \text{attention mix } A^{(i)} \in \mathbb{R}^d \quad (12)$$

$$\tilde{A}^{(i)} = A^{(i)} W_O^{(i)}, \quad W_O^{(i)} \in \mathbb{R}^{d \times d} \quad // \text{single-head attention output} \quad (13)$$

$$X^{(i)} = \text{N}(X^{(i-1)} + \tilde{A}^{(i)}), \quad i = 1, \dots, l \quad // \text{residual + normalization} \quad (14)$$

$$Z = X^{(l)} W_O + b_O, \quad W_O \in \mathbb{R}^{d \times V}, b_O \in \mathbb{R}^V \quad (15)$$

$$\hat{Y} = \text{softmax}(Z) \quad (16)$$

**Experiments with one-hot models (section 4).** The theoretical analysis is driven by experiments on models equipped with frozen, one-hot embeddings and uniform attention, the latter obtained by

<sup>4</sup>Available at <https://github.com/THU-KEG/MAVEN-FACT>.

806 setting the attention-key matrix  $K$  to the zero matrix. Under these conditions the *columns* of the  
807 attention value–output product  $KV^T$  map directly to individual vocabulary items, exposing a clear  
808 block structure in the matrix (fig. 1). As detailed in the main text, the vocabulary is organized so that  
809 indices 1–20 encode input subject embeddings, 21–40 input attribute embeddings, 41–44 positional  
810 embeddings, 45–64 output subject embeddings, and 65–84 output attribute embeddings.

811 **Methodology: interpreting one-hot embeddings.** Figure 2 contrasts two sequences—a correct  
812 one (top row) and an incorrect one (bottom row)—by showing the final-layer activations before  
813 projecting to the logit space. The one-hot embeddings make the activation patterns in that layer  
814 interpretable. We display the activations for the raw representations (left), after layer normalization  
815 (middle), and after applying the unembedding matrix and the softmax transformation (right). Observe  
816 the differing  $y$ -axis scales: normalization substantially magnifies the component corresponding to the  
817 correct answer in the “true” sequence, while the effect is far less pronounced for the false sequence.  
818 The model that produced fig. 1 was trained with SGD, learning rate 1.0 and batch size 16. The output  
819 matrix was fixed to identity, and only the value matrix was learned, from zero initialization.

820 **Experiments with fully-trained models (section 5):** In section 5, we train all components, including  
821 the input embeddings and the  $K$  attention matrix. The model is trained for 50,000 batches of size  
822 128 and is optimized with the Adam optimizer [Kingma and Ba, 2015] with a learning weight of  $1e-4$   
823 and a weight decay of  $1e-5$ . We do not include biases in the attention modules, and use RMSNorm as  
824 layer normalization. We run all experiments on 4 NVIDIA GeForce GTX 1080 GPUs. Training a  
825 single model lasts up to half an hour.

## 826 C Additional Experiments

827 In the main text we concentrated on a single-layer model ( $l = 1$ ) with a true-attribute probability of  
828  $\rho = 0.99$ . Here we extend the analysis to additional settings.

829 Our primary focus was the linear separability at the second-subject token,  $x'$ , where the model  
830 predicts the second attribute. This is the only position where the truth signal is *behaviorally* relevant.  
831 Nevertheless, the theory also predicts a linear truth encoding at the first-attribute token  $y$ , owing to the  
832 fixed attention pattern. When the attention  $KV$  matrix is learned, however, this need not occur—the  
833 model can rely exclusively on the attention paid to  $x'$  and leave  $y$  uninformative. The same theory  
834 further implies that a linear truth direction should eventually emerge for any true-sentence rate  $\rho$ ,  
835 even though the gradient magnitude (and therefore the speed of emergence) does depend on  $\rho$ .

836 **Varying the true sentence rate,  $\rho$ .** In fig. 6b we vary  $\rho$  across five random seeds and measure linear  
837 separability at both token positions. As predicted, when the attention pattern is learned, separability  
838 is much stronger at the second subject than at the first attribute. The time to emergence grows as  $\rho$   
839 increases, yet linear encoding still appears even at the extreme setting of  $\rho = 0.999$ . Developing a  
840 theory that precisely predicts this  $\rho$ -dependent timing is left to future work.

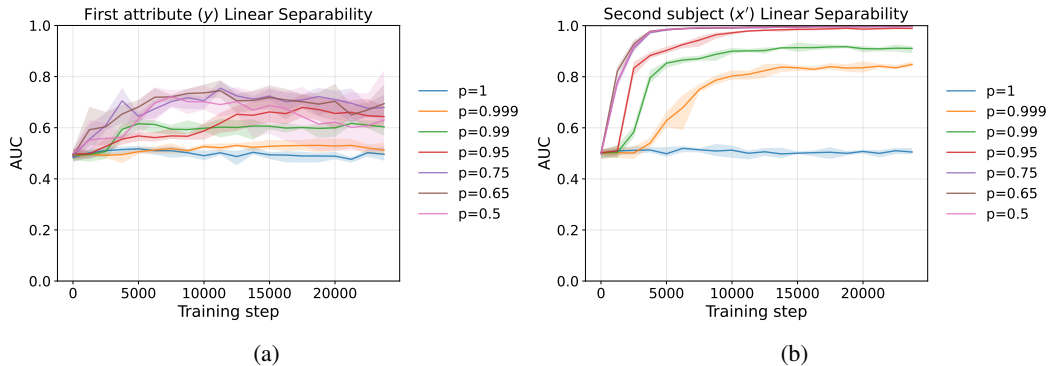


Figure 6: Dependency of linear separability on  $\rho$ .

841 **Dependency on  $d_{\text{model}}$  and  $|\mathcal{S}|$ .** In fig. 7 we plot the linear separability at the final checkpoint, for  
842 different hidden sizes and number of facts to memorize ( $\rho = 0.99$ ,  $l = 1$  are fixed). With the exception

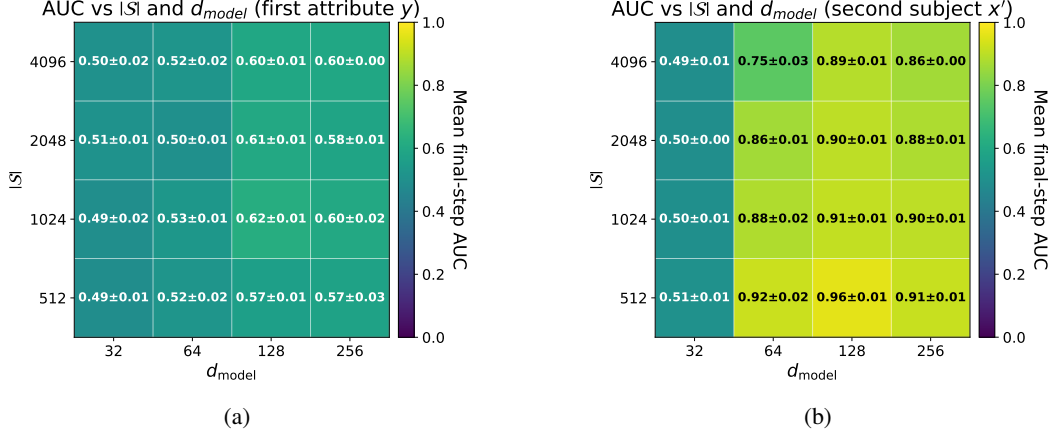


Figure 7: Dependency of linear separability on  $d_{\text{model}}$  and  $|\mathcal{S}|$ .

843 of  $d_{\text{model}} = 32$ , the separability persists over the second subject  $x'$  for different combinations of these  
844 parameters.

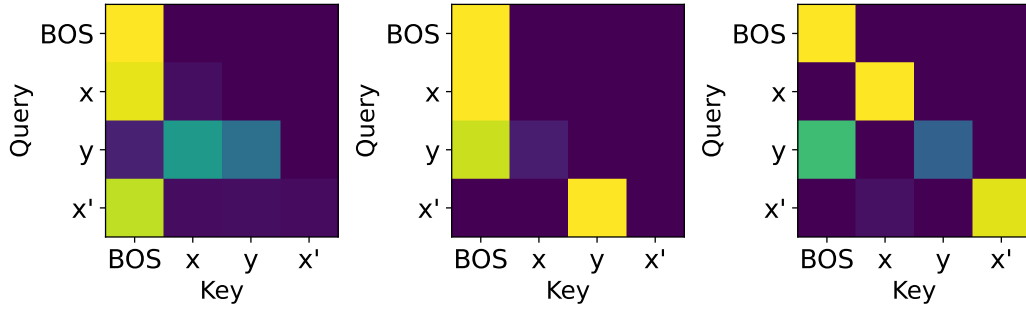


Figure 8: attention patterns of a 3-layer model.

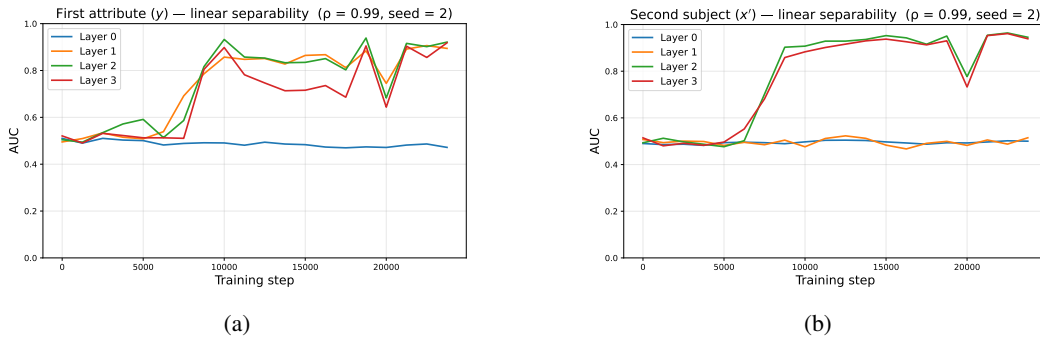


Figure 9: Linear separability across layers for a 3-layer model; linear separability on the  $x'$  token is created after *copying* the signal from the  $y$  token in the second layer.

845 **Additional layers.** As we discuss in the main-text (section 5), in a model with a single self-attention  
846 layer, it is the second attribute ( $x'$ ) token that attends to both  $x$  and  $y$ . With more layers, there are  
847 additional strategies. For instance,  $y$  may attend to both  $x$  and itself in the first layer, in the same way  
848  $x'$  attends to both  $x$  and  $y$  in the theoretical 1-layer model; then, in the next layer,  $x'$  attends to  $y$ ,  
849 copies the signal and create a linear separation that persists the last layer. This is the mechanism that  
850 emerges in 4/5 random initializations of a 3-layer model, and is clearly manifested in the attention  
851 patterns (fig. 8) and in the linear classification accuracy across layers (fig. 9).

852 **Bridging the gap between the fully-trainable model and the toy model.** Our theoretical analysis  
853 (appendix D) is motivated by the structured patterns that emerge in the attention kernel—the  $OV$   
854 matrix—when it is visualized (fig. 1). To test whether a comparable mechanism appears when we  
855 employ dense embeddings and allow the  $KV$  matrices to train freely (thus removing the enforced  
856 uniform attention over  $x, y$ ), we train a model with a large hidden dimension but only a small set of  
857 facts to memorize ( $|\mathcal{S}| = 32$  and  $d_{\text{model}} = 512$ ). We freeze the randomly-initialized dense embeddings  
858 and train all other parameters. The limited number of subjects makes the memorization patterns  
859 easier to inspect, while the high dimensionality approximates the regime of mutually orthogonal  
860 embeddings required by the theory.

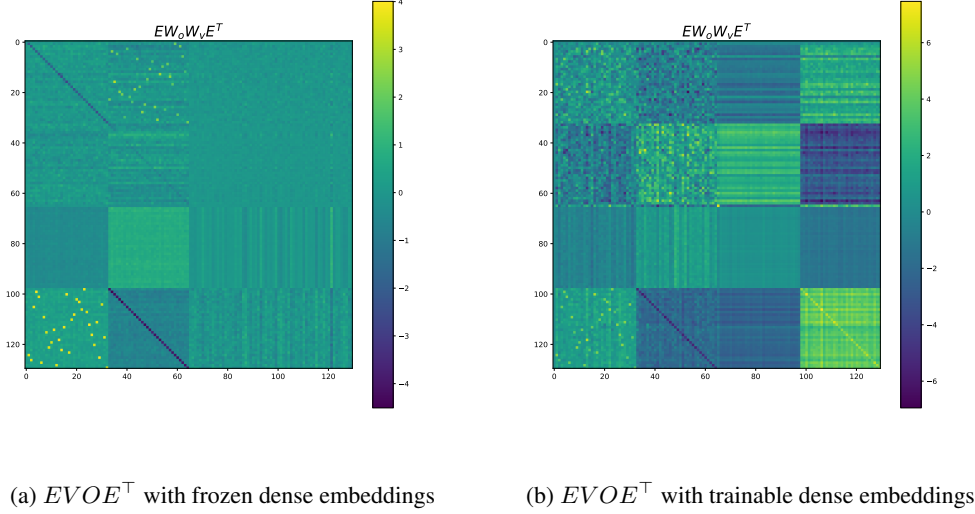


Figure 10: Visualization of the attention matrix with dense embeddings.

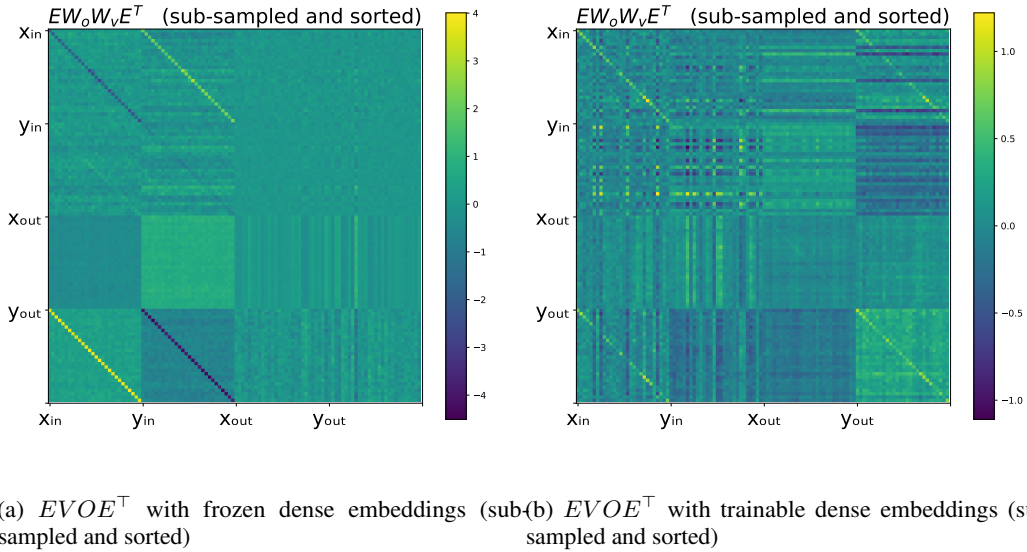


Figure 11: Visualization of the attention matrix with dense embeddings.

861 Because the model now uses dense embeddings—so individual coordinates no longer correspond  
862 directly to vocabulary items—we do not expect an obvious block structure in the raw  $OV$  matrix.  
863 Instead, following Dar et al. [2023], we visualize  $EVOE^\top$ , where  $E$  concatenates the input and

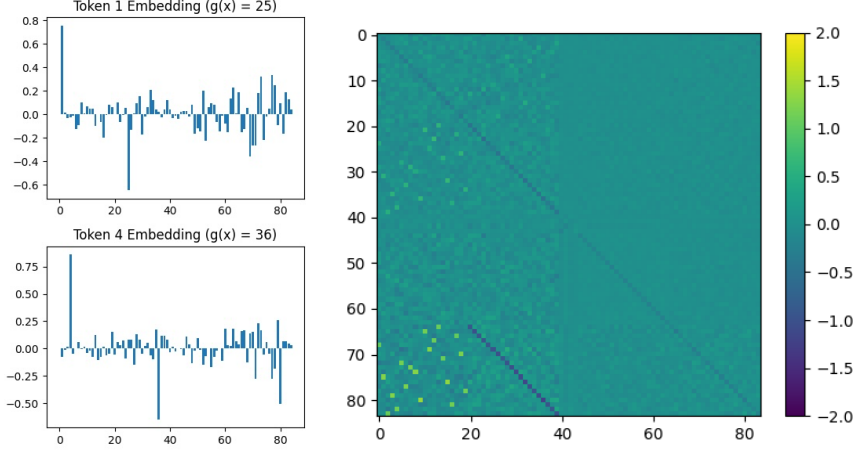


Figure 12: Visualization of learned embeddings and value matrix for a model as in Section 4 with learned embeddings, initialized to one-hot.

output embedding matrices. This operation computes the pairwise similarities between embeddings as induced by the  $VO$  transformation. Concretely,  $(EVOE^\top)_{ij} = E_i^\top V, O, E_j$  measures how strongly the value vector elicited by symbol  $i$  aligns with the output direction that scores symbol  $j$ , so every cell again describes a relation between concrete symbols, exactly what the raw  $OV$  matrix showed when the embeddings were one-hot. The resulting heat-map (fig. 10a) exhibits a strikingly similar pattern to that observed with frozen one-hot embeddings and a fixed attention pattern, suggesting that the dense model converges to a similar underlying mechanism. In contrast, when we do train the embeddings, the pattern partially disappears, as parts of the memorization can occur in the embeddings themselves (fig. 10b). In general, there is much more variability between runs and hyperparameters when training the embeddings, where some hyperparameter choices do not show a pattern that is highly similar to the idealized one.

With a full set of  $|\mathcal{S}| = d_{\text{model}} = 512$  tokens, the global pattern is hard to spot at first glance. If we instead sub-sample  $28x$  tokens, retain only their partners  $g(x)$ , and then sort the rows/columns, the latent memorization re-emerges: the lower-left block collapses into a clear diagonal (the previously random pattern in the leftmost lower block in fig. 10a is transformed into a diagonal due to the sorting). This diagonal appears whether the embeddings are frozen or trainable (see figs. 11a and 11b).

**One possible circuit with learned embeddings.** We now present one possible circuit that we found when initializing with the one-hot embeddings, in a simplified architecture with uniform attention as in Section 4. We still denote  $e_x, e_y, u_x, u_y$  the one-hot embeddings as in Section 4, which only refer to the initialization in this setting with learned embeddings. After training, we may visualize the learned embeddings and interpret them as linear combinations of the initial one-hot embeddings, as shown in Figure 12. Denoting  $\tilde{e}_x, \tilde{e}_y, \tilde{u}_x, \tilde{u}_y$  the embeddings after training, the circuit we found looks as follows:

$$\begin{aligned}\tilde{e}_x &= e_x - e_{g(x)} \\ \tilde{e}_y &= e_y - e_{g^{-1}(y)} \\ \tilde{u}_x &= \sum_x u_x - \sum_y u_y \\ \tilde{u}_y &= u_y + e_{g^{-1}(y)} \\ W &= \sum_x (u_{g(x)} - e_x) e_x^\top - \sum_y (e_y + u_y) e_y^\top.\end{aligned}$$

The approximation  $\tilde{e}_x = e_x - e_{g(x)}$ , for instance, follows from the two large positive and negative spikes in the left part of fig. 12, for indices 1 and 25/36. Similar to our analysis of Section 4, we compute the quantity  $W(\tilde{e}_x + \tilde{e}_y)$ , which appears in the residual stream for both token  $y$  and token  $x'$ :

$$W(\tilde{e}_x + \tilde{e}_y) = u_{g(x)} - e_x + e_{g(x)} + u_{g(x)} - e_y - u_y - u_y + e_{g^{-1}(y)}$$

890 We observe that this vanishes when  $y = g(x)$ , suggesting that a similar mechanism as in the fixed  
891 embeddings case studied in Section 4 is at play, where layer-norm can lead to sharper predictions for  
892 true sequences, as well as provide a truth direction.

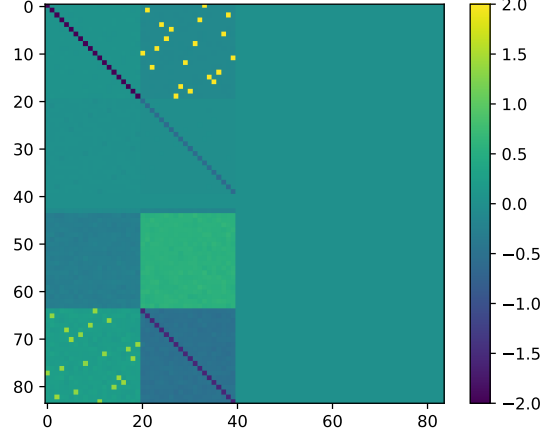


Figure 13: Structure of the value matrix  $W$  when training without positional embeddings.

## D Theoretical analysis

This section contains theoretical analysis and proofs for the results in Section 4.

### D.1 Training dynamics

We now provide some theoretical insights on the training dynamics in the simple one-layer model of Section 4. We further simplify the model here by removing positional embeddings. Figure 13 shows that the model still learns the relevant blocks even without positional embeddings, though some of the uniform distributions on unembeddings are now absorbed in other blocks.

The lemma below highlights the structure of the gradient for a softmax classification model consisting of a linear model followed by a layer-norm operation.

**Lemma 1.** Consider the model  $F_W(x) = U \cdot \mathbf{N}(a_x + Wb_x) \in \mathbb{R}^{2N}$ , with  $\mathbf{N}(v) = v/\|v\|$ , and the following cross-entropy population loss on some distribution over  $(x, y)$ :

$$L(W) = \mathbb{E}_{x,y}[-\log \mathcal{S}(F_W(x))_y], \quad (17)$$

where  $y$  is the label and  $\mathcal{S}$  the softmax operation. The gradient with respect to  $W$  is then given by:

$$\nabla L(W) = \sum_{k=1}^{2N} \mathbb{E}_{x,y} \left[ \frac{\mathcal{S}(U \cdot \mathbf{N}(v_x))_k - \mathbf{1}\{y = k\}}{\|v_x\|} \mathbf{P}_{(v_x/\|v_x\|)u_k} b_x^\top \right], \quad (18)$$

with  $v_x = a_x + Wb_x$  and where  $\mathbf{P}_\theta = I - \theta\theta^\top$  is the projection onto the tangent space at  $\theta \in \mathbb{S}^d$ .

Let us decompose the population loss as

$$L(W) = L_1(W) + L_2(W) + L_3(W), \quad (19)$$

where  $L_t(W)$  is the next-token prediction loss for predicting  $z_{t+1}$  from  $z_{1:t}$ , with  $z_{1:4} = (x, y, x', y')$ . We show the following result.

**Theorem 4.** Consider the following algorithm, with step-size  $\eta = N/\rho$ , and initialization  $W_0 = 0$ :

1. Set  $W_1 = W_0 - \eta \nabla L_1(W_0)$

2. Set  $W_2 = W_1 - \eta \nabla L_1(W_1)$

3. Set  $W_3 = W_2 - \eta \nabla L_3(W_2)$

Then, we have

$$W_3 = \sum_{x=1}^N (\beta_1 u_{g(x)} - \alpha_1 e_x) e_x^\top + \sum_y (\alpha_2 e_{g^{-1}(y)} - \beta_2 u_y) e_y^\top + o(1), \quad (20)$$

where  $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$  can be found in the proof.

915 *Proof.* Let us decompose each loss into contributions from true and false sequences, which follows  
 916 from the fact that the data distribution is a mixture of the two:

$$L_i(W) = \rho L_i^T(W) + (1 - \rho) L_i^F(W).$$

917 **Step 1.** In the first step, we take a gradient step only on the loss  $L_1$  for the prediction of the second  
 918 token  $y$  at the first token  $x$ , starting from initialization  $W_0 = 0$ . Recall that this model takes the  
 919 form  $F(x) = U \cdot \mathbf{N}(e_x + W e_x)$ , so that in the notation of Lemma 1 we have  $a_x = b_x = v_x = e_x$ .

920 We begin with the gradient on true sequences:

$$\begin{aligned} -\eta \nabla L_1^T(W_0) &= -\eta \sum_{k=1}^{2N} \mathcal{S}(0)_k u_k \mathbb{E}_x[e_x^\top] + \eta \mathbb{E}_x[u_{g(x)} e_x^\top] \\ &= \frac{\eta}{N} \sum_{x=1}^N u_{g(x)} e_x^\top - \frac{\eta}{2N^2} \sum_{z=1}^{2N} \sum_{x=1}^N u_z e_x^\top \\ &= \frac{\eta}{N} \sum_{x=1}^N u_{g(x)} e_x^\top + O(\eta/N^2). \end{aligned}$$

921 On false sequences, we have

$$\begin{aligned} -\eta \nabla L_1^F(W_0) &= -\eta \mathbb{E}_x \left[ \sum_{k=1}^{2N} \mathcal{S}(0)_k u_k e_x^\top \right] + \eta \mathbb{E}_{x,y} [u_y e_x^\top] \\ &= \frac{\eta}{N^2} \sum_{x=1}^N \sum_{y=N+1}^{2N} u_y e_x^\top - \frac{\eta}{2N^2} \sum_{z=1}^{2N} \sum_{x=1}^N u_z \\ &= O(\eta/N^2), \end{aligned}$$

922 using the fact that  $x$  and  $y$  are independent. With  $\eta = N/\rho$ , we obtain

$$W_1 = W_0 - \eta \nabla L_1(W_0) = \sum_{x=1}^N u_{g(x)} e_x^\top + O(1/N).$$

923 **Step 2.** For the second step taken at  $W = W_1$ , we will assume  $v_x = e_x + u_{g(x)}$ , so that  $\|v_x\| = \sqrt{2}$ .<sup>5</sup>  
 924 We also denote  $\sigma_{x,k} := \mathcal{S}(U \cdot \mathbf{N}(v_x))_k$ , noting that we have  $\sigma_{x,k} = O(1/N)$  for all  $x$  and  $k$ . On true  
 925 sequences, we have

$$\begin{aligned} -\eta \nabla L_1^T(W_1) &= \frac{\eta}{N\sqrt{2}} \sum_{x=1}^N \left( u_{g(x)} e_x^\top - \frac{1}{2} (e_x + u_{g(x)}) e_x^\top \right) - \frac{\eta}{N\sqrt{2}} \sum_{x=1}^N \sum_{k=1}^{2N} \sigma_{x,k} \left( u_k - \frac{\delta_{k,g(x)}}{2} v_x \right) e_x^\top \\ &= \frac{\eta}{2\sqrt{2}N} \sum_{x=1}^N (u_{g(x)} - e_x) e_x^\top + O(\eta/N^2), \end{aligned}$$

926 where  $\delta_{k,g(x)} = \mathbf{1}\{k = g(x)\}$  denotes the Kronecker delta. For false sequences, we have

$$\begin{aligned} -\eta \nabla L_1^F(W_1) &= \frac{\eta}{\sqrt{2}} \mathbb{E}_{x,y} \left[ \left( I - \frac{v_x v_x^\top}{2} \right) u_y e_x^\top \right] - \frac{\eta}{N\sqrt{2}} \sum_{x=1}^N \sum_{k=1}^{2N} \sigma_{x,k} \left( u_k - \frac{\delta_{k,g(x)}}{2} v_x \right) e_x^\top \\ &= \frac{\eta}{N^2\sqrt{2}} \sum_{x=1}^N \sum_{y=N+1}^{2N} \left( u_y - \frac{\delta_{y,g(x)}}{2} v_x \right) e_x^\top - \frac{\eta}{N\sqrt{2}} \sum_{x=1}^N \sum_{k=1}^{2N} \sigma_{x,k} \left( u_k - \frac{\delta_{k,g(x)}}{2} v_x \right) e_x^\top \\ &= O(\eta/N^2). \end{aligned}$$

927 With  $\eta = N/\rho$ , this yields

$$W_2 = W_1 - \eta \nabla L_1(W_1) = \sum_{x=1}^N (\alpha u_{g(x)} - e_x) e_x^\top + O(1/N),$$

928 with  $\alpha = 1 + \frac{1}{2\sqrt{2}}$ .

---

<sup>5</sup>This is true up to terms that vanish in the  $N \rightarrow \infty$  limit, but we will ignore them here for simplicity. We note that with more care, these can be incorporated in the analysis, leading to the same block structure.



929 **Step 3.** The third step takes one gradient step on the loss  $L_3$  at the third token, i.e., predicting  $y'$   
 930 from  $(x, y, x')$ . The model now takes the form  $F(x, y, x') = U \cdot \mathbf{N}(e_{x'} + \frac{1}{3}W(e_x + e_y + e_{x'}))$ .

931 The gradient of the loss on  $y'$  is given as in (18), where we assume<sup>6</sup>

$$\begin{aligned} v_{x,y,x'} &= e_{x'} + \frac{1}{3}W_2(e_x + e_y + e_{x'}) \\ &= \frac{2}{3}e_{x'} - \frac{1}{3}e_x + \frac{\alpha}{3}u_{g(x)} + \frac{\alpha}{3}u_{g(x')} =: v_{x,x'}. \end{aligned}$$

932 We have  $\|v_{x,x'}\| = \frac{1}{3}\sqrt{5 + 2\alpha^2}$  for  $x \neq x'$  and  $\|v_{x,x'}\| = \frac{1}{3}\sqrt{1 + 2\alpha^2}$  for  $x = x'$ . Note that we  
 933 once again have  $\sigma_{x,y,x',k} := \mathcal{S}(U \cdot \mathbf{N}(v_{x,y,x'}))_k = O(1/N)$ . On true sequences, we have

$$-\eta \nabla L_3^T(W_2) = \eta \mathbb{E}_{x,x'} \left[ \frac{1}{3\|v_{x,x'}\|} \left( I - \frac{v_{x,x'}v_{x,x'}^\top}{\|v_{x,x'}\|^2} \right) u_{g(x')}(e_x + e_{g(x)} + e_{x'})^\top \right] \quad (21)$$

$$- \eta \sum_{k=1}^{2N} \mathbb{E}_{x,x'} \left[ \frac{\sigma_{x,g(x),x',k}}{3\|v_{x,x'}\|} \left( I - \frac{v_{x,x'}v_{x,x'}^\top}{\|v_{x,x'}\|^2} \right) u_k(e_x + e_{g(x)} + e_{x'})^\top \right] \quad (22)$$

934 It is easy to check that the second term is of order  $O(\eta/N^2)$ . For the first term, we have

$$\begin{aligned} &\eta \mathbb{E}_{x,x'} \left[ \frac{1}{3\|v_{x,x'}\|} \left( I - \frac{v_{x,x'}v_{x,x'}^\top}{\|v_{x,x'}\|^2} \right) u_{g(x')}(e_x + e_{g(x)} + e_{x'})^\top \right] \\ &= \eta \mathbb{E}_{x,x'} \left[ \frac{1}{3\|v_{x,x'}\|} u_{g(x')}e_x^\top \right] - \eta \mathbb{E}_{x,x'} \left[ \frac{\alpha(1 + \delta_{g(x),g(x')})}{9\|v_{x,x'}\|^3} v_{x,x'}(e_x + e_{g(x)} + e_{x'})^\top \right] + O(\eta/N^2) \\ &= \frac{\eta\beta_1}{N} \sum_{x=1}^N u_{g(x)}e_x^\top - \eta \mathbb{E}_{x,x'} [\gamma_{x,x'}v_{x,x'}(e_x + e_{g(x)} + e_{x'})^\top] + O(\eta/N^2), \end{aligned}$$

935 with

$$\beta_1 = \mathbb{E}_x \left[ \frac{1}{3\|v_{x,1}\|} \right] \quad \text{and} \quad \gamma_{x,x'} = \frac{\alpha(1 + \delta_{g(x),g(x')})}{9\|v_{x,x'}\|^3}.$$

936 We have

$$\begin{aligned} &-\eta \mathbb{E}_{x,x'} [\gamma_{x,x'}v_{x,x'}(e_x + e_{g(x)} + e_{x'})^\top] \\ &= -\eta \mathbb{E}_{x'} [\mathbb{E}_{x'} [\gamma_{x,x'}v_{x,x'}|x](e_x + e_{g(x)})^\top] - \eta \mathbb{E}_{x'} [\mathbb{E}_x [\gamma_{x,x'}v_{x,x'}|x']e_{x'}^\top] \\ &= \frac{\eta\beta_2}{N} \sum_{x=1}^N (e_x - \alpha u_{g(x)})(e_x + e_{g(x)})^\top - \frac{\eta\beta_2}{N} \sum_{x=1}^N (2e_x + \alpha u_{g(x)})e_x^\top + O(\eta/N^2) \\ &= -\frac{\eta\beta_2}{N} \sum_{x=1}^N e_x e_x^\top + \frac{\eta\beta_2}{N} \sum_{y=N+1}^{2N} (e_{g^{-1}(y)} - \alpha u_y)e_y^\top + O(\eta/N^2), \end{aligned}$$

937 with

$$\beta_2 = \frac{1}{3}\mathbb{E}_{x'} [\gamma_{1,x'}] = \frac{1}{3}\mathbb{E}_x [\gamma_{x,1}] = \frac{1}{3N}\gamma_{1,1} + \frac{N-1}{3N}\gamma_{1,2}.$$

938 We have thus shown

$$-\eta \nabla L_3^T(W_2) = \frac{\eta}{N} \sum_{x=1}^N (\beta_1 u_{g(x)} - \beta_2 e_x)e_x^\top + \frac{\eta\beta_2}{N} \sum_{y=N+1}^{2N} (e_{g^{-1}(y)} - \alpha u_y)e_y^\top + O(\eta/N^2). \quad (23)$$

939 For false sequences, it can be checked that  $\eta \nabla L_3^F(W_2) = O(\eta/N^2)$ . Thus, taking step-size  $\eta = N/\rho$   
 940 yields

$$\begin{aligned} W_3 &= W_2 - \eta \nabla L_3(W_2) \\ &= (\alpha + \beta_1) \sum_{x=1}^N u_{g(x)}e_x^\top - (1 + \beta_2) \sum_{x=1}^N e_x e_x^\top + \beta_2 \sum_{y=N+1}^{2N} (e_{g^{-1}(y)} - \alpha u_y)e_y^\top + O(1/N). \end{aligned}$$

941

□

<sup>6</sup>Once again, this is only true up to vanishing terms in  $N$ , which we ignore here for simplicity.

## 942 D.2 Proof of Theorem 1

943 Suppose we are given  $(x, y, x')$ , where we assume for simplicity that  $x \neq x'$  and  $g(x') \neq y$ . Denote  
 944 by  $f_W(z_{1:t})$  the output of the model in (2) before applying the LN and the unembedding layer. Then,  
 945 we have that:

$$\begin{aligned} f_W(x, y, x') &= e_{x'} + p_3 + \frac{1}{3}\bar{\gamma} \left( \sum_y u_y - \sum_x u_x \right) + \\ &+ \frac{1}{3} \left( -\alpha_1 e_x + \beta_1 u_{g(x)} + \alpha_2 e_{g^{-1}(y)} - \beta_2 u_y - \alpha_1 e_{x'} + \beta_1 u_{g(x')} \right) \end{aligned} \quad (24)$$

946 Denote by  $c_1 := 2 + \frac{\bar{\gamma}^2(2N-2)+2\alpha_1^2+\beta_1^2}{9}$  and  $c_2 := 2 + \frac{\bar{\gamma}^2(2N-3)+2\alpha_1^2+\beta_1^2}{9}$ . for a true sample where  
 947  $y = g(x)$  we have that:

$$\|f_W(x, g(x), x')\|^2 = c + (\beta_1 - \beta_2 + \bar{\gamma})^2 + (\beta_1 + \bar{\gamma})^2.$$

948 Hence, after applying the LN and unembedding layer we have that:

$$\begin{aligned} (F_W(x, g(x), x'))_{g(x')} &= \frac{\beta_1 + \bar{\gamma}}{3\sqrt{c_1 + (\beta_1 - \beta_2 + \bar{\gamma})^2 + (\beta_1 + \bar{\gamma})^2}} \\ \max_{y' \neq g(x')} (F_W(x, g(x), x'))_{y'} &= \frac{\bar{\gamma} + \max(0, \beta_1 - \beta_2)}{3\sqrt{c_1 + (\beta_1 - \beta_2 + \bar{\gamma})^2 + (\beta_1 + \bar{\gamma})^2}} \end{aligned}$$

949 For a false sample where  $y \neq g(x)$  we have that:

$$\|f_W(x, g(x), x')\|^2 = c_2 + 2(\beta_1 + \bar{\gamma})^2 + (-\beta_2 + \bar{\gamma})^2.$$

950 Hence, after applying the LN and unembedding layer we have that:

$$\begin{aligned} (F_W(x, y, x'))_{g(x')} &= \frac{\beta_1 + \bar{\gamma}}{3\sqrt{c_2 + 2(\beta_1 + \bar{\gamma})^2 + (-\beta_2 + \bar{\gamma})^2}} \\ \max_{y' \neq g(x')} (F_W(x, y, x'))_{y'} &= \frac{\beta_1 + \bar{\gamma}}{3\sqrt{c_2 + 2(\beta_1 + \bar{\gamma})^2 + (-\beta_2 + \bar{\gamma})^2}}. \end{aligned}$$

951 Plugging in these terms finishes the proof.

## 952 D.3 Proof of Theorem 2

953 *Proof.* We first describe the output of the model in (2) before applying LN. Denote by  $v_T, v_F \in$   
 954  $\mathbb{R}^{4N+3}$  these outputs for true and false samples respectively. Recall that a true sample  $(x, y)$  is when  
 955  $y = g(x)$  and false otherwise. Then, we have that:

$$v_T = e_y + p_2 + \frac{1}{2} \left( (\alpha_2 - \alpha_1)e_x + (\beta_1 - \beta_2)u_y + (\gamma_1 - \gamma_2) \cdot \left( \sum_y u_y - \sum_x u_x \right) \right) \quad (25)$$

$$v_F = e_y + p_2 + \frac{1}{2} \left( -\alpha_1 e_x + \alpha_2 u_{g^{-1}(y)} + \beta_1 u_{g(x)} - \beta_2 u_y + (\gamma_1 - \gamma_2) \cdot \left( \sum_y u_y - \sum_x u_x \right) \right) \quad (26)$$

956 We will first show that without adding N the samples above cannot be separated for general  $x$  and  $y$ .

957 Assume otherwise, that there exists a linear separator  $w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{pmatrix}$  with  $w_1, \dots, w_4 \in \mathbb{R}^N, w_5 \in \mathbb{R}^3$

958 and bias term  $b \in \mathbb{R}$  such that  $\langle w, v_T \rangle - b \geq 0$  and  $\langle w, v_F \rangle - b < 0$  for every true or false sample

959 respectively. We slightly abuse notation and write  $\langle w_1, e_x \rangle$  as  $\left\langle \begin{pmatrix} w_1 \\ 0_{3N+3} \end{pmatrix}, e_x \right\rangle$ , and similarly when  
 960 multiplying  $w_2$  by  $e_y$ ,  $w_3$  by  $u_x$ ,  $w_4$  by  $u_y$  and  $w_5$  by  $p_t$ .

$$c := \frac{1}{2} \left\langle (\gamma_1 - \gamma_2) \cdot \left( \sum_y u_y - \sum_x u_x \right), w_3 + w_4 \right\rangle + \langle w_5, p_2 \rangle$$

961 the terms in the inner products that are independent of the sample. Then, using the linear separator on  
 962 these four samples we have:

$$b \leq (\alpha_2 - \alpha_1) \langle e_{x_i}, w_1 \rangle + \langle e_{y_i} w_2 \rangle + (\beta_1 - \beta_2) \langle u_{y_i}, w_4 \rangle + c \quad (27)$$

$$b \leq (\alpha_2 - \alpha_1) \langle e_{x_j}, w_1 \rangle + \langle e_{y_j} w_2 \rangle + (\beta_1 - \beta_2) \langle u_{y_j}, w_4 \rangle + c \quad (28)$$

$$b \geq \alpha_2 \langle e_{x_i}, w_1 \rangle - \alpha_1 \langle e_{x_j}, w_1 \rangle + \langle e_{y_i}, w_2 \rangle + \beta_1 \langle u_{y_j}, w_4 \rangle - \beta_2 \langle u_{y_i}, w_4 \rangle + c \quad (29)$$

$$b \geq \alpha_2 \langle e_{x_j}, w_1 \rangle - \alpha_1 \langle e_{x_i}, w_1 \rangle + \langle e_{y_j}, w_2 \rangle + \beta_1 \langle u_{y_i}, w_4 \rangle - \beta_2 \langle u_{y_j}, w_4 \rangle + c. \quad (30)$$

963 Adding up (29) and (30) we have that:

$$2b - 2c \geq (\alpha_2 - \alpha_1) \langle e_{x_j}, w_1 \rangle + \langle e_{y_j} w_2 \rangle + (\beta_1 - \beta_2) \langle u_{y_j}, w_4 \rangle + \quad (31)$$

$$+ (\alpha_2 - \alpha_1) \langle e_{x_i}, w_1 \rangle + \langle e_{y_i} w_2 \rangle + (\beta_1 - \beta_2) \langle u_{y_i}, w_4 \rangle, \quad (32)$$

964 which is a contradiction to (27) and (28). This means that there is no linear separator, regardless of  
 965 the values of the parameters, which proves the first item.

966 Assume there is layer normalization after the prediction as in (2). This means that the output of the  
 967 model is  $\frac{v}{\|v\|}$ . Consider the linear predictor  $w = p_2$ , and a bias term  $b$  that will be determined later.

968 Then, the output of the linear predictor is exactly  $\langle w, v \rangle = \frac{1}{\|v\|}$ .

969 We will now calculate the norm of both true and false samples. For a true sample  $(x, g(x))$  we have  
 970 that:

$$\|v_T\|^2 = 2 + (\alpha_2 - \alpha_1)^2 + (\gamma_1 - \gamma_2)^2 \cdot (2N - 1) + (\gamma_1 - \gamma_2 + \beta_1 - \beta_2)^2. \quad (33)$$

971 For a negative sample  $(x, y)$  with  $g(x) \neq y$  we have:

$$\|v_F\|^2 = 2 + \alpha_1^2 + \alpha_2^2 + (\gamma_1 - \gamma_2)^2 \cdot (2N - 2) + (\gamma_1 - \gamma_2 + \beta_1)^2 + (\gamma_1 - \gamma_2 - \beta_2)^2. \quad (34)$$

972 There exists a linear separator as long as  $\frac{1}{\|v_F\|} - \frac{1}{\|v_T\|} \neq 0$ . Since the vectors  $v_T$  and  $v_F$  are both  
 973 non-zero, this is equivalent to  $\|v_T\|^2 \neq \|v_F\|^2$ . By the above calculation, we have that:

$$\begin{aligned} & \|v_F\|^2 - \|v_T\|^2 \\ &= \alpha_1^2 + \alpha_2^2 - (\alpha_1 - \alpha_2)^2 - (\gamma_1 - \gamma_2)^2 + (\gamma_1 - \gamma_2 + \beta_1)^2 + (\gamma_1 - \gamma_2 - \beta_2)^2 - (\gamma_1 - \gamma_2 + \beta_1 - \beta_2)^2 \\ &= 2\alpha_1\alpha_2 + 2\beta_1\beta_2. \end{aligned}$$

974 This shows that if  $2\alpha_1\alpha_2 + 2\beta_1\beta_2 \neq 0$  then we have a linear separation between true and false  
 975 samples.

976 Further assuming that  $\alpha_1 = \alpha_2$ ,  $\beta_1 = \beta_2$ ,  $\gamma_1 = \gamma_2$  we have that  $\|v_T\|^2 = 2$  and  $\|v_F\|^2 =$   
 977  $2 + 2\alpha_2 + 2\beta_2$ . To find the optimal margin for this predictor we pick:

$$b = \frac{1}{2} \cdot \left( \frac{1}{\|v_T\|} - \frac{1}{\|v_F\|} \right) = \frac{1}{2\sqrt{2}} \left( 1 - \frac{1}{\sqrt{1 + \alpha^2 + \beta^2}} \right).$$

978 We will now prove that there is linear separation after predicting the  $x'$  token. Using the output of the  
 979 model as in (2) we get:

$$v_T = C + \frac{1}{3} ((\alpha_2 - \alpha_1)e_x + (\beta_1 - \beta_2)u_y - \alpha_1 e_{x'} + \beta_1 u_{g(x')}) \quad (35)$$

$$v_F = C + \frac{1}{3} (-\alpha_1 e_x + \alpha_2 u_{g^{-1}(y)} + \beta_1 u_{g(x)} - \beta_2 u_y - \alpha_1 e_{x'} + \beta_1 u_{g(x')}) , \quad (36)$$

980 where  $C = e_{x'} + p_3 + \frac{\hat{\gamma}}{3} \cdot (\sum_y u_y - \sum_x u_x)$ . We can now calculate:

$$\|v_T\|^2 = 2 + \frac{1}{9} ((\alpha_2 - \alpha_1)^2 + (\beta_1 - \beta_2 + \bar{\gamma})^2 + \alpha_1^2 + (\beta_1 + \bar{\gamma})^2 + (2N - 2)\bar{\gamma}^2) \quad (37)$$

$$\|v_F\|^2 = 2 + \frac{1}{9} (2\alpha_1^2 + \alpha_2^2 + 2(\beta_1 + \bar{\gamma})^2 + (\bar{\gamma} - \beta_2)^2 + (2N - 3)\bar{\gamma}^2) . \quad (38)$$

981 We now have that:

$$\begin{aligned} \|v_F\|^2 - \|v_T\|^2 &= \frac{1}{9} \cdot (\alpha_1^2 + \alpha_2^2 + (\beta_1 + \bar{\gamma})^2 + (\bar{\gamma} - \beta_2)^2 - (\alpha_2 - \alpha_1)^2 - (\beta_1 - \beta_2 + \bar{\gamma})^2 - \bar{\gamma}^2) \\ &= \frac{2}{9} (\alpha_1 \alpha_2 + \beta_1 \beta_2) . \end{aligned}$$

982 By a similar argument to the previous case, if  $\alpha_1 \alpha_2 + \beta_1 \beta_2 \neq 0$  then there is linear separation  
 983 between true and false samples. Further assuming that  $\alpha_1 = \alpha_2$ ,  $\beta_1 = \beta_2$  and  $\bar{\gamma} = 0$ , to find the  
 984 optimal margin for the predictor we pick:

$$b = \frac{1}{2} \cdot \left( \frac{1}{\|v_T\|} - \frac{1}{\|v_F\|} \right) = \frac{\alpha^2 + \beta^2}{9\sqrt{4 + \frac{8}{9}(\alpha^2 + \beta^2) + \frac{1}{27}(\alpha^2 + \beta^2)^2}} .$$

985

□