

Eigenvalue Calibration for Semantic Embeddings of Large Language Models

Sebastian G. Gruber^{†1}

Nassim Walha^{†2,3,4}

Francis Bach⁶

Florian Buettner^{2,3,4,5}

¹ESAT-PSI, KU Leuven, Belgium

²German Cancer Research Center (DKFZ), Heidelberg, Germany

³German Cancer Consortium (DKTK), Germany

⁴Goethe University Frankfurt, Germany

⁵Frankfurt Cancer Institute, Germany

⁶PSL Research University / Inria, France.

Abstract

Uncertainty quantification is central to the reliable deployment of large language models (LLMs), and eigenvalues of semantic embeddings have recently emerged as a key tool in state-of-the-art methods. However, conventional calibration results developed for classification probabilities cannot be directly transferred to eigenvalues. We address this gap by proposing a novel framework for calibrating the eigenvalues of semantic embeddings. We interpret LLMs combined with semantic embeddings of their generated answers as density matrix predictors, and we propose a novel approach to calibrate density matrix predictors by applying temperature scaling to their eigenvalues. We establish entropy–risk equivalence under calibration, derive a central calibration inequality specific to eigenvalues, and prove that temperature-scaled eigenvalues optimize calibration when minimizing proper score risks. Experiments on a variety of real-world settings show that current LLMs are systematically overconfident, and validate our theoretical findings. Together, these results advance the foundations and practice of uncertainty quantification for semantic embeddings.

1 INTRODUCTION

Uncertainty quantification has become a cornerstone for assessing the reliability of modern machine learning models, particularly large language models (LLMs) [Shorinwa et al., 2025]. In high-stakes applications, calibrated uncertainty estimates are essential for downstream decision-making, model comparison, and human–AI collaboration

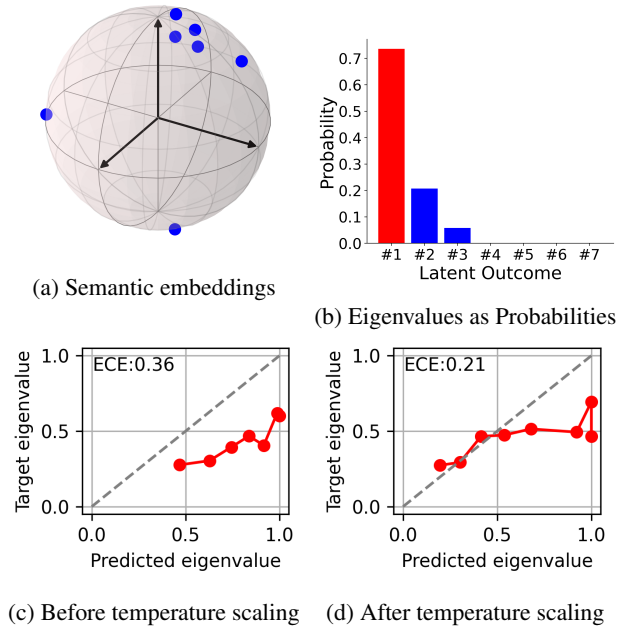


Figure 1: Normalised semantic embeddings reside in a hypersphere (Figure 1a). We interpret the eigenvalues of the respective density matrix as probabilities of latent outcomes (Figure 1b). Large language models are overconfident in their predicted maximum eigenvalue (Figure 1c), which can be adjusted via temperature scaling (Figure 1d) resulting in a lower expected calibration error (ECE), and, thus, more reliable uncertainties.

[†]Equal contribution. Alphabetical order.

[Silva Filho et al., 2023, Maier-Hein et al., 2024]. Recent advances have shown that semantic embeddings of LLM outputs provide a powerful basis for uncertainty quantification, enabling fine-grained measures of predictive confidence [Gruber and Buettner, 2024, Nikitin et al., 2024, Walha et al., 2026]. In particular, the eigenvalues constructed from embeddings capture uncertainty through entropy quantities, and have already been adopted in state-of-the-art LLM uncertainty quantification methods [Nikitin et al., 2024, Walha et al., 2026].

Despite this progress, a fundamental gap remains: conventional calibration results developed for classification probabilities cannot be directly transferred to eigenvalues. In classification, calibration aligns predicted probabilities with empirical frequencies, providing interpretable confidence estimates [Murphy, 1973]. In contrast, the eigenvalues of density matrices (constructed from embeddings) encode latent outcome probabilities in a different mathematical space, raising the question of how calibration should be defined, analyzed, and optimized in this setting. Addressing this gap is essential to ensure that embedding-based uncertainty quantification methods are both theoretically principled and practically reliable.

In this paper, we provide the first comprehensive study of eigenvalue calibration for density matrix predictors in general and semantic embeddings of LLMs in particular. To connect our theory with practice, we interpret LLMs as density matrix predictors and evaluate our framework extensively on real-world settings. Figure 1 presents an overview of our setup, and a motivating example from our evaluations is given as follows. For the TriviaQA question “What links do Bollywood, Hollywood and Lollywood have?”, the Phi-4 Mini model generates 20 candidate responses. Among these, 8 correspond to the correct answer “film-making industry”, while the remaining 12 incorrectly focus on “film-making in the Indian subcontinent.” Because this incorrect response is the most frequent one and semantically similar to the correct generations, the model exhibits overconfidence, with a maximum eigenvalue of the predicted density matrix $\lambda_{\max} = 0.81$. This is the predicted eigenvalue plotted on the x-axis of Figure 1c and Figure 1d. After applying temperature calibration with the optimal value $T = 2.51$, the predicted eigenvalue is reduced to $\lambda_{\max} = 0.29$, mitigating overconfidence and more accurately reflecting the unreliability of the predicted answer.

In summary, our **contributions** are as follows:

- We introduce a novel notion of calibration suitable for density matrix predictors, including eigenvalue-based uncertainty quantification via semantic embeddings in Section 3. Specifically, calibration is required such that the average entropy predicts the model risk.
- We theoretically and empirically demonstrate how proper scores and temperature scaling can be used

to optimise the calibration of eigenvalues, and show how current large language models are systematically overconfident in Section 4.

- We propose a novel approach to plot reliability diagrams for transparent insights into the calibration of eigenvalues, which verifies our other results in Section 5.

2 BACKGROUND

In this section, we give an overview of the background necessary to state our contributions. Throughout this work, we assume a supervised learning setup, with input-target random variables (X, Y) following a joint distribution $\mathbb{P}_{XY}$ over a sample space $\mathcal{X} \times \mathcal{Y}$.

First, we give an introduction to uncertainty calibration, followed by proper scores, density matrices, and Bregman matrix divergences.

Calibration. Uncertainty calibration in classification is crucial for real-world applications [Silva Filho et al., 2023]. We define a model $f: \mathcal{X} \rightarrow \mathcal{P}$ as **canonically calibrated** based on a joint distribution $\mathbb{P}_{XY}$ and a set $\mathcal{P}$ of potential target distributions if and only if

$$\mathbb{P}(Y \mid f(X)) \stackrel{a.s.}{=} f(X), \quad (1)$$

where *a.s.* refers to *almost surely* [Vaicenavicius et al., 2019]. Note that canonical calibration is also defined beyond classification under a more general set of distributions $\mathcal{P}$ and a general target space $\mathcal{Y}$ [Gruber and Buettner, 2022]. In Section 3, we introduce matrix and eigenvalue calibration arising from canonical calibration.

Proper Scores. A loss function of the form $S: \mathcal{P} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ for which holds that

$$\mathbb{E}_{Y \sim Q} [S(P, Y)] \geq \mathbb{E}_{Y \sim Q} [S(Q, Y)], \quad \forall P, Q \in \mathcal{P}, \quad (2)$$

with respect to a set of distributions $\mathcal{P}$ is referred to as proper scoring rule, or, in short, **proper score** [Gneiting and Raftery, 2007]. It is known that the respective **entropy function** $H_S: \mathcal{P} \rightarrow \mathbb{R}$ defined via $H_S(P) := \mathbb{E}_{Y \sim P} [S(P, Y)]$ is concave. A common example is the log score, which leads to the Shannon entropy as entropy function [Gneiting and Raftery, 2007].

Density matrices and eigenvalues. We use density matrices (also referred to as density operators), which are positive semi-definite matrices with unit trace [Wilde, 2013]. A matrix M is said to be positive semi-definite (p.s.d.) if and only if for all $a \in \mathbb{R}^d$ holds $a^\top M a \geq 0$. Therefore, we define the space of all p.s.d. $d \times d$ matrices via $\mathbb{H}_d := \{M \in \mathbb{R}^{d \times d} \mid M \text{ p.s.d.}\}$. The respective space of density matrices is then defined by $\mathbb{H}_d^\Delta := \{D \in \mathbb{H}_d \mid \text{tr } D = 1\}$.

For any density matrix in $\mathbb{H}_d^\Delta$ holds that its eigenvalues are elements of the simplex $\Delta_d := \{p \in \mathbb{R}_{\geq 0}^d \mid \sum_i p_i = 1\}$. We also make use of spectral functions of the form $\mathbf{f}: \mathbb{H}_d \rightarrow \mathbb{R}^{d \times d}$, which are defined based on a function $f: \mathbb{R} \rightarrow \mathbb{R}$ via $\mathbf{f}(M) := \sum_{i=1}^d f(\lambda_i) e_i e_i^\top$, where $\lambda_1, \dots, \lambda_d$ and $e_1, \dots, e_d$ are the eigenvalues and eigenvectors of M . The maximum eigenvalue of a density matrix $D \in \mathbb{H}_d^\Delta$ is equal to its spectral norm $\|\cdot\|_2$, i.e., $\lambda_{\max}(D) = \|D\|_2$ [Boyd and Vandenberghe, 2004]. Thus, computing the maximum eigenvalue is a convex operation, which is relevant for Section 3.

In classification, given a one-hot encoded random variable Y_{oh} , we can recover its (marginal) distribution as a probability vector $\mathbb{P}_{Y_{\text{oh}}}$ via

$$\mathbb{P}_{Y_{\text{oh}}} = \mathbb{E}[Y_{\text{oh}}] \in \Delta_d. \quad (3)$$

Given another random variable X , its conditional distribution is recovered via

$$\mathbb{P}_{Y_{\text{oh}}|X} = \mathbb{E}[Y_{\text{oh}} \mid X], \quad (4)$$

which maps into the simplex Δ_d . In a similar way, we can construct marginal and conditional density matrices based on random variables residing in a hypersphere [Wilde, 2013]. The **(marginal) density matrix** of a random variable $\mathbf{Y}$ with outcomes in the d -dimensional hypersphere $S_d := \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ is given by

$$\mathbb{D}_{\mathbf{Y}} := \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] \in \mathbb{H}_d^\Delta. \quad (5)$$

Note that, generally, $\mathbb{D}_{\mathbf{Y}}$ does not characterize the distribution of $\mathbf{Y}$. Further, its **conditional density matrix** with an additional random variable X is defined as

$$\mathbb{D}_{\mathbf{Y}|X} := \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top \mid X], \quad (6)$$

which maps into $\mathbb{H}_d^\Delta$. Similar to probability vectors, it also holds for density matrices that $\mathbb{E}[\mathbb{D}_{\mathbf{Y}|X}] = \mathbb{D}_{\mathbf{Y}}$.

Since the eigenvalues of a density matrix are non-negative and sum up to one, they can be interpreted as probabilities of non-observable (i.e. “latent”) pure states in quantum mechanics [Wilde, 2013]. In the context of semantic embeddings, *we propose to interpret eigenvalues as probabilities of semantic latent outcomes* (cf. Figure 1a and 1b). For example, “It’s Paris.” and “The answer is Paris.” are literally different answers but they are semantically the same (latent) answer covered by the same eigenvalue.

Bregman matrix divergences. It is a known fact that proper scores induce Bregman divergences [Ovcharov, 2018]. In this work, we offer the first formal definition of proper scores defined for density matrices, which, as we show in a later section, induce Bregman matrix divergences as defined in [Kulis et al., 2009]. Given a convex and differentiable function $\phi: \mathbb{H}_d \rightarrow \mathbb{R}$, the generated **Bregman matrix divergence** $\text{Div}_\phi: \mathbb{H}_d \times \mathbb{H}_d \rightarrow \mathbb{R}_{\geq 0}$ is given by

$$\text{Div}_\phi(x, y) := \phi(x) - \phi(y) - \text{tr}[\nabla\phi(y)(x - y)^\top]. \quad (7)$$

We can now present the contributions of our work. We start with a novel notion of calibration regarding density matrices and its eigenvalues.

3 EIGENVALUE CALIBRATION

To connect our theoretical contributions with practical applications, it is of central importance to identify LLMs in combination with semantic embeddings as density matrix predictors.

LLMs as density matrix predictors. Let $\mathcal{X}_{\text{text}}$ be the space of language and $\mathcal{P}_{\text{text}}$ be a set of distributions with support $\mathcal{X}_{\text{text}}$. Let $f: \mathcal{X}_{\text{text}} \rightarrow \mathcal{P}_{\text{text}}$ be an LLM, which allows to sample answers $a \sim f(x)$ given an input query $x \in \mathcal{X}_{\text{text}}$, and let $e: \mathcal{X}_{\text{text}} \rightarrow \mathbb{R}^d$ be a semantic embedding model. From here on, we use the associated density matrix predictor $d: \mathcal{X}_{\text{text}} \rightarrow \mathbb{H}_d^\Delta$ given by

$$d(x) := \mathbb{E}_{a \sim f(x)} [e(a) e(a)^\top]. \quad (8)$$

In essence, the density matrix predictor is simply the expected outer product of the answers’ semantic embeddings. In practice, the predictor d can be approximated via the estimator $\hat{d}(x) := \frac{1}{m} \sum_{i=1}^m e(a_i) e(a_i)^\top$ given sampled answers $a_1, \dots, a_m \sim f(x)$.

Similarly to Figure 1b, Walha et al. [2026] interpret the eigenvalues of $\hat{d}$ as probabilities over different latent semantic outcomes within the sampled answers. In the following, we motivate this interpretation: The first eigenvector of $\hat{d}$ is the 1-D semantic direction that minimizes residual variance when the embedded answers are projected onto it, i.e., it represents the dominant latent semantic direction. The corresponding eigenvalue measures the squared aggregate mass of the answers along that direction, so it is naturally interpretable as confidence concentrated on a latent semantic mode. Further eigenvectors are built orthogonally to previous ones, which matches the orthogonality of latent semantic directions. Two extremes make this concrete: if all sampled answers are semantically identical, then $\lambda_{\max} = 1$ and all other eigenvalues are 0, which maximizes confidence; if they are mutually orthogonal, then all non-zero eigenvalues are equal with $\lambda_{\max} = \lambda_1 = \lambda_2 = \dots = \lambda_m = \frac{1}{m}$, thus maximizing entropy.

Definition 1 (Matrix and Eigenvalue Calibration). *We say a density matrix predictor d is **matrix calibrated** w.r.t. the joint distribution $\mathbb{P}_{XY}$ if*

$$\mathbb{D}_{Y|d(X)} \stackrel{a.s.}{=} d(X), \quad (9)$$

*and it is **eigenvalue calibrated** if*

$$\lambda_{\max}(\mathbb{D}_{Y|\Lambda_X}) \stackrel{a.s.}{=} \Lambda_X, \quad (10)$$

with $\Lambda_X := \lambda_{\max}(d(X))$.

In other words, similar to when we say a model is canonically calibrated when its predicted distribution matches the target distribution (given the prediction), we say a density matrix predictor is matrix calibrated if its predicted matrix matches the target density matrix. For eigenvalue calibration, we condition on the predicted maximum eigenvalue. However, even though eigenvalue calibration seems simpler than matrix calibration, we argue based on our following results that matrix calibration is more meaningful in practice.

Following the previous definitions, we can create a relationship between the different calibration notions.

Theorem 1 (Eigenvalue Calibration Inequality). *Let f be an LLM and d its respective density matrix predictor. If f is canonically calibrated then d is matrix calibrated. Further, if d is matrix calibrated, then*

$$\lambda_{\max}(\mathbb{D}_{Y|d(X)}) \stackrel{a.s.}{=} \Lambda_X, \quad (11)$$

with $\Lambda_X := \lambda_{\max}(d(X))$, but also

$$\lambda_{\max}(\mathbb{D}_{Y|\Lambda_X}) \stackrel{a.s.}{\leq} \Lambda_X. \quad (12)$$

The proof is located in Appendix C and makes use of Gruber [2024, Theorem 1] in combination with the convexity of the spectral norm. Theorem 1 clarifies that matrix calibration follows from canonical calibration, similar to other calibration notions in classification [Gruber and Buettner, 2022, Gupta and Ramdas, 2022]. However, it also offers the surprising fact that eigenvalue calibration does not follow, which impacts our ability to compute reliability diagrams for calibration diagnostics. Therefore, in a later section, we propose to rather use Equation 11 to compute meaningful reliability diagrams.

We now introduce proper matrix scores, which connect matrix calibration with risk minimization.

Definition 2. A function $\mathbf{S}: \mathbb{H}_d^\Delta \rightarrow \mathbb{R}^{d \times d}$ is defined to be a *proper matrix score* if

$$\text{tr}[\mathbf{S}(M_{\text{pred}}) M_{\text{target}}] \geq \text{tr}[\mathbf{S}(M_{\text{target}}) M_{\text{target}}] \quad (13)$$

for all $M_{\text{pred}}, M_{\text{target}} \in \mathbb{H}_d^\Delta$.

Our definition of proper matrix scores is a subclass of proper scores, which is usable for semantic embeddings, as the following shows.

Proposition 1. A proper matrix score $\mathbf{S}: \mathbb{H}_d^\Delta \rightarrow \mathbb{R}^{d \times d}$ generates a proper score $S: \mathcal{P}_d \times \mathcal{Y}_d \rightarrow \mathbb{R}$ with respect to the set of distributions $\mathcal{P}_d := \{P \mid \int_{\mathcal{S}_d} \|x\|_2^2 dP(x) < \infty\}$ defined on the hypersphere and $\mathcal{Y}_d = \mathbb{R}^d$. The respective proper score is given by

$$S(P, y) := y^\top \mathbf{S}(\mathbb{E}_{x \sim P}[xx^\top]) y. \quad (14)$$

The proof is presented in Appendix C.

The respective **matrix entropy function** $H_S: \mathbb{H}_d^\Delta \rightarrow \mathbb{R}$ is given by

$$H_S(M) := \text{tr}[S(M) M]. \quad (15)$$

Further, we can define the associated **divergence function** $D_S: \mathbb{H}_d^\Delta \times \mathbb{H}_d^\Delta \rightarrow \mathbb{R}_{\geq 0}$ via

$$D_S(M_1, M_2) := \text{tr}[\mathbf{S}(M_2) M_1 - \mathbf{S}(M_1) M_1]. \quad (16)$$

We offer the following statements regarding associated entropy and divergence functions.

Proposition 2. Given a proper matrix score $\mathbf{S}$, it holds that H_S is concave. Further, if H_S is also differentiable, then

$$D_S = \text{Div}_{-H_S}. \quad (17)$$

This follows from the fact that a proper matrix score is a proper score, for which Ovcharov [2018] shows similar results, and since a density matrix is constructed from a distribution in a linear manner (cf. Appendix C). This result connects our definition of proper matrix scores to the already existing definition of Bregman matrix divergences.

Further, since we assume a supervised learning setup with joint distribution $\mathbb{P}_{X\mathbf{Y}}$ and density matrix predictor d , we are interested in the respective **risk** given a proper matrix score $\mathbf{S}$, which we define via

$$\mathcal{R}_S(d) := \mathbb{E}[\mathbf{Y}^\top \mathbf{S}(d(X)) \mathbf{Y}]. \quad (18)$$

The special case of a proper matrix score, which we use in our experiments, is the matrix version of the log score, also known as cross entropy loss. The matrix log score is given by $\mathbf{S}_{\log}(M) := -\log(M)$, where $\log$ is used as a spectral function. It follows that the associated entropy function is $H_{\mathbf{S}_{\log}}(M) = -\sum_{i=1}^d \lambda_i \log \lambda_i$, where $\lambda_1, \dots, \lambda_d$ are the eigenvalues of $M \in \mathbb{H}_d^\Delta$. This function is also known as von Neumann entropy [Wilde, 2013], and is essentially the Shannon entropy of the respective eigenvalues. It is used for state-of-the-art results in LLM uncertainty quantification [Nikitin et al., 2024]. The related divergence function is given by $D_{\mathbf{S}_{\log}}(M_1, M_2) = \text{tr}[\log(M_1) M_1 - \log(M_2) M_1]$, which recovers the quantum relative entropy [Wilde, 2013]. The matrix log score will be used in our experiments due to its prominence.

We now offer a central result, which connects risk and entropy under matrix calibration.

Lemma 1. Given a proper matrix score $\mathbf{S}$ and a density matrix predictor d , which is matrix calibrated for a joint distribution $\mathbb{P}_{X\mathbf{Y}}$, it holds that

$$\underbrace{\mathbb{E}[H_S(d(X))]}_{\text{Expected Uncertainty}} = \underbrace{\mathcal{R}_S(d)}_{\text{Risk}}. \quad (19)$$

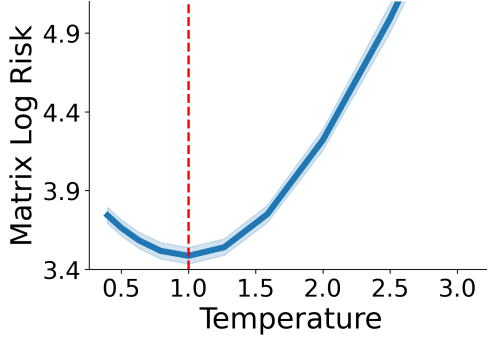


Figure 2: The matrix version of the cross entropy risk correctly determines the ground truth temperature (red line) for $n = 100$ real-world semantic embeddings.

The proof is located in Appendix C. Lemma 1 shows the importance of matrix calibration for uncertainty quantification. It indicates that under calibration, the expected uncertainty given by the entropy function is an accurate representation of the respective risk. However, exact calibration is infeasible in practice. Therefore, we offer the following result, which proves that Lemma 1 may also be achieved in an approximate manner.

Theorem 2 (Entropy-Risk Convergence). *Let $d_1, d_2, \dots$ be a sequence of density matrix predictors such that $\lim_{n \rightarrow \infty} d_n$ is matrix calibrated. Then, for a continuous proper matrix score $\mathbf{S}$, it holds*

$$\underbrace{\mathbb{E}[H_{\mathbf{S}}(d_n(X))]}_{\text{Expected Uncertainty}} - \underbrace{\mathcal{R}_{\mathbf{S}}(d_n)}_{\text{Risk}} \rightarrow 0 \quad (20)$$

for $n \rightarrow \infty$.

We have now established matrix calibration and its relevance for uncertainty quantification via semantic embeddings. Next, we are proposing methodology, which allows to optimize matrix calibration via risk minimization in a post-hoc manner, similar to well-established procedures in classification [Gruber and Buettner, 2022].

4 OPTIMISING MATRIX CALIBRATION

In classification, it is a common procedure to transform a classifier’s predicted probabilities via a simple transformation to improve its calibration [Guo et al., 2017]. If the transformation is injective and the optimization objective is a proper score-based risk, then a respective calibration error is provably optimized [Gruber and Buettner, 2022]. However, this result cannot be readily transferred to our setup since the injectivity needs to hold from distribution space to distribution space. Transforming distributions into density matrices is not injective in general. This requires us to propose an analogous result, which changes the perspective from distributions towards density matrices.

To achieve this, we first translate the essential calibration-sharpness decomposition of proper scores introduced by Bröcker [2009] to proper matrix scores.

Lemma 2 (Calibration-Sharpness Decomposition). *Let $\mathbf{S}$ be a proper matrix score and d be a density matrix predictor for a joint distribution $\mathbb{P}_{X\mathbf{Y}}$. Assuming all integrals are finite, it holds*

$$\underbrace{\mathcal{R}_{\mathbf{S}}(d)}_{\text{Risk}} = \underbrace{\text{Cal}_{\mathbf{S}}(d)}_{\text{Calibration}} - \underbrace{\mathbb{E}[D_{\mathbf{S}}(\mathbb{D}_{\mathbf{Y}|d(X)}, \mathbb{D}_{\mathbf{Y}})]}_{\text{Sharpness}} + \underbrace{H_{\mathbf{S}}(\mathbb{D}_{\mathbf{Y}})}_{\text{Noise}}, \quad (21)$$

with the respective calibration error being defined by

$$\text{Cal}_{\mathbf{S}}(d) := \mathbb{E}[D_{\mathbf{S}}(\mathbb{D}_{\mathbf{Y}|d(X)}, d(X))]. \quad (22)$$

Since the calibration error $\text{Cal}_{\mathbf{S}}$ is induced by a proper score, we follow Gruber and Buettner [2022] and refer to it as proper calibration error (cf. Appendix C for details and proof).

Theorem 3 (Calibration Optimization via Risk). *Given a proper matrix score $\mathbf{S}: \mathbb{H}_d^{\Delta} \times \mathcal{Y}^d \rightarrow \mathbb{R}$, a density matrix predictor $d: \mathcal{X} \rightarrow \mathbb{H}_d^{\Delta}$ and an injective function $h: \mathbb{H}_d^{\Delta} \rightarrow \mathbb{H}_d^{\Delta}$, it holds*

$$\underbrace{\mathcal{R}_{\mathbf{S}}(h \circ d) - \mathcal{R}_{\mathbf{S}}(d)}_{\text{Change in Risk}} = \underbrace{\text{Cal}_{\mathbf{S}}(h \circ d) - \text{Cal}_{\mathbf{S}}(d)}_{\text{Change in Calibration}}. \quad (23)$$

Gruber and Buettner [2022] show that an injective transformation of a model output changes exclusively the calibration term of a proper score. Theorem 3 translates this result to density matrices.

Proposition 3 (Matrix Temperature Scaling). *The function $h_{\text{TS}}: \mathbb{H}_d^{\Delta} \rightarrow \mathbb{H}_d^{\Delta}$ defined via*

$$h_{\text{TS}}(M) := \frac{1}{\text{tr } M^{\alpha}} M^{\alpha} = \frac{1}{\sum_{i=1}^d \lambda_i^{\alpha}} \sum_{i=1}^d \lambda_i^{\alpha} e_i e_i^{\top} \quad (24)$$

is injective for $\alpha > 0$.

The proof for injectivity is given in Appendix C. The exponent α is used as a spectral function, which makes it act on the eigenvalues similar to how conventional temperature scaling transforms probability vectors [Guo et al., 2017]. We refer to $\frac{1}{\alpha}$ as the temperature (parameter) based on convention in the literature [Guo et al., 2017, Kull et al., 2019, Tomani et al., 2022].

Experiments. To verify our theory in practice, we conduct experiments for real-world settings of modern open source LLMs and question answering tasks. All source code for our experiments is available at https://github.com/MLO-lab/matrix_eigenvalue_calibration.

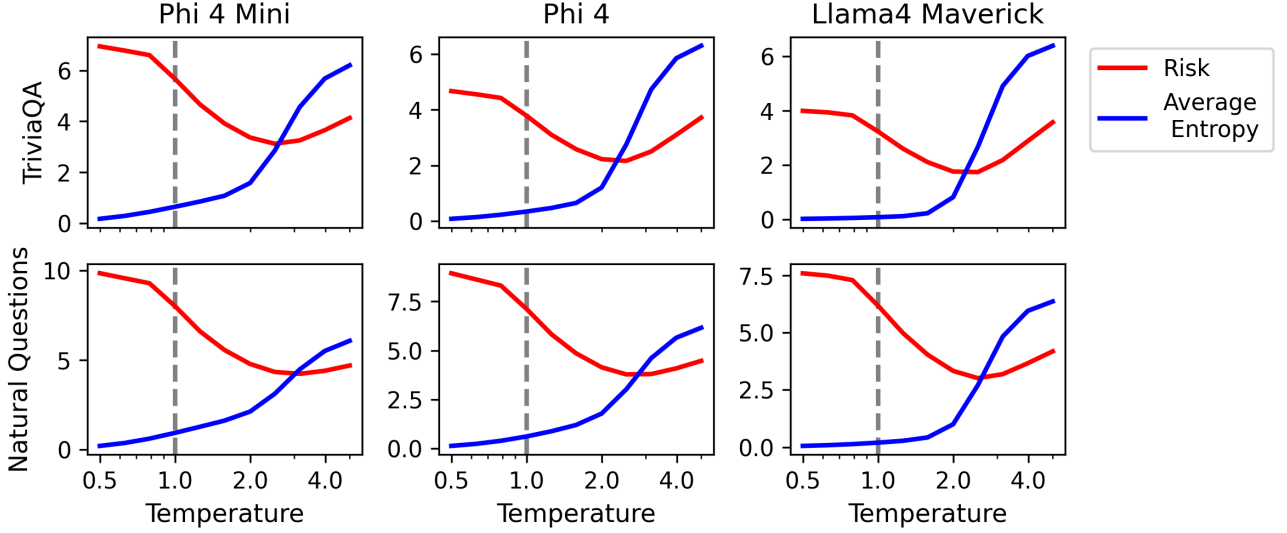


Figure 3: According to Theorem 3, the optimal temperature for calibration is indicated via the minimum risk. Here, the average entropy matches the empirical risk at the optimal temperature across all settings. This indicates that temperature scaling is sufficient to yield well-calibrated eigenvalues such that Theorem 2 holds. Further, in all settings, the optimal temperature is greater than one (grey dashed line). Thus, all models are systematically overconfident and require adjustment.

Table 1: AUROC values ($\uparrow$) based on fuzzy answer correctness with maximum eigenvalue and entropy as uncertainty scores before and after temperature scaling. Using the optimal temperature based on risk minimization improves the AUROC in most of the cases.

Dataset	Model	Eigenvalue	Eigenvalue TS	Entropy	Entropy TS
Natural Questions	Llama4 Maverick	0.66 ± 0.001	0.67 ± 0.001	0.668 ± 0.001	0.663 ± 0.001
	Phi 4	0.742 ± 0.001	0.794 ± 0.0	0.764 ± 0.001	0.795 ± 0.001
	Phi 4 Mini	0.768 ± 0.001	0.778 ± 0.001	0.776 ± 0.001	0.774 ± 0.001
TriviaQA	Llama4 Maverick	0.69 ± 0.001	0.688 ± 0.001	0.684 ± 0.001	0.685 ± 0.001
	Phi 4	0.814 ± 0.001	0.82 ± 0.001	0.818 ± 0.001	0.819 ± 0.001
	Phi 4 Mini	0.808 ± 0.0	0.814 ± 0.001	0.814 ± 0.0	0.805 ± 0.001

We conduct experiments with the TriviaQA [Joshi et al., 2017] and Natural Questions [Kwiatkowski et al., 2019] question answering datasets. As answering LLMs, we use the open source models Phi 4 [Abdin et al., 2024], Phi 4 Mini [Abouelenin et al., 2025], and Llama4 Maverick [AI, 2025]. As semantic embedding model, we use all-mpnet-base-v2 [Song et al., 2020].

For each dataset, we randomly sample subsets from the validation split to conduct our experiments. A development subset of 300 examples per dataset is reserved for temperature scaling, where the temperature parameter is optimized via matrix log risk (see Figure 3). Once the optimal temperature is determined, evaluation is performed on separate subsets of approximately 1700 examples each. These evaluation subsets are used to construct reliability diagrams and

compute the AUROC scores. For each input x , we generate $m = 20$ candidate responses from the LLM and embed them to estimate the corresponding density matrix $d(x)$. Following prior work on LLM uncertainty [Kuhn et al., 2023, Walha et al., 2026], responses are sampled at temperature 0.5 to balance diversity and accuracy. More details are given in Appendix A.

First, we verify that the matrix version of the cross entropy loss $S_{\log}$ does indeed find the optimal temperature. For this, we use $n = 100$ random samples from the TriviaQA dataset and construct a constant predictor with $\frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i \mathbf{Y}_i^T$ from the target semantic embeddings. A change in temperature for this predictor should increase the risk since it is constructed directly from the target samples. As can be seen in Figure 2, this is indeed the case, which confirms that temperatures too

high and too low are indicated via an increased empirical risk.

Next, we conduct temperature scaling on the predicted density matrices of the LLMs. We compare the empirical risk of the log score with the average entropy across a variety of temperature parameters. The results across all settings are shown in Figure 3.

First, it is predicted by Theorem 2 that the expected uncertainty quantified via the entropy is a predictor of the model risk, if the model converges towards matrix calibration. This prediction is precisely observed in all settings since the line of the average entropy and the risk touch at the minimum risk. Further, Theorem 3 states that the temperature with the smallest risk is also the temperature for optimal calibration. Therefore, Figure 3 shows that both of these theoretical findings support each other in practice and empirically verifies our theory in a practical setting.

The second finding is that the risk-optimal temperatures are all above one. This concludes that the models are systematically overconfident across all settings. Therefore, it is likely that overconfidence is a common occurrence in modern LLMs, which requires adjustment for reliable uncertainties, especially if we want the average entropy (as used in state-of-the-art methods; cf. [Nikitin et al., 2024]) to predict the model’s risk.

To further validate our theoretical findings, we compare the risk $\mathcal{R}_S$ and the calibration error Cal_S , as the matrix scaling temperature is increased. The plots presented in Figure 9 (Appendix B) confirm that both quantities are minimized at the same scaling temperature across all models and datasets, as predicted by Theorem 3.

Further, we are also interested in the implications of matrix calibration regarding downstream tasks. One example is the usage of uncertainty scores for detecting answer correctness, which is evaluated via the AUROC [Kuhn et al., 2023, Gruber and Buettner, 2024, Nikitin et al., 2024, Walha et al., 2026]. Since there is no theoretical connection between calibration and AUROC, it is possible that calibration could have a negative impact on the downstream task. We compare the AUROC values in all settings before and after temperature scaling via risk minimization in Table 1. As can be seen, even though we do not optimize temperature scaling towards the AUROC, improving matrix calibration has mostly a beneficial effect on detecting the model’s answer correctness. Note that other state-of-the-art uncertainty measures for LLMs such as semantic entropy [Kuhn et al., 2023] and kernel language entropy [Nikitin et al., 2024] cannot be affected by matrix temperature scaling, as they are independent of the predicted density matrix and depend only on the generated answers directly, which is orthogonal to our approach. As shown in [Walha et al., 2026], these approaches are outperformed by the entropy baseline. We further evaluate both approaches on TriviaQA in our

setting and report the results in Table 2. Indeed, the density-matrix-based entropy and maximum eigenvalue approaches outperform semantic entropy and kernel language entropy, further supporting the claim that these approaches achieve competitive, state-of-the-art performance. This motivates the need for a principled calibration method tailored to them.

Table 2: AUROC values ($\uparrow$) based on fuzzy answer correctness for Kernel Language Entropy (KLE) and Semantic Entropy (SE) on TriviaQA.

Model	KLE	SE
Llama4 Maverick	0.674 $\pm$ 0.001	0.673 $\pm$ 0.001
Phi 4	0.807 $\pm$ 0.001	0.798 $\pm$ 0.001
Phi 4 Mini	0.797 $\pm$ 0.000	0.789 $\pm$ 0.001

Finally, we evaluate the effect of matrix temperature scaling on classical correctness calibration. The results are presented in Appendix B.3.

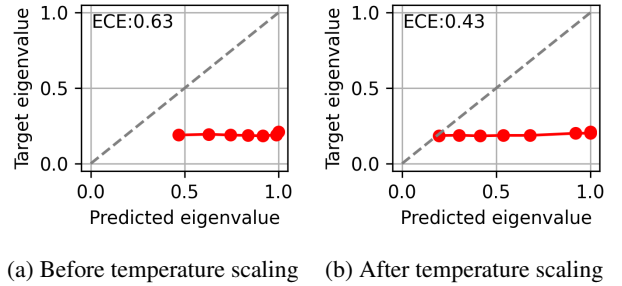


Figure 4: Reliability diagrams according to eigenvalue calibration as in Eq. 10. The “plateau” of eigenvalues along the x-axis is expected following Theorem 1. Comparing the distribution before and after temperature scaling indicates that the predicted eigenvalues are adjusted towards the diagonal, which reduces the respective ECE.

5 RELIABILITY DIAGRAMS FOR EIGENVALUES

In classification, it is a common procedure to optimize calibration via risk minimization and to assess a model’s calibration via reliability diagrams. There, reliability diagrams are constructed by binning the probability predictions of the model and then computing the target frequency per bin. In the following, we transfer this procedure to eigenvalues.

Theorem 1 states that matrix calibration only translates to calibrated eigenvalues if we condition on the predicted density matrix. Otherwise, we end up with an inequality. The empirical implications of this theoretical result can be observed when plotting reliability diagrams for the maximum eigenvalue without information on the predicted density

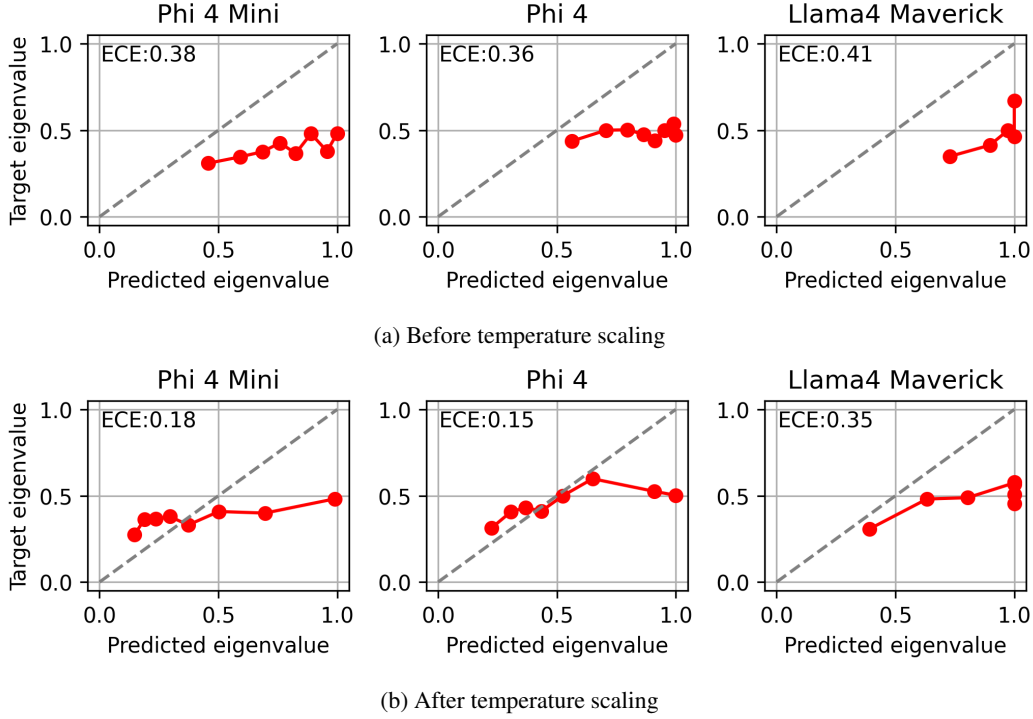


Figure 5: Reliability diagrams according to Algorithm 1 of LLMs before and after temperature scaling. Figure 5a shows that all models are systematically overconfident in their predicted eigenvalue. This overconfidence is reduced via temperature scaling as seen in Figure 5b. The respective ECE is also reduced in consequence. This improves the correctness of predicted eigenvalues by the models.

matrix. Here, we bin according to the predicted maximum eigenvalue and then compute the target eigenvalue in each bin. The resulting reliability diagram for Phi 4 Mini on TriviaQA is depicted in Figure 4. The weighted gap between the diagonal line and the empirical target eigenvalues is included as the expected calibration error (ECE) value. We identify that temperature scaling reduces some of the model overconfidence, which is also indicated via the ECE value.

Ideally, we want a reliability diagram for Equation 11, i.e., we need to condition on information about the predicted confidence matrix beyond the maximum eigenvalue. To achieve this, we propose Algorithm 1, which uses hierarchical clustering as an additional step during the binning procedure. The clustering is done based on the predicted density matrices, which, in consequence, includes their information for computing the target eigenvalues. We then compute the average target eigenvalue across clusters to “collapse” the information again on a single axis for plotting. Therefore, instead of plotting $\lambda_{\max}(\mathbb{D}_{Y|\lambda_{\max}(d(X))})$ on the y-axis, we plot $\mathbb{E}[\lambda_{\max}(\mathbb{D}_{Y|d(X)}) | \lambda_{\max}(d(X))]$. This is in line with Equation 11. The resulting reliability diagrams for $B = 8$ bins and $C = 5$ clusters are depicted in Figure 5 (Natural Questions) and Figures 1c & 1d (Phi 4 Mini on TriviaQA). Here, it is more pronounced how temperature scaling improves the calibration of the LLMs, both

visually and in the ECE value. We also provide the corresponding reliability diagrams for TriviaQA on all models in Appendix B, which further confirm this observation.

Further, we validate the choice of $B = 8$ and $C = 5$ for Algorithm 1 using a sensitivity analysis. We also compare the performance of matrix temperature scaling to a sampling temperature calibration baseline for LLMs from Lamb et al. [2025]. Finally, we reproduce the main results of this paper using a different embedding model, and achieve similar outcomes. All the details and results of these experiments are included in Appendix B.

6 CONCLUSION

We introduced a principled framework for calibrating eigenvalues of semantic embeddings, bridging a key gap between conventional calibration and state-of-the-art uncertainty quantification methods for LLMs. Our results establish entropy–risk equivalence under calibration, prove a central calibration inequality for eigenvalues, and show that temperature scaling reduces calibration error. Evaluations confirmed that modern LLMs are systematically overconfident and that recalibration allows for interpreting average entropy as risk, besides improving several other metrics. Therefore, eigenvalue calibration opens new directions for

Algorithm 1: Reliability diagram for eigenvalues of density matrix predictions.

Input: Target embeddings $Y \in \mathbb{R}^{n \times d}$ and predicted density matrices $D \in \mathbb{R}^{n \times d \times d}$ (data instances n , embedding dimension d), bin number B , cluster number C .

Output: predicted eigenvalues $\lambda_{\text{preds}} \in \mathbb{R}^B$, target eigenvalues $\lambda_{\text{targets}} \in \mathbb{R}^B$

Compute max eigenvalues

$E \leftarrow \{\lambda_{\max}(D_i) \mid i \in 1..n\}$;

Compute equal mass bins $\mathbf{B}_1, \dots, \mathbf{B}_B \subset E$;

for $b \in 1..B$ **do**

 Select pred. density matrices d_b w.r.t. $\mathbf{B}_b$;

 Compute corr. matrix $M \leftarrow \text{corr}(d_b)$;

 Compute hierarchical clusters

$\mathbf{C}_1, \dots, \mathbf{C}_C \leftarrow \text{clusters}(M)$;

for $c \in 1..C$ **do**

 Select target embeddings Y_c w.r.t. $\mathbf{C}_c$;

 Compute target eigenvalue

$\mathbf{T}_{b,c} \leftarrow \lambda_{\max} \left(\frac{1}{m_{\text{sel}}} Y_c Y_c^\top \right)$;

end

$\lambda_{\text{preds},b} \leftarrow \text{avg}(\mathbf{B}_b)$;

$\lambda_{\text{targets},b} \leftarrow \text{avg}(\mathbf{T}_b)$;

end

return $\lambda_{\text{preds}}, \lambda_{\text{targets}}$;

embedding-based uncertainty quantification towards more reliable AI systems.

Acknowledgements

Co-funded by the European Union (ERC, TAIPO, 101088594 to FB). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal intelligence. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025. Accessed: 2025-10-02.

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. jina-embeddings-v5-text: Task-targeted embedding distillation. *arXiv preprint arXiv:2602.15547*, 2026.

Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, Jul 2009.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Sebastian G Gruber. *A novel framework for uncertainty quantification via proper scores for classification and beyond*. PhD thesis, Universitätsbibliothek Johann Christian Senckenberg, 2024.

Sebastian G. Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. In *Advances in Neural Information Processing Systems*, 2022.

Sebastian G. Gruber and Florian Buettner. A bias-variance-covariance decomposition of kernel scores for generative models. In *International Conference on Machine Learning*, pages 16460–16501, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics*, pages 1601–1611, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, 10(2), 2009.

- Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12295–12305, 2019.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.
- Tom A. Lamb, Desi R. Ivanova, Philip Torr, and Tim G. J. Rudner. Semantic-level confidence calibration of language models via temperature scaling. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.
- Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024.
- Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595 – 600, 1973.
- Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Martinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, 2024.
- Evgeni Y. Ovcharov. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1):53 – 79, 2018.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.
- Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9): 3211–3260, 2023.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, pages 16857–16867, 2020.
- Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European Conference on Computer Vision*, pages 555–569, 2022.
- Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, pages 3459–3467, 2019.
- Nassim Walha, Sebastian G Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. Fine-grained uncertainty decomposition in large language models: A spectral approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26090–26098, 2026.
- Mark M Wilde. *Quantum Information Theory*. Cambridge University Press, 2013.

Appendix

Sebastian G. Gruber^{†1}

Nassim Walha^{†2,3,4}

Francis Bach⁶

Florian Buettner^{2,3,4,5}

¹ESAT-PSI, KU Leuven, Belgium

²German Cancer Research Center (DKFZ), Heidelberg, Germany

³German Cancer Consortium (DKTK), Germany

⁴Goethe University Frankfurt, Germany

⁵Frankfurt Cancer Institute, Germany

⁶PSL Research University / Inria, France.

In this Appendix, we offer experimental details in Appendix A, additional results in Appendix B, and missing proofs from the main paper in Appendix C.

A EXTENDED EXPERIMENTAL DETAILS

In this section, we discuss more details on the experimental setup. The source code for all experiments is openly available at https://github.com/ML0-lab/matrix_eigenvalue_calibration.

For each development subset (of size $n = 300$) and each model, we use $m = 100$ sampled answers. For each test subset (of size $n = 1700$) and each model, we use $m = 20$ sampled answers.

For computing the log score, we add a small $\epsilon=1e-10$ to all eigenvalues equal to zero, to avoid computing the logarithm of zero.

To compute AUROC values in Table 1, we require a binary indicator of model correctness for each question. This is obtained by generating a single “standard” answer at temperature 0.1 and comparing it against the ground-truth reference using a fuzzy matching criterion based on the ROUGE-L score Kuhn et al. [2023].

The error bars for the risks and entropy values in all figures are computed via their standard deviation. The error bars for the AUROC values are computed via bootstrap sampling of $B = 20$ subsets.

B ADDITIONAL EXPERIMENTS AND RESULTS

In this section, we present additional experimental results.

B.1 EMPIRICAL ESTIMATOR CONVERGENCE

We assess the empirical estimator convergence by sampling a subset of answers repeatedly from a set of 300 sampled answers from an LLM. In Figure 6 are shown the results, where we include the standard deviation of the estimator. As can be seen, the estimator has a positive bias but converges quickly to the ground truth. Based on these results, we picked a sampling size of $m = 20$ in our main experiments.

B.2 RELIABILITY RESULTS FOR TRIVIAQA

We compare the change of eigenvalue calibration based on temperature scaling via risk minimization for TriviaQA in the same manner as Natural Questions in Section 4. The results are depicted in Figure 7. As can be seen, the results are similar to the ones discussed in the main paper and support our developed theory and methodology. Specifically, temperature scaling adjusts the calibration of the LLMs, which are systematically overconfident.

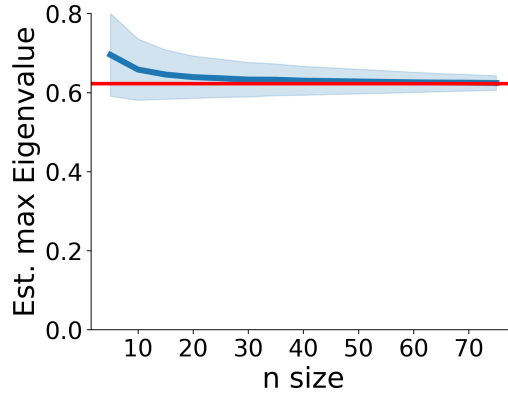


Figure 6: Estimating the maximum eigenvalue has a positive bias but converges quickly to the ground truth (red line) for $m \geq 20$ data instances.

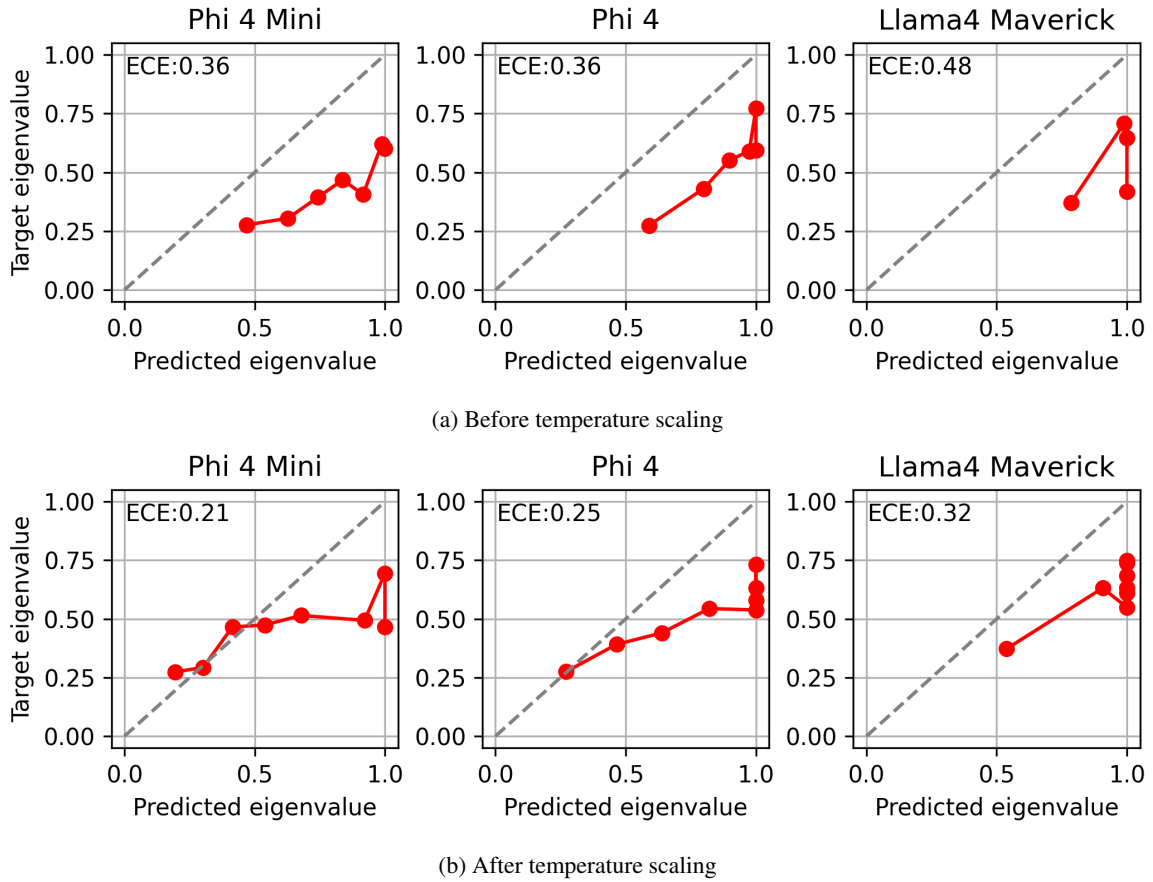


Figure 7: Reliability diagrams for TriviaQA according to Algorithm 1 of LLMs before and after temperature scaling. Figure 7a shows that all models are systematically overconfident in their predicted eigenvalue. This overconfidence is reduced via temperature scaling as seen in Figure 7b. The respective ECE is also reduced in consequence. This improves the correctness of predicted eigenvalues by the models.

B.3 CALIBRATION TOWARDS CORRECTNESS

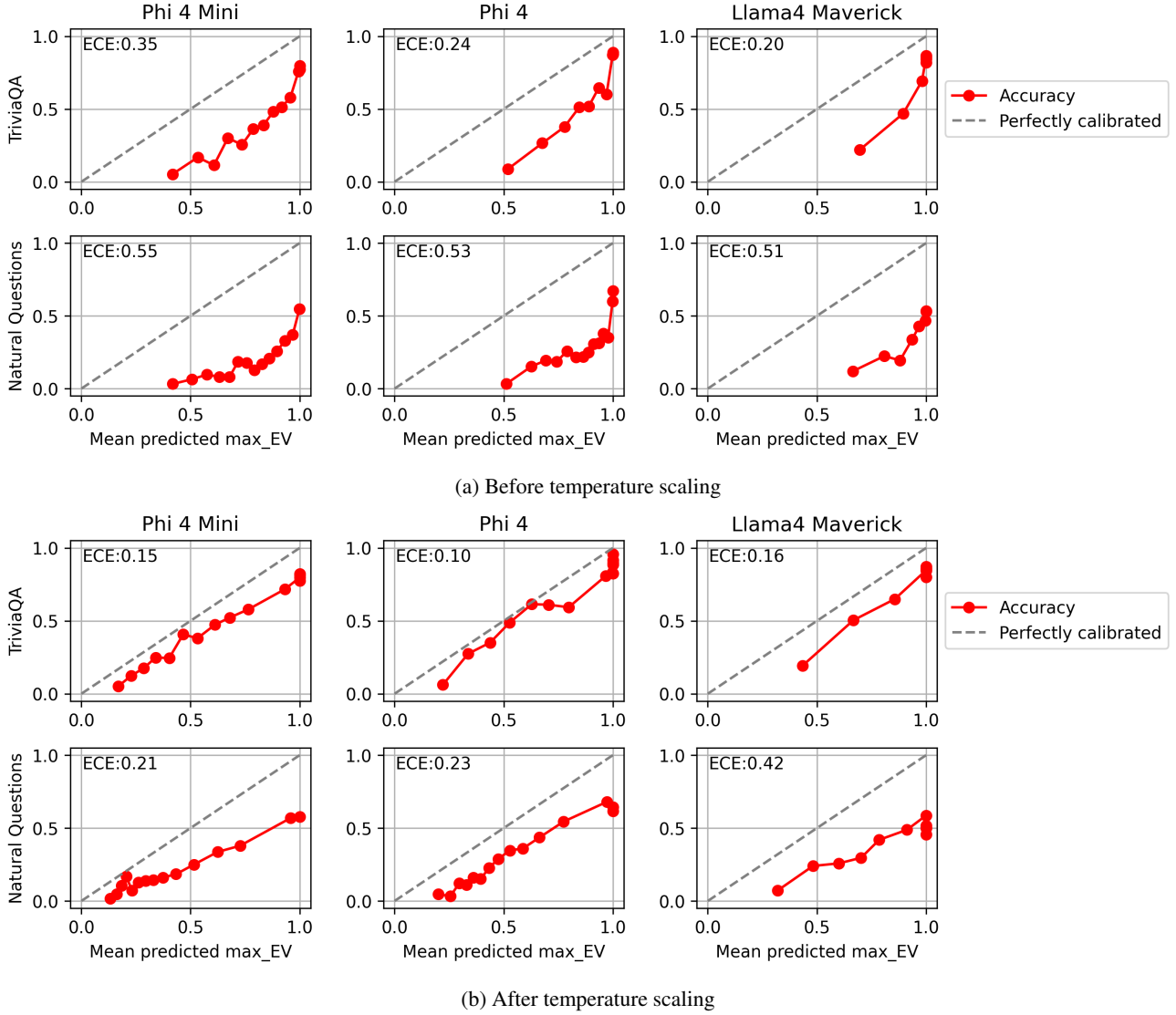


Figure 8: Reliability diagrams for conventional answer correctness. Matrix temperature scaling also has beneficial effects on how representative predicted eigenvalues are for the answer correctness, even though we do not directly optimize towards this objective.

As an additional downstream task analysis, we also evaluate how calibrating eigenvalues influences the predictiveness of the eigenvalues regarding the likelihood of answer correctness. For this, we construct conventional reliability diagrams with the predicted eigenvalue on the x-axis and the average correctness given only this prediction on the y-axis. The results are seen in Figure 8. Note that we still optimize temperature scaling via the risk of a matrix score, i.e., we do not change the optimization objective from the main paper. As we can see, calibrating via risk minimization also improves the calibration with the answer correctness likelihood, even though we do not directly optimize towards this objective.

Based on these results and the AUROC values in Table 1, we conclude that optimizing matrix calibration via risk minimization does not deteriorate any downstream tasks, and, in most cases, even improves on them.

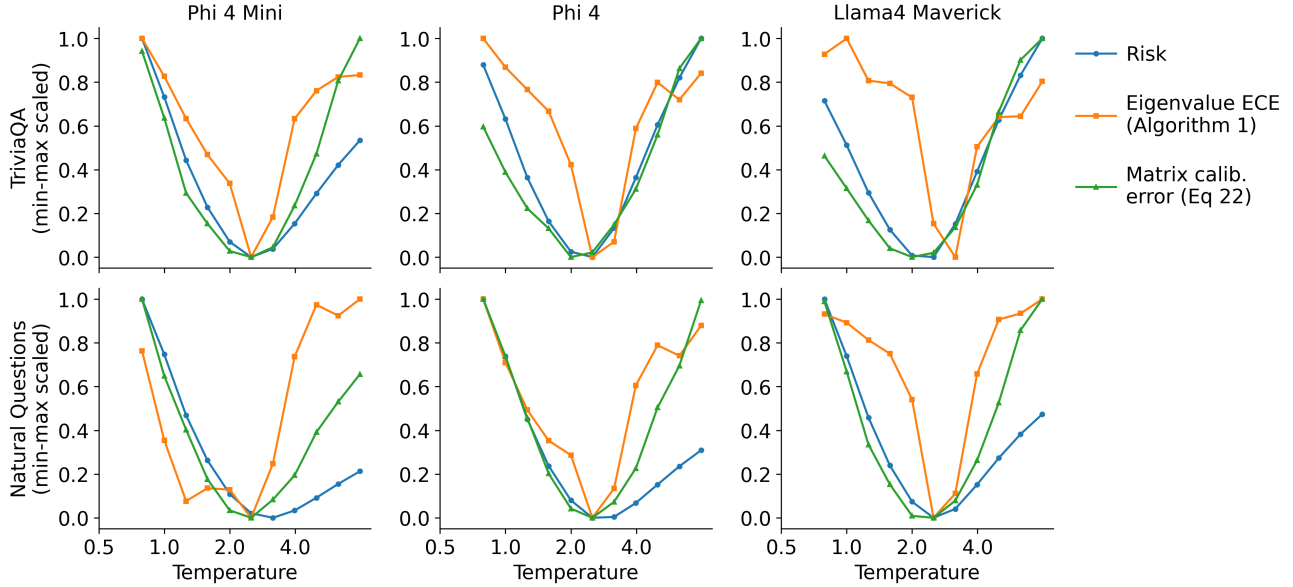


Figure 9: Scaled values of the risk $\mathcal{R}_S$, matrix calibration error Cal_S , and the eigenvalue ECE (as defined in Algorithm 1). As predicted by Theorem 3, all quantities are co-minimized at essentially the same temperature (with small deviations due to estimation noise).

B.4 RISK, CALIBRATION ERROR, AND EIGENVALUE ECE ARE CO-MINIMIZED

By Theorem 3, minimizing the risk $\mathcal{R}_S$ via matrix temperature scaling is equivalent to minimizing the matrix calibration error Cal_S . Theorem 1 further suggests that this equivalence should also manifest as a minimal eigenvalue ECE, as defined in Algorithm 1.

We apply matrix temperature scaling across all models and datasets in our evaluation, tracking each of these quantities throughout. The results, shown in Figure 9, confirm the theoretical prediction. To estimate the matrix calibration error Cal_S , we cluster the predicted density matrices $d(X)$ across the full dataset, which is required to obtain a reliable estimate of $\mathbb{D}_{Y|d(X)}$. Estimating the eigenvalue ECE via Algorithm 1, by contrast, requires a bin-then-cluster procedure, in which clustering of $d(X)$ is restricted to within each bin. Consequently, the eigenvalue ECE estimate is inherently noisier than the matrix calibration error estimate.

Finally, this experiment complements Figure 2 by showing that the matrix generalization of the cross-entropy loss recovers the optimal temperature not only for constant density matrix predictors, but also for realistic LLM-based density matrix predictors.

B.5 ADDITIONAL ANALYSIS FOR ALGORITHM 1

Table 3: Eigenvalue ECE (as computed by Algorithm 1) across different values of B (number of bins) and C (number of clusters) on the Natural Questions dataset using Phi 4 Mini. ECE is stable across a wide range of (B, C) once $C \geq 5$.

B	6			8			10		
C	3	5	7	3	5	7	3	5	7
Eigenvalue ECE	0.293	0.198	0.170	0.277	0.188	0.220	0.236	0.222	0.211

We propose to evaluate the sensitivity of Algorithm 1 to the choice of the binning and clustering parameters B and C . To this end, we run the algorithm for different values of these parameters and evaluate the eigenvalue ECE. The results are shown in Table 3. The low number of clusters $C = 3$ is the only failure mode in our analysis, where over-merging makes the

per-cluster target eigenvalues near-uniform (consistent with Figure 4). For $C \geq 5$, this problem disappears and we observe stable eigenvalue ECE values across different configurations. Our default values $B = 8$, $C = 5$ are chosen on that basis and do not require tuning per dataset or model.

Table 4: Within-cluster and within-bin pairwise similarity of predicted density matrices $d(X)$ across all models and datasets. Matrices grouped in one cluster are substantially more aligned with each other than matrices in one bin, highlighting the effectiveness of our clustering to approximate the conditioning on $d(X)$.

Dataset	Model	Within cluster pairwise similarity	Within bin pairwise similarity
Natural Questions	Llama4 Maverick	0.490	0.036
	Phi 4	0.351	0.033
	Phi 4 Mini	0.362	0.060
TriviaQA	Llama4 Maverick	0.401	0.046
	Phi 4	0.421	0.046
	Phi 4 Mini	0.318	0.064

We also perform a clustering-coherence check showing that density matrices grouped together by Algorithm 1 are substantially more aligned with each other than matrices that merely share a confidence bin, supporting the claim that the hierarchical clustering we use is a practical approximation to conditioning on $d(X)$. The results are presented in Table 4.

B.6 SAMPLING TEMPERATURE CALIBRATION BASELINE

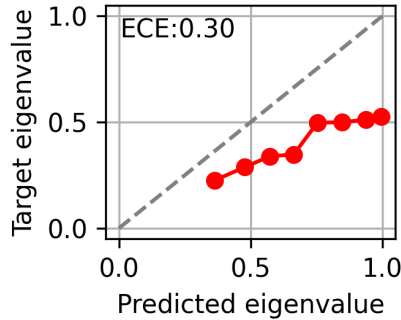
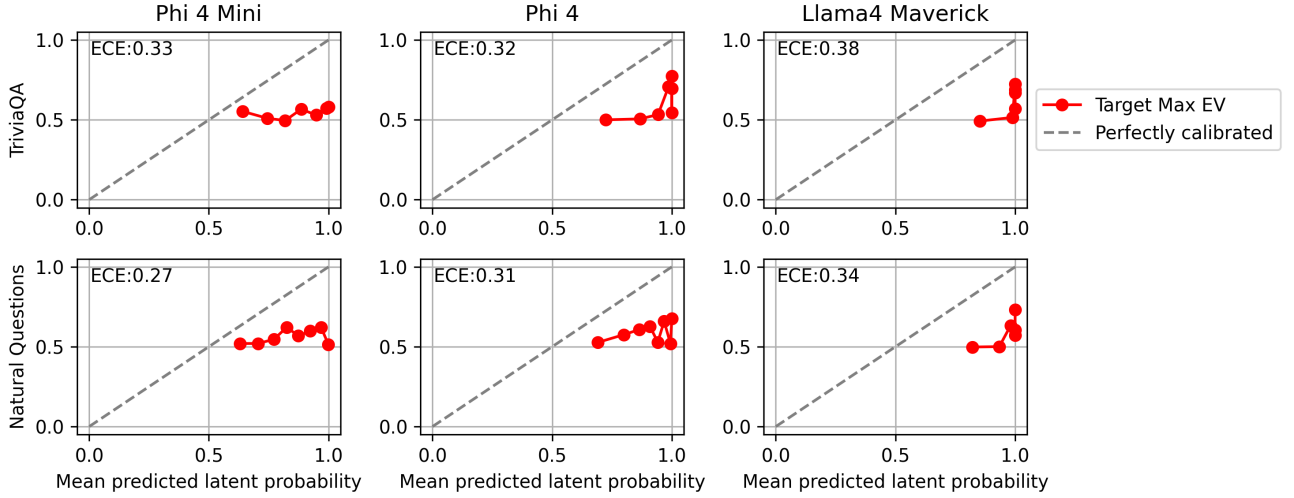
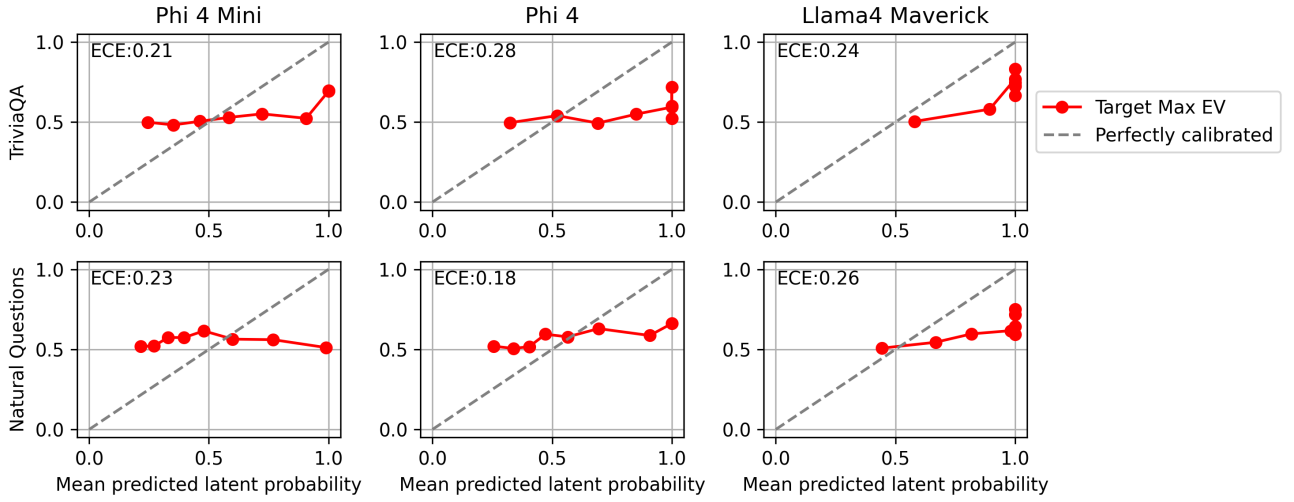


Figure 10: Eigenvalue-based reliability diagram (as computed by Algorithm 1) after applying the sampling temperature calibration baseline for the Figure 1c and Figure 1d setup (TriviaQA using Phi 4 Mini). Calibrating sampling temperature does reduce the eigenvalue ECE but the performance is limited compared to our matrix temperature scaling approach.

In this paper, we provide a novel notion of matrix calibration, and therefore there are no existing baselines that are designed for matrix calibration. Nevertheless, there are already existing works in the LLM calibration literature that consider other notions of calibration. For example, Lamb et al. [2025] compute the optimal sampling temperature that minimizes the calibration error against correctness, using the probability of the largest cluster from semantic entropy [Kuhn et al., 2023] as a confidence measure. We evaluate how this approach affects matrix calibration and compare it to matrix temperature scaling. First, similarly to Lamb et al. [2025], we identify the optimal sampling temperature t^* by minimizing the expected calibration error against correctness. Then, we use the temperature t^* to generate answers and provide predicted density matrices $d(X)$, for which we can evaluate matrix calibration using Algorithm 1. The resulting reliability diagram is depicted in Figure 10 and corresponds to the TriviaQA and Phi 4 Mini setup. Compared to the uncalibrated setup (Figure 1c) sampling temperature calibration [Lamb et al., 2025] reduced eigenvalue ECE from 0.36 to 0.30. However, as expected, matrix temperature scaling (Figure 1d) outperforms this baseline and reduces eigenvalue ECE even further to 0.21.



(a) Before temperature scaling



(b) After temperature scaling

Figure 11: Reliability diagrams according to Algorithm 1 of LLMs before and after temperature scaling using `jinaai/jina-embeddings-v5-text-nano` embeddings. Similarly to the main results, LLMs are always overconfident, and matrix temperature scaling effectively reduces this overconfidence, achieving lower eigenvalue ECEs.

Table 5: AUROC values ($\uparrow$) based on fuzzy answer correctness with maximum eigenvalue and entropy as uncertainty scores before and after temperature scaling, using the `jinaai/jina-embeddings-v5-text-nano` embedding model. Using the optimal temperature based on risk minimization improves the AUROC in many cases.

Dataset	Model	Eigenvalue	Eigenvalue TS	Entropy	Entropy TS
Natural Questions	Llama4 Maverick	0.658 $\pm$ 0.0	0.646 $\pm$ 0.0	0.646 $\pm$ 0.0	0.642 $\pm$ 0.0
	Phi 4	0.744 $\pm$ 0.001	0.795 $\pm$ 0.0	0.763 $\pm$ 0.001	0.797 $\pm$ 0.0
	Phi 4 Mini	0.765 $\pm$ 0.001	0.764 $\pm$ 0.001	0.768 $\pm$ 0.001	0.748 $\pm$ 0.001
TriviaQA	Llama4 Maverick	0.686 $\pm$ 0.001	0.699 $\pm$ 0.001	0.696 $\pm$ 0.001	0.697 $\pm$ 0.001
	Phi 4	0.816 $\pm$ 0.001	0.825 $\pm$ 0.001	0.825 $\pm$ 0.001	0.818 $\pm$ 0.001
	Phi 4 Mini	0.798 $\pm$ 0.001	0.799 $\pm$ 0.001	0.801 $\pm$ 0.0	0.79 $\pm$ 0.001

B.7 ADDITIONAL EMBEDDING MODEL

We reproduce the main experiments of the paper using another embedding model `jinaai/jina-embeddings-v5-text-nano` from Akram et al. [2026]. The reliability diagrams before and after matrix temperature scaling are depicted in Figure 11, and the AUROC values are provided in Table 5. Overall, we can reach the same conclusion: LLMs are always overconfident before calibration, and matrix temperature scaling helps reduce this overconfidence considerably, while also improving AUROCs of spectral uncertainty quantification methods in many cases. These results are embedder-agnostic and they are driven by the LLM’s answer distribution rather than by the particular embedding model we use.

C MISSING PROOFS

In this section, we present the missing proofs from the main paper.

C.1 PROOF OF THEOREM 1

Theorem 4 (Theorem 1 restated). *Let f be an LLM and d its respective density matrix predictor. If f is canonically calibrated then d is matrix calibrated. Further, if d is matrix calibrated, then*

$$\lambda_{\max}(\mathbb{D}_{Y|d(X)}) \stackrel{a.s.}{=} \Lambda_X, \quad (25)$$

with $\Lambda_X := \lambda_{\max}(d(X))$, but also

$$\lambda_{\max}(\mathbb{D}_{Y|\Lambda_X}) \stackrel{a.s.}{\leq} \Lambda_X. \quad (26)$$

As a further minor assumption, which we omitted for brevity, we require that all integrals in the following are finite.

Proof. We write $P := f(X)$ for brevity. We write $c = \int_{S_d} yy^\top dp(y)$ for any $p \in \mathcal{P}$ (which we assume exists in $\mathbb{R}^{d \times d}$). For every $p \in \text{supp } \mathbb{P}_P$ holds

$$\begin{aligned} \mathbb{P}(Y | P = p) &= p \\ &\stackrel{(i)}{\implies} \mathbb{E}_P[\mathbb{P}(Y | P) | d(X) = c] = \mathbb{E}[P | d(X) = c] \\ &\implies \mathbb{P}(Y | d(X) = c) = \mathbb{E}[P | d(X) = c] \\ &\stackrel{(i)}{\implies} \int_{S_d} yy^\top d\mathbb{P}_{Y|d(X)=c}(y) = \int_{S_d} yy^\top d\mathbb{E}[P | d(X) = c](y) \\ &\iff \mathbb{D}_{Y|d(X)=c} = \mathbb{E}\left[\int_{S_d} yy^\top dP(y) | d(X) = c\right] \\ &\iff \mathbb{D}_{Y|d(X)=c} = \mathbb{E}[d(X) | d(X) = c] \\ &\iff \mathbb{D}_{Y|d(X)=c} = c \\ &\implies \lambda_{\max}(\mathbb{D}_{Y|d(X)=c}) = \lambda_{\max}(c). \end{aligned} \quad (27)$$

Since $p \in \text{supp } \mathbb{P}_P$ is arbitrary, it follows from the last line that $\lambda_{\max}(\mathbb{D}_{Y|d(X)}) \stackrel{a.s.}{=} \Lambda_X$.

To show the inequality, note that $\lambda_{\max}$ is a convex function, which leads to

$$\mathbb{E}[\lambda_{\max}(\mathbb{D}_{Y|d(X)}) | \Lambda_X] \geq \lambda_{\max}(\mathbb{E}[\mathbb{D}_{Y|d(X)} | \Lambda_X]) = \lambda_{\max}(\mathbb{D}_{Y|\Lambda_X}). \quad (28)$$

Therefore, $\lambda_{\max}(\mathbb{D}_{Y|\Lambda_X}) \stackrel{a.s.}{\leq} \Lambda_X$. This is essentially a special case of [Gruber, 2024, Theorem 1]. $\square$

C.2 PROOF OF PROPOSITION 1

Proposition 4 (Proposition 1 restated). *A proper matrix score $\mathbf{S}: \mathbb{H}_d^\Delta \rightarrow \mathbb{R}^{d \times d}$ generates a proper score $S: \mathcal{P}_d \times \mathcal{Y}_d \rightarrow \mathbb{R}$ with respect to the set of distributions $\mathcal{P}_d := \left\{ P \mid \int_{S_d} \|x\|_2^2 dP(x) < \infty \right\}$ defined on the hypersphere and $\mathcal{Y}_d = \mathbb{R}^d$. The respective proper score is given by*

$$S(P, y) := y^\top \mathbf{S} (\mathbb{E}_{x \sim P} [xx^\top]) y. \quad (29)$$

Proof. First, note that S maps $\mathcal{P}_d \times \mathcal{Y}_d \rightarrow \mathbb{R}$. Now, we only need to show that for any $P, Q \in \mathcal{P}$ it holds $\mathbb{E}_{Y \sim Q} [S(P, Y)] \geq \mathbb{E}_{Y \sim Q} [S(Q, Y)]$. We will write $\mathbf{D}_Q := \mathbb{E}_{x \sim Q} [xx^\top]$ for any $Q \in \mathcal{P}_d$. It holds

$$\begin{aligned} \mathbb{E}_{Y \sim Q} [S(Q, Y)] &= \mathbb{E}_{Y \sim Q} [Y^\top \mathbf{S} (\mathbb{E}_{x \sim Q} [xx^\top]) Y] \\ &= \text{tr} (\mathbf{S} (\mathbf{D}_Q) \mathbb{E}_{Y \sim Q} [YY^\top]) \\ &= \text{tr} (\mathbf{S} (\mathbf{D}_Q) \mathbf{D}_Q) \\ &\leq \text{tr} (\mathbf{S} (\mathbf{D}_P) \mathbf{D}_Q) \\ &= \text{tr} (\mathbf{S} (\mathbb{E}_{x \sim P} [xx^\top]) \mathbb{E}_{Y \sim Q} [YY^\top]) \\ &= \mathbb{E}_{Y \sim Q} [Y^\top \mathbf{S} (\mathbb{E}_{x \sim P} [xx^\top]) Y] \\ &= \mathbb{E}_{Y \sim Q} [S(P, Y)], \end{aligned} \quad (30)$$

where the inequality follows from the definition of a proper matrix score. $\square$

C.3 PROOF OF PROPOSITION 2

Proposition 5 (Proposition 2 restated). *Given a proper matrix score $\mathbf{S}$, it holds that $H_{\mathbf{S}}$ is concave. Further, if $H_{\mathbf{S}}$ is also differentiable, then*

$$D_{\mathbf{S}} = \text{Div}_{-H_{\mathbf{S}}} . \quad (31)$$

Proof. First, we show concavity of $H_{\mathbf{S}}$:

$$\begin{aligned} H_{\mathbf{S}} (\lambda M_1 + (1 - \lambda) M_2) &= \text{tr} S (\lambda M_1 + (1 - \lambda) M_2) \lambda M_1 + \text{tr} S (\lambda M_1 + (1 - \lambda) M_2) (1 - \lambda) M_2 \\ &\geq \text{tr} S (M_1) \lambda M_1 + \text{tr} S (M_2) (1 - \lambda) M_2 \\ &= \lambda H_{\mathbf{S}} (M_1) + (1 - \lambda) H_{\mathbf{S}} (M_2), \end{aligned} \quad (32)$$

for any $\lambda \in (0, 1)$ and $M_1, M_2 \in \mathbb{H}_d^\Delta$.

Now comes the divergence part. For any $M_1, M_2 \in \mathbb{H}_d^\Delta$ it holds

$$\begin{aligned} D_{\mathbf{S}} (M_1, M_2) &= \text{tr} [\mathbf{S} (M_2) M_1 - \mathbf{S} (M_1) M_1] \\ &= \text{tr} [\mathbf{S} (M_2) M_1] - H_{\mathbf{S}} (M_1) \\ &= \text{tr} [\mathbf{S} (M_2) (M_1 - M_2)] - H_{\mathbf{S}} (M_1) + H_{\mathbf{S}} (M_2). \end{aligned} \quad (33)$$

Since $D_{\mathbf{S}} (M_1, M_2) \geq 0$ it follows that $H_{\mathbf{S}} (M_1) \geq \text{tr} [\mathbf{S} (M_2) (M_1 - M_2)] + H_{\mathbf{S}} (M_2)$. In addition to $-H_{\mathbf{S}}$ being convex, it follows that $\mathbf{S}$ is a subgradient of $-H_{\mathbf{S}}$, i.e., $\mathbf{S} (M_1) \in \partial(-H_{\mathbf{S}})(M_1)$. Since $-H_{\mathbf{S}}$ is differentiable, any subgradient is equal to the gradient, i.e., $\mathbf{S} = \nabla H_{\mathbf{S}}$. Therefore

$$D_{\mathbf{S}} (M_1, M_2) = \text{tr} [\nabla H_{\mathbf{S}} (M_2) (M_1 - M_2)] - H_{\mathbf{S}} (M_1) + H_{\mathbf{S}} (M_2) = \text{Div}_{-H_{\mathbf{S}}} (M_1, M_2). \quad (34)$$

$\square$

C.4 PROOF OF LEMMA 1

Lemma 3 (Lemma 1 restated). *Given a proper matrix score $\mathbf{S}$ and a density matrix predictor d , which is matrix calibrated for a joint distribution $\mathbb{P}_{X\mathbf{Y}}$, it holds that*

$$\underbrace{\mathbb{E}[H_{\mathbf{S}}(d(X))]}_{\text{Expected Uncertainty}} = \underbrace{\mathcal{R}_{\mathbf{S}}(d)}_{\text{Risk}}. \quad (35)$$

As a further, unstated minor assumption, we assume that all integrals used in the proof exist.

Proof. We omit *almost surely* in the following equations. Note that from d matrix calibrated follows that $\mathbb{D}_{Y|d(X)} = d(X)$. Further,

$$\begin{aligned} \mathbb{E}_{X\mathbf{Y}}[S(d(X), Y)] &= \mathbb{E}_X[\mathbb{E}_{\mathbf{Y}}[S(d(X), Y) | d(X)]] \\ &= \mathbb{E}_X[\mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_{Y|d(X)}}[S(d(X), Y)]] \\ &= \mathbb{E}_X\left[\int_{\mathcal{Y}_d} S(d(X), y) d\mathbb{P}_{Y|d(X)}(y)\right] \\ &= \mathbb{E}_X\left[\int_{\mathcal{Y}_d} y \mathbf{S}(d(X)) y^{\top} d\mathbb{P}_{Y|d(X)}(y)\right] \\ &= \mathbb{E}_X\left[\int_{\mathcal{Y}_d} \text{tr}(S(d(X)) yy^{\top}) d\mathbb{P}_{Y|d(X)}(y)\right] \\ &= \mathbb{E}_X\left[\text{tr}\left(S(d(X)) \int_{\mathcal{Y}_d} yy^{\top} d\mathbb{P}_{Y|d(X)}(y)\right)\right] \\ &= \mathbb{E}_X[\text{tr}(S(d(X)) \mathbb{D}_{Y|d(X)})] \\ &= \mathbb{E}_X[\text{tr}(S(d(X)) d(X))] \\ &= \mathbb{E}_X[H_{\mathbf{S}}(d(X))], \end{aligned} \quad (36)$$

where the matrix calibration assumption is used in the second-to-last equation. $\square$

C.5 PROOF OF THEOREM 2

Theorem 5 (Theorem 2 restated). *Let $d_1, d_2, \dots$ be a sequence of density matrix predictors such that $\lim_{n \rightarrow \infty} d_n$ is matrix calibrated. Then, for a continuous proper matrix score $\mathbf{S}$, it holds*

$$\underbrace{\mathbb{E}[H_{\mathbf{S}}(d_n(X))]}_{\text{Expected Uncertainty}} - \underbrace{\mathcal{R}_{\mathbf{S}}(d_n)}_{\text{Risk}} \longrightarrow 0 \quad (37)$$

for $n \rightarrow \infty$.

Proof. Since $\mathbf{S}$ is continuous, so is $H_{\mathbf{S}}$. We use S to denote the respective proper score based on $\mathbf{S}$. It holds

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}_X[H_{\mathbf{S}}(d_n(X))] &= \mathbb{E}_X\left[H_{\mathbf{S}}\left(\lim_{n \rightarrow \infty} d_n(X)\right)\right] \\ &= \mathbb{E}_{X\mathbf{Y}}\left[S\left(\lim_{n \rightarrow \infty} d_n(X), \mathbf{Y}\right)\right] \\ &= \lim_{n \rightarrow \infty} \mathbb{E}_{X\mathbf{Y}}[S(d_n(X), \mathbf{Y})], \end{aligned} \quad (38)$$

where the second-to-last equality follows from the assumption and Lemma 1. $\square$

C.6 PROOF OF LEMMA 2

Lemma 4 (Lemma 2 restated). *Let $\mathbf{S}$ be a proper matrix score and d be a density matrix predictor for a joint distribution $\mathbb{P}_{X\mathbf{Y}}$. Assuming all integrals are finite, it holds*

$$\underbrace{\mathcal{R}_S(d)}_{\text{Risk}} = \underbrace{\text{Cal}_S(d)}_{\text{Calibration}} - \underbrace{\mathbb{E} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, \mathbb{D}_Y)]}_{\text{Sharpness}} + \underbrace{H_{\mathbf{S}} (\mathbb{D}_Y)}_{\text{Noise}}, \quad (39)$$

with the respective calibration error being defined by

$$\text{Cal}_S(d) := \mathbb{E} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, d(X))]. \quad (40)$$

Proof. In the following, we use the definitions of the risk, divergence, and entropy associated with $\mathbf{S}$.

$$\begin{aligned} & \mathbb{E}_{X\mathbf{Y}} [\mathbf{Y}^\top \mathbf{S}(d(X)) \mathbf{Y}] \\ &= \mathbb{E}_{d(X)} [\mathbb{E}_{\mathbf{Y}} [\mathbf{Y}^\top \mathbf{S}(d(X)) \mathbf{Y} \mid d(X)]] \\ &= \mathbb{E}_{d(X)} [\text{tr} [\mathbf{S}(d(X)) \mathbb{D}_{Y|d(X)}]] \\ &= \mathbb{E}_{d(X)} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, \mathbb{D}_Y)] - \mathbb{E} [H_{\mathbf{S}} (\mathbb{D}_{Y|d(X)})] \\ &= \mathbb{E}_{d(X)} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, d(X))] - \mathbb{E} [H_{\mathbf{S}} (\mathbb{D}_{Y|d(X)})] + H_{\mathbf{S}} (\mathbb{D}_Y) - H_{\mathbf{S}} (\mathbb{D}_Y) \\ &= \mathbb{E}_{d(X)} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, d(X))] - \mathbb{E}_{d(X)} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, \mathbb{D}_Y)] - H_{\mathbf{S}} (\mathbb{D}_Y). \end{aligned} \quad (41)$$

□

C.7 PROOF OF THEOREM 3

Theorem 6 (Theorem 3 restated). *Given a proper matrix score $S: \mathbb{H}_d^\Delta \times \mathcal{Y}^d \rightarrow \mathbb{H}_d^\Delta$, a density matrix predictor $d: \mathcal{X} \rightarrow \mathbb{H}_d^\Delta$ and an injective function $h: \mathbb{H}_d^\Delta \rightarrow \mathbb{H}_d^\Delta$, it holds*

$$\underbrace{\mathcal{R}_S(h \circ d) - \mathcal{R}_S(d)}_{\text{Change in Risk}} = \underbrace{\text{Cal}_S(h \circ d) - \text{Cal}_S(d)}_{\text{Change in Calibration}}. \quad (42)$$

Proof. According to Lemma 2, the only difference between the risk and the respective calibration error is the sharpness and entropy terms. The entropy term is independent of the prediction. Therefore, we only need to show that

$$\mathbb{E}_{d(X)} [D_{\mathbf{S}} (\mathbb{D}_{Y|d(X)}, \mathbb{D}_Y)] = \mathbb{E}_{d(X)} [D_{\mathbf{S}} (\mathbb{D}_{Y|h(d(X))}, \mathbb{D}_Y)]. \quad (43)$$

This follows since from the injectivity of h follows $\mathbb{P}(Y \mid h(d(X))) = \mathbb{P}(Y \mid d(X))$ from which follows that $\mathbb{D}_{Y|h(d(X))} = \mathbb{D}_{Y|d(X)}$. This is analogous to [Gruber and Buettner, 2022]. □

C.8 PROOF OF PROPOSITION 3

Proposition 6 (Proposition 3 restated). *The function $h_{\text{TS}}: \mathbb{H}_d^\Delta \rightarrow \mathbb{H}_d^\Delta$ defined via*

$$h_{\text{TS}}(M) := \frac{1}{\text{tr } M^\alpha} M^\alpha = \frac{1}{\sum_{i=1}^d \lambda_i^\alpha} \sum_{i=1}^d \lambda_i^\alpha e_i e_i^\top \quad (44)$$

is injective for $\alpha > 0$.

Proof. Note that matrix temperature scaling is a “stretching” and “squeezing” of the ellipse represented by M along its eigenvectors.

Let $M_1, M_2 \in \mathbb{H}_d^\Delta$ such that $M_1 \neq M_2$. Let $M_1 = U_1 \Lambda_1 U_1^\top$ and $M_2 = U_2 \Lambda_2 U_2^\top$ be their eigenvalue decompositions. If $\Lambda_1 = \Lambda_2$, then $U_1 \neq U_2$, from which follows

$$h_{\text{TS}}(M_1) = U_1 \frac{1}{\text{tr } \Lambda_1^\alpha} \Lambda_1^\alpha U_1^\top \neq U_2 \frac{1}{\text{tr } \Lambda_1^\alpha} \Lambda_1^\alpha U_2^\top = h_{\text{TS}}(M_2), \quad (45)$$

i.e., we cannot stretch and squeeze an ellipse into another when they “point” in different directions.

Since from $\frac{1}{\text{tr } \Lambda_1^\alpha} \Lambda_1^\alpha = \frac{1}{\text{tr } \Lambda_2^\alpha} \Lambda_2^\alpha$ follows that $c\Lambda_1 = \Lambda_2$ with $c = \left(\frac{\text{tr } \Lambda_2^\alpha}{\text{tr } \Lambda_1^\alpha} \right)^{\frac{1}{\alpha}}$, it is required that $c = 1$ otherwise Λ_1 or Λ_2 would not have normalized eigenvalues. Reversing the implication, it follows that if $\Lambda_1 \neq \Lambda_2$, then $\frac{1}{\text{tr } \Lambda_1^\alpha} \Lambda_1^\alpha \neq \frac{1}{\text{tr } \Lambda_2^\alpha} \Lambda_2^\alpha$. From this follows

$$h_{\text{TS}}(M_1) = U_1 \frac{1}{\text{tr } \Lambda_1^\alpha} \Lambda_1^\alpha U_1^\top \neq U_1 \frac{1}{\text{tr } \Lambda_2^\alpha} \Lambda_2^\alpha U_1^\top = h_{\text{TS}}(M_2), \quad (46)$$

when $U_1 = U_2$ and $\Lambda_1 \neq \Lambda_2$. Therefore, h_{TS} is injective. $\square$