The Energy to Say No: Pre-Generation Abstention for Safety-Critical Medical RAG

Ravi Shankar[1] **Sheng Wong**[1] **Lin Li**[2] **Magdalena Bachmann**[3] **Alex Silverthorne**[3]
Beth Albert[1] **Gabriel Davis Jones**[1]*
[1] OxDHL, Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK
[2] OATML, Department of Computer Science, University of Oxford, Oxford, UK
[3] Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK

Abstract

Retrieval-augmented generation (RAG) systems require reliable abstention mechanisms to avoid generating harmful responses, particularly in safety-critical domains such as women's health where incorrect answers can lead to serious consequences. However, existing confidence estimation approaches often fail to provide adequate safety guarantees for pre-generation decision making. We introduce the *Margin-Structured Energy-Based Model (MS-EBM)*, a framework that learns smooth energy landscapes over dense semantic representations of guideline-derived questions, enabling systems to make principled abstention decisions before generation occurs. Using identical in-batch negatives for training and validation, we evaluate MS-EBM against softmax-based confidence estimation and non-parametric baselines including k-NN, ODIN, and Mahalanobis distance across three out-of-distribution scenarios: *Hard*, *Easy*, and *Mixed* splits. Results demonstrate substantial improvements in abstention quality, with MS-EBM achieving AUROC scores of 0.946, 0.977, and 0.961 on Hard, Easy, and Mixed splits respectively, compared to 0.895, 0.937, and 0.916 for softmax baselines. The model also significantly reduces false positive rates, achieving FPR@95TPR of 41.3% versus 69.4% on Hard splits. Comprehensive ablation studies reveal that heterogeneous negative sampling, combining both hard and easy negatives, proves essential for robust out-of-distribution generalisation, while curriculum design shows minimal impact once diverse negatives are included. Analysis through risk-coverage curves and energy-gap distributions confirms that MS-EBM's scoring provides more reliable confidence signals than probability-based approaches, offering a scalable and interpretable foundation for building safer RAG systems.

1 Introduction

Large language models (LLMs) coupled with retrieval-augmented generation (RAG) Lewis et al. [2020] are being piloted for clinical decision support Thirunavukarasu et al. [2023], Nori et al. [2023], Moor et al. [2023], He et al. [2025b]. They can synthesise guidance across large corpora, yet they also generate fluent errors when inputs fall outside scope or when retrieved evidence is sparse or misleading Ji et al. [2023], Huang et al. [2025]. In safety-critical care, such failures erode trust and can cause harm Miotto et al. [2022]. Robust abstention is therefore a first-order requirement Chow [1970], Geifman and El-Yaniv [2017], Liu et al. [2020], Kamath et al. [2020], Wen et al. [2025]: the system should recognise when not to answer and instead expand retrieval, escalate, or defer to a human expert.

*Corresponding author: gabriel.jones@wrh.ox.ac.uk

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

RAG is particularly promising where the knowledge corpus can be strictly controlled. In healthcare, evidence-based guidelines, drug formularies, and institution-specific protocols can be curated, versioned, and indexed so that generation is constrained to cited sources with known provenance and update cycles Singhal et al. [2023], Xiong et al. [2025]. This limits reliance on memorised web text, reduces parametric drift, enables auditability, and supports time-bounded answers Bommasani et al. [2021]. In women's health, for example, authoritative guidance from bodies such as Royal College of Obstetricians and Gynaecologists (RCOG), National Institute for Health and Care Excellence (NICE), and World Health Orgnaisation (WHO) can anchor responses to accepted standards of care. Nevertheless, reliability depends on recall and scope: if relevant material is not retrieved or the query falls outside the indexed corpus, the model may generalise beyond evidential support He et al. [2025a]. In these settings, abstention and explicit self-assessment of evidential sufficiency are critical Kadavath et al. [2022], Asai et al. [2024].

In healthcare applications, two distinct classes of query could trigger abstention:

- **(i) Domain-irrelevant prompts** that fall outside healthcare expertise, for example questions related to finance or economics, where the correct behaviour is immediate redirection; and
- **(ii) Domain-relevant but out-of-scope queries** relative to the model's training or validation, such as applying pregnancy-specific diabetes or hypertension protocols to non-pregnant adults, paediatric gynaecology when training covered only adult care, oncological queries requiring therapies, dosing, or trial evidence not present in the corpus, site-specific policies absent from the training data, or modality shifts such as image-first triage when the model was trained only on text. These near-distribution queries are hazardous because they are semantically close to in-scope content and can elicit persuasive but unsafe answers Hendrycks et al. [2022], Li et al. [2025].

Abstention should be the default unless retrieved evidence is sufficiently specific and in-distribution (ID). To address this we experimented a margin-structured abstention framework built on energy-based modelling. Our method combines a novel loss function with a semi-structured negative sampling strategy to explicitly separate in-domain (ID) queries from both trivial irrelevancies and clinically plausible confusables. The resulting energy scores provide calibrated abstention decisions that are robust in the challenging regime of near-distribution OOD queries (as discussed in category (ii) Domain-relevant but out-of-scope queries above).

Our contributions are threefold:

1. **Margin-structured abstention for medical RAG:** We introduce the Margin-Structured Energy-Based Model (MS-EBM), a dual-head framework trained with the Energy-Calibrated Semi-Contrastive Triplet Loss (EC-SCTL). This novel loss combines similarity and energy margins with auxiliary hinges, jointly optimised under a semi-structured negative sampling strategy. The resulting pre-generation abstention layer is model-agnostic and consistently outperforms probability- and density-based baselines on hard, safety-critical queries.
2. **Fair hard-case benchmarking**: We evaluate on easy, hard, and mixed OOD splits under corrected methodology, showing that while non-parametric baselines like kNN excel on easy OOD, our MS-EBM is substantially more robust on safety-critical hard queries where simple density methods collapse. This benchmark design highlights the conditions where density or probability-based confidence suffices and where structured energy separation is required.
3. **Ablation-driven insights**: We showcase the robustness arising from the synergy between heterogeneous negative exposure (easy, mid-range, and hard negatives) and margin-based energy shaping. This provides methodological guidance for abstention-aware RAG in safety-critical domains such as women's health.

2 Related work

Prior work in women's health underscores the need for principled abstention Draelos et al. [2025]. In a head-to-head evaluation with questions from the UK RCOG, ChatGPT achieved moderate accuracy on basic science but only coin-toss performance on clinical reasoning, while often expressing high confidence irrespective of correctness, indicating unreliable self-assessment Bachmann et al.

[2024]. Meaning-level uncertainty signals (for example semantic entropy) offer a more discriminative path: *semantic entropy* outperformed perplexity for identifying unreliable outputs on obstetrics and gynaecology questions, and achieved expert-validated discrimination approaching ceiling, supporting the value of pre-generation uncertainty checks and deferral mechanisms Penny-Dimri et al. [2025]. However, computing semantic entropy typically requires first generating multiple candidate responses to estimate the meaning distribution, which adds latency and compute cost.

Foundational work on abstention frames the problem as a reject option in statistical decision theory, and selective prediction formalises the coverage–risk trade-off with standard risk–coverage evaluation Chow [1970], El-Yaniv and Wiener [2010], Geifman and El-Yaniv [2017, 2019]. Post-hoc calibration and conformal prediction can support abstention, although they do not shape the representation space during training Guo et al. [2017], Angelopoulos and Bates [2021], Kumar et al. [2019]. Out-of-distribution (OOD) detection methods include maximum softmax probability, ODIN, and Mahalanobis distance, with surveys detailing pitfalls and best practice Hendrycks and Gimpel [2017], Liang et al. [2018], Lee et al. [2018], Mohseni et al. [2020], Yang et al. [2024], Fort et al. [2021]. Energy-based models interpret predictions via an energy landscape in which in-distribution samples receive low energy, and explicit energy training has improved OOD detection; margin-based and contrastive variants further shape the energy function LeCun et al. [2006], Grathwohl et al. [2020], Liu et al. [2020], Wang et al. [2022].

In RAG, conditioning on retrieved passages improves factuality, yet systems can still over-commit on OOD inputs. Prior work explores selective QA, self-evaluation, and retrieval-aware abstention Lewis et al. [2020], Kamath et al. [2020], Kadavath et al. [2022], Asai et al. [2024], Margatina et al. [2023]. Self-evaluation methods typically rely on the LLM to generate candidate answers or partial generations before abstaining, which adds latency and couples abstention to the model's parametric behaviour Kadavath et al. [2022], Li et al. [2022]. Self-RAG, for instance, trains a critic to decide whether retrieval is sufficient only after generation has begun Asai et al. [2024].

Contrastive learning, where a model is trained to distinguish between similar and dissimilar pairs of data, benefits from informative negative mining, widely studied in metric learning and dense retrieval Schroff et al. [2015], Wu et al. [2018], Hermans et al. [2017], Karpukhin et al. [2020], Xiong et al. [2021], Robinson et al. [2021]. For medical QA, integrating external corpora helps calibrate a broad not-our-domain boundary; common OOD pools include MedMCQA and SQuAD Rajpurkar et al. [2018], Mutabazi et al. [2021], Pal et al. [2022], Rajpurkar et al. [2016]. Non-parametric scores such as the kth-neighbour similarity also provide strong abstention baselines when thresholds are fixed on validation and reused at test Sun et al. [2022], Berthelot et al. [2019], Hendrycks and Gimpel [2017], Liang et al. [2018], Liu et al. [2020], Ren et al. [2019].

Advancing existing literature, our approach inserts an abstention layer entirely *before* generation. The energy-based abstainer operates directly over dense embeddings of the query, using a smooth energy landscape with controlled negative exposure (hard, easy, and mixed ablations) to separate in-distribution from near-domain confusables. Low-energy queries proceed to generation, while high-energy queries trigger abstention or escalation, enabling a model-agnostic, pre-generation abstention layer that integrates with any RAG pipeline.

3 Methods

3.1 Data Preparation

We used four main sources of data: (1) *in-domain anchor questions*, derived from a corpus of best-practice clinical guidelines (see Appendix A1 Table 4 for a excerpt of the guidelines used for the corpus creation) in obstetrics and gynaecology curated by clinicians. Questions were generated from this corpus using ChatGPT-4o, with a subset subsequently validated by the clinical team. From this pool, we selected a representative set of 100K anchors using TF–IDF features and MiniBatchKMeans clustering. (2) *hard negatives*, synthetically generated with a controlled prompt that preserved the original question's structure and intent (e.g., diagnosis, management, screening) while substituting obstetrics/gynaecology terms with analogues from other specialties (e.g., uterus $\rightarrow$ prostate, CA-125 $\rightarrow$ PSA), thereby producing medically plausible but domain-shifted confounders; (3) *external OOD examples*, drawn from publicly available datasets, including the MedMCQA multi-subject medical QA dataset Pal et al. [2022] and the Stanford Question Answering Dataset (SQuAD) Rajpurkar et al. [2016]; and (4) a *reserve in-domain corpus* used for mid-range negative sampling. All texts

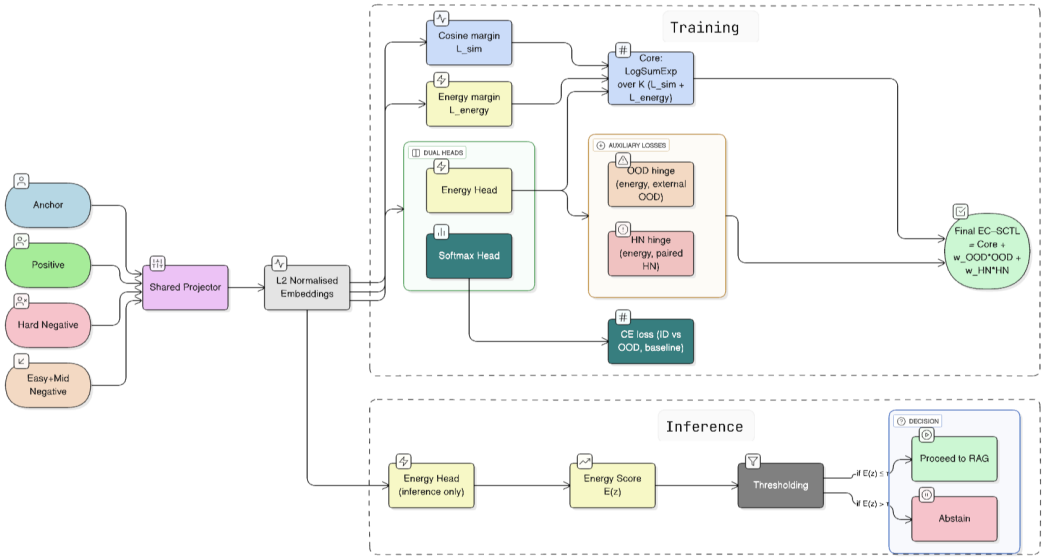

Figure 1: Training and inference pipeline of MS-EBM with semi-structured negative sampling and EC-SCTL loss

were embedded using the `BAAI/bge-m3` encoder Chen et al. [2024] and L2-normalised to ensure consistent cosine similarity across the corpus.

3.2 Positive and Negative Pair Generation

For each anchor question, the positive was selected using reciprocal-nearest-neighbour (RNN) filtering, retaining a pair only if each was the other's top-1 neighbour under cosine similarity. This strict reciprocity ensured semantically aligned positives while removing duplicates and one-sided matches Jarvis and Patrick [1973]. Mid-range negatives were drawn from similarity bands excluded if trivially easy or excessively hard to maintain a balance between informativeness and diversity. Hard negatives were defined as *confusable examples*, i.e., semantically close but clinically misleading text. These were generated using the prompt-based procedure described above with the `google/medgemma-4b-it` model Sellergren et al. [2025], pairing each anchor with one such hard negative to guarantee consistent coverage of confusing samples across splits.

In addition, we constructed an external OOD pool consisting of MedMCQA and SQuAD queries. Sampling from this pool during training encouraged the model to assign higher energy to irrelevant samples, directly improving abstention behaviour. Each training tuple therefore contained an anchor, its RNN-filtered positive, one MedGEMMA-derived hard negative, optional mid-range negatives, and OOD negatives. Since the number of negatives varied across anchors, tuples were padded and masked to enable efficient batch operations.

3.3 Model Architecture

Our model MS-EBM(see Figure 1), adopts a dual-branch design that combines a shared representation space with two task-specific scoring heads. A *projector network* first maps the 1024-dimensional input embeddings (pre-computed using `BAAI/bge-m3`) into a 256-dimensional latent space. The projector consists of two fully connected layers ($1024 \to 512 \to 256$), with a GELU activation on the hidden layer, followed by L2 normalisation of the output vector. This normalisation ensures that similarity comparisons are stable and that vector norms do not dominate the training objective.

On top of this shared latent space, we define two alternative heads:

- The *energy head*, a two-layer feedforward network ($256 \to 256 \to 1$) with GELU activation, produces a scalar *energy score* for each input. Within the energy-based modelling framework,

this score is optimised to separate in-domain anchors and positives from hard negatives, mid-range negatives, and out-of-distribution (OOD) samples.

- The *softmax head*, a two-layer feedforward classifier ($256 \to 256 \to 2$) with GELU activation, outputs logits for binary classification between in-domain and out-of-domain classes. This provides a probabilistic baseline against which the abstention-aware energy formulation can be compared.

During training, the projector weights are shared across all tuple elements (anchor, positive, and the various negative types). This ensures that all comparisons, whether cosine similarity, energy difference, or classification margin, are carried out in a consistent representation space. Both heads can thus be trained and evaluated fairly on identical embeddings, allowing for a direct comparison of softmax versus MS-EBM abstention.

3.4 Semi-Structured Negative Sampling

Each training mini-batch provides a synchronised tuple $(\mathbf{z}_A, \mathbf{z}_P, \mathbf{z}_{HN})$ for every anchor, where the hard negative $\mathbf{z}_{HN}$ is paired 1-to-1 with its anchor by construction. In addition, we form a heterogeneous candidate pool of negatives per anchor by concatenating: (i) mid-range similarity negatives and (ii) easy negatives, depending on the ablation configuration. To avoid leakage, these samples exclude items already assigned to training, validation, or test splits.

From this pool, we draw exactly k_{mine} negatives per anchor *uniformly at random* without replacement. These sampled negatives are used in the loss via a LogSumExp aggregation over per-negative terms. The projector parameters are shared across all tuple elements to ensure that similarity and energy are computed in a single, consistent latent space.

This procedure guarantees (a) a deterministically paired hard negative for each anchor, and (b) a fixed, fair exposure budget of k_{mine} additional negatives per anchor drawn from a heterogeneous pool.

3.5 Loss Functions

Energy-Calibrated Semi-Contrastive Triplet Loss (EC-SCTL). For each anchor-positive pair $(\mathbf{z}_A, \mathbf{z}_P)$, we use (i) its deterministically paired hard negative $\mathbf{z}_{HN}$, and (ii) K additional sampled negatives $\mathbf{z}_N$ drawn from the heterogeneous pool. The loss combines similarity and energy terms, drawing on contrastive learning Schroff et al. [2015], Hermans et al. [2017], hard-negative mining Wu et al. [2018], Karpukhin et al. [2020], and energy-based modelling Grathwohl et al. [2020], Liu et al. [2020].

Cosine similarity enforces relative closeness of anchor and positive:

$$\mathcal{L}_{\text{sim}} = \text{softplus}_T\Big(m_{\text{sim}} + \cos(\mathbf{z}_A, \mathbf{z}_N) - \cos(\mathbf{z}_A, \mathbf{z}_P)\Big),$$
$$\text{softplus}_T(x) = T \cdot \log\big(1 + \exp(x/T)\big), \quad T = \texttt{softplus\_temp}.$$

The energy term encourages positives to have lower energy than negatives:

$$\mathcal{L}_{\text{energy}} = \lambda \cdot \text{softplus}_T(E_P - E_N + m_E).$$

The K sampled negatives are aggregated with a LogSumExp:

$$\mathcal{L}_{\text{core}} = \tfrac{1}{T} \log \sum_{k=1}^{K} \exp\Big(T\big(\mathcal{L}_{\text{sim},k} + \mathcal{L}_{\text{energy},k}\big)\Big).$$

Here T interpolates between mean pooling ($T \to 0$) and max pooling ($T \to \infty$), allowing the model to adapt between averaging across all negatives and focusing on the hardest ones.

Two auxiliary hinge terms stabilise training:

$$\mathcal{L}_{\text{OOD}} = \text{softplus}_T\Big(m_{\text{OOD}} + E(\mathbf{z}_A) - \max_j E(\mathbf{z}_{\text{clean},j})\Big),$$
$$\mathcal{L}_{\text{HN}} = \text{softplus}_T\Big(m_{\text{HN}} + E(\mathbf{z}_A) - E(\mathbf{z}_{HN})\Big).$$

Here $\mathbf{z}_{\text{clean},j}$ are the *clean OOD negatives* drawn from the external pool, and the hinge uses the hardest one (highest energy).

The final objective is:

$$\mathcal{L}_{\text{EC-SCTL}} = \mathcal{L}_{\text{core}} + w_{\text{OOD}}\,\mathcal{L}_{\text{OOD}} + w_{\text{HN}}\,\mathcal{L}_{\text{HN}}.$$

Softmax baseline loss. The baseline classifier is trained with cross-entropy between in-domain ($y = 0$) and OOD ($y = 1$) classes:

$$\mathcal{L}_{\text{softmax}} = -\sum_{c\in\{0,1\}} y_c \log p_c,$$
$$p_c = \text{softmax}(\mathbf{z})_c.$$

3.6 Training and Evaluation Strategy

All models were trained under a consistent protocol to ensure fair comparison (see Table 3 in Appendix for more details on key hyperparameters and settings). For both the MS-EBM and softmax models, input embeddings were projected into a shared 256-dimensional latent space. Optimisation used AdamW (Learning rate: 5×10^{-4}, weight decay: 2×10^{-4}) with cosine annealing over 20 epochs and batch size 1024. Each anchor was paired with its positive and a synchronised hard negative, and further supplemented with k_{mine} additional sampled negatives. This ensured that both the energy and softmax heads were exposed to the same negatives per batch. Validation loss was monitored throughout training, and the checkpoint with the lowest value was retained.

The k-NN baseline was built on the same pre-computed BGE-M3 embeddings, indexed with FAISS [Douze et al., 2025] for cosine similarity, using $k = 5$ neighbours. For Energy–OOD (on the softmax head), we compute the OOD score as

$$-T \log \sum \exp\left(\frac{\text{logit}_{\text{ID}}}{T}\right)$$

from the binary ID/OOD logits, tuning T on the validation split to minimise DetErr and then fixing the threshold for test. For ODIN (temperature + embedding perturbation), we take a small step in the input *embedding* space to increase the ID probability and score

$$1 - \Pr(\text{ID} \mid T)$$

on the perturbed input; (T, ε) are selected on validation (with ℓ_∞ step and unit-norm re-normalisation) and the resulting threshold is fixed for test. For Mahalanobis (single-class), we fit a Gaussian to ID-train projector features $z = \text{projector}(x)$ (with shrinkage), and use the squared Mahalanobis distance

$$(z - \mu)^\top \Sigma^{-1} (z - \mu)$$

as the OOD score; the decision threshold is chosen on validation to minimise DetErr and applied unchanged at test.

Abstention thresholds were set on the validation split and then fixed for evaluation on the test set. For softmax and non-parametric baselines, thresholds were chosen to minimise detection error or to achieve 95% TPR. For the MS-EBM, we report results using raw energy scores without any post-hoc calibration, in order to isolate the contribution of the energy head itself. We evaluate two operating points: (i) the detection error threshold τ_{DetErr}, defined as

$$\mathrm{DetErr} = \tfrac{1}{2}\big(\mathrm{FPR}(\tau) + \mathrm{FNR}(\tau)\big),$$

which minimises the average of false positive and false negative rates; and (ii) the τ_{95} threshold, corresponding to the operating point closest to 95% TPR. For each method we report AUROC, AUPR, the false positive rate at τ_{95} ($\mathrm{FPR@95}$), and the detection error (DetErr) (see Appendix Table 2).

To assess design choices, we conducted ablations on the MS-EBM by varying negative exposure: `hard only`, `easy only`, `no hard`, `no easy`, and `all negatives`. These controlled ablations isolate the role of hard negatives in robustness to difficult queries.

Table 1: Comparison of all methods under the full training configuration (Easy+Mid+Hard negatives).

Method	Hard		Mixed	
	AUROC	FPR@95	AUROC	FPR@95
MS-EBM	**0.946**	**41.3**	**0.961**	**18.2**
Softmax	0.895	69.4	0.916	68.9
kNN (k=5)	0.855	0.6	0.926	0.6
Energy–OOD	0.895	69.5	0.916	68.7
ODIN	0.895	69.4	0.916	68.9
Mahalanobis	0.903	49.7	0.951	33.2

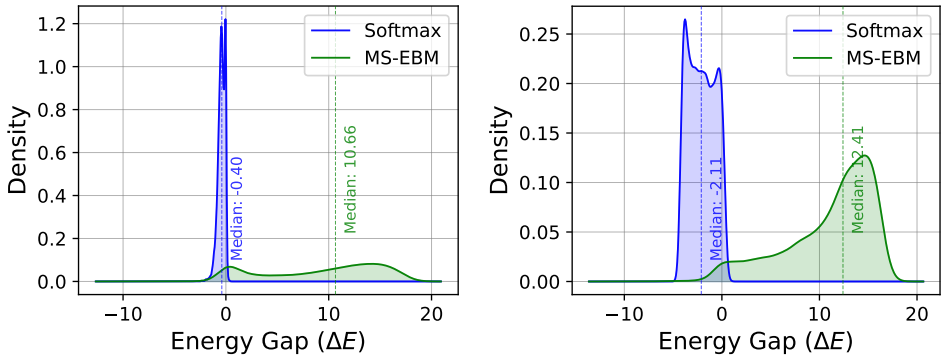

Figure 2: Energy–gap distributions (Hard and Mixed splits). MS-EBM shows clearer ID–OOD separation.

4 Results

Table 1 reports AUROC and FPR@95 on the *Hard* and *Mixed* splits, focusing on the safety-critical settings where near–domain confusions arise. Additional metrics (AUPR, DetErr) and results on the *Easy* split are provided in the Appendix (see Table: 2).Our results suggest three main finding as follows:

First, whenever training included hard negatives (`all negatives`, `hard only`, `hard+easy`), the MS-EBM consistently achieved higher AUROC and lower FPR@95 than softmax on the hard-OOD split. For example, in the full-data setting (`all negs`), AUROC improved from 0.895 (Softmax) to 0.946 (MS-EBM), while FPR@95 dropped from 69.4% to 41.3%. This shows that shaping the energy landscape yields a more reliable abstention signal for near-domain confusables. In Figure 2 and 4, the energy–gap distributions show that Softmax collapses near zero margin, while MS-EBM shifts the entire distribution rightwards, yielding consistent separation between positives and negatives across both Hard and Mixed splits.

Second, heterogeneous negatives were essential. When trained only on hard negatives (`hard_only`), both models failed on clean OOD: AUROC fell below 0.80 and FPR@95 exceeded 97%. Adding easy negatives (`all_negs` or `hard+easy`) restored strong performance (AUROC $>$ 0.98, FPR@95 $<$ 0.04), indicating that mixtures of hard and easy negatives are necessary to stabilise decision boundaries.

Third, non-parametric baselines such as k-NN excelled on clean OOD but collapsed on hard OOD. On easy OOD, k-NN achieved near-perfect AUROC ($>$ 0.998) and very low DetErr ($\sim$0.015), outperforming parametric models. On mixed OOD it reached AUROC 0.926, but on hard OOD AUROC dropped to 0.855 with DetErr above 0.42, showing that simple density fails against confusable examples where the MS-EBM retained clear margins.

Ablations (see Table 2, Figures 4, and 5 in Appendix) further confirmed that robustness requires both hard and easy negatives. Removing hard negatives (`no hard`, `easy only`) or easy negatives (`no easy`) sharply degraded mixed-OOD performance. For instance, AUROC dropped from 0.961 (`all negs`) to 0.708 (`no hard`) and 0.743 (`no easy`). This validates the design choice of pairing each

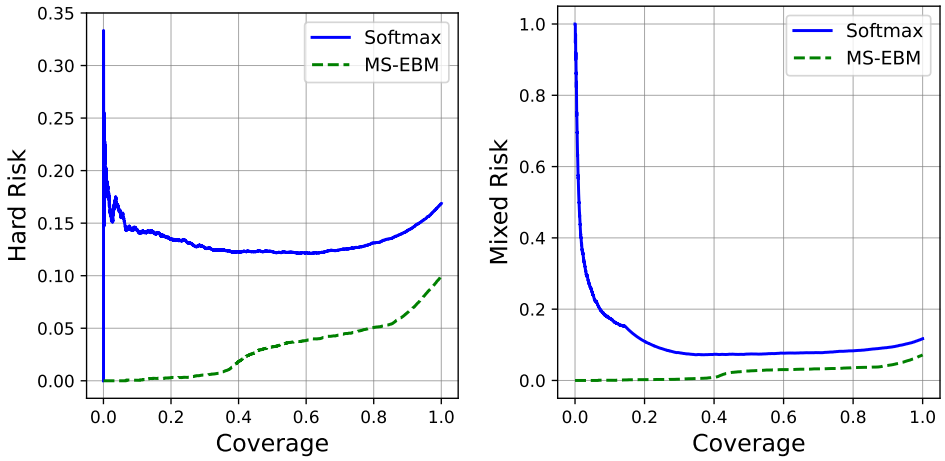

Figure 3: Risk–coverage curves comparing Softmax and MS-EBM under the full training configuration. Left: Hard split; Right: Mixed split. Lower selective risk at higher coverage indicates better abstention.

anchor with a synchronised hard negative while sampling additional easy and mid-range negatives. Risk-coverage curves (in Figure 3) show that the MS-EBM maintains lower selective risk across all coverage levels and achieves a sharper separation between ID and OOD queries.

5 Discussion

Our results suggest that the experimented MS-EBM offers a more reliable abstention mechanism than probability- or density-based approaches, especially on semantically hard, near-distribution queries. By shaping the latent space to separate in-domain anchors from confusable negatives, our method reduces over-commitment on unsafe cases.

Three implications follow. First, shaping the energy landscape provides a different abstention signal than softmax probabilities Grathwohl et al. [2020], Liu et al. [2020]: it enforces structured separation robust to near-distribution queries. Second, heterogeneous negatives are essential. Exposure to both easy and hard cases allows the model to reject trivial irrelevancies while also distinguishing clinically misleading confusions Wu et al. [2018], Karpukhin et al. [2020]. Third, non-parametric methods such as kNN Hendrycks and Gimpel [2017], Sun et al. [2022] perform well on clean OOD but fail on subtle semantic confusions, underscoring the value of parametric shaping.

These lessons are especially salient in healthcare, where overconfident errors carry direct clinical risks. Evaluation of ChatGPT against RCOG exams found only moderate accuracy and coin-toss clinical reasoning, with high confidence irrespective of correctness Bachmann et al. [2024]. Related work such as self-evaluation Kadavath et al. [2022] and Self-RAG Asai et al. [2024] still operate during or after generation, whereas our abstainer is model-agnostic and sits entirely before decoding. While semantic entropy, also operating post generation, offers a stronger uncertainty signal Penny-Dimri et al. [2025], it requires generating multiple outputs, adding latency. Our approach is complementary: MS-EBM provides a fast, pre-generation filter, while semantic entropy can further refine borderline situations that afford latency.

Compared to existing methods, MS-EBM requires no calibration, token-level likelihoods, or hand-crafted safeguards. It scales to large corpora and remains robust to near-domain confusions. Methodologically, this work contributes an energy-based scoring head trained over dense embeddings, a semi-structured negative sampling strategy, and corrected benchmarking that separates clean and hard OOD. For medicine, it offers a practical abstention mechanism that reduces unsafe overconfidence. For computer science, it establishes MS-EBM as a parametric alternative to perplexity, semantic entropy, and softmax baselines, with ablation-driven insights into negative exposure(see Table 2, Figures 2, and 3 in Appendix).

Nonetheless, limitations remain. The dataset is synthetically generated and restricted to English guideline-derived corpora, and synthetic hard negatives may not capture the full spectrum of real-world clinical diversity or query noise such as typos, abbreviations, grammatical variation, and code-switching. The abstention mechanism currently operates at the embedding level, without end-to-end integration, and thresholds are static. Because anchors and hard negatives are synthetic/guideline-derived, distributions may diverge from clinician queries; extending the negative pool to include naturally perturbed queries could further improve robustness. In practice, thresholds should be periodically re-established on a small validation set when guidelines or corpora change, with ongoing monitoring of risk–coverage under rolling windows and scheduled refresh of hard negatives or light fine-tuning to mitigate drift. Future work should extend to multilingual corpora, refine semi-hard negative mining, evaluate in clinical workflows, and explore hybrid systems combining MS-EBM with existing methods. Engagement with regulatory frameworks (e.g., EU AI Act, FDA) will also be essential to formalise abstention as a safety mechanism.

6 Conclusion

We presented MS-EBM, a pre-generation abstention framework for retrieval-augmented question answering that jointly learns similarity and energy margins within a shared latent space. Through semi-structured negative sampling and explicit margin design, MS-EBM reliably separates in-domain queries from confusable near-distribution and out-of-scope cases, providing a more dependable abstention signal than probability- or density-based methods. As it operates entirely before generation, the approach is model-agnostic and scalable, making it particularly suited for safety-critical domains such as healthcare. Looking ahead, extending MS-EBM to multilingual corpora, refining adaptive negative mining, and integrating clinician-in-the-loop evaluation will be important steps toward deployment. Together, these advances chart a path toward retrieval-augmented systems that are not only more accurate, but also demonstrably safer, fairer, and more trustworthy in real-world deployment.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. doi: 10.48550/arXiv.2107.07511. URL https://arxiv.org/abs/2107.07511.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2310.11511. URL https://arxiv.org/abs/2310.11511.

Magdalena Bachmann, Ioana Duta, Emily Mazey, William Cooke, Manu Vatish, and Gabriel Davis Jones. Exploring the capabilities of chatgpt in women's health: obstetrics and gynaecology. *npj Women's Health*, 2:26, 2024. doi: 10.1038/s44294-024-00028-w. URL https://doi.org/10.1038/s44294-024-00028-w.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. doi: 10.48550/arXiv.1905.02249. URL https://proceedings.neurips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258. URL https://arxiv.org/abs/2108.07258.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL https://arxiv.org/abs/2402.03216.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054427. URL https://ieeexplore.ieee.org/document/1054427.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2025. URL https://arxiv.org/abs/2401.08281.

Rachel L. Draelos, Samina Afreen, Barbara Blasko, Tiffany L. Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L. Gardner, Lauren Herrmann, Aswathy Vaikom House, Stephanie Kass, Marianne Kavan, Kirshma Khemani, Amanda Koire, Lauren M. McDonald, Zahraa Rabeeah, and Amy Shah. Large language models provide unsafe answers to patient-posed medical questions. *arXiv preprint arXiv:2507.18905*, 2025. doi: 10.48550/arXiv.2507.18905. URL https://arxiv.org/abs/2507.18905.

Ran El-Yaniv and Yair Wiener. *On the foundations of noise-free selective classification*, volume 11. 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 34, pages 7064–7077, 2021. doi: 10.48550/arXiv.2106.03004. URL https://proceedings.neurips.cc/paper/2021/hash/3941c4358616274ac2436eacf67fae05-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887, 2017. doi: 10.48550/arXiv.1705.08500. URL https://papers.nips.cc/paper/2017/hash/2862c844b8c3dbb60c94d7d5c4a87a85-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2151–2159, 2019. doi: 10.48550/arXiv.1901.09192. URL https://proceedings.mlr.press/v97/geifman19a.html.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. doi: 10.48550/arXiv.1912.03263. URL https://openreview.net/forum?id=HkxzxONtDB.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017. doi: 10.48550/arXiv.1706.04599. URL https://proceedings.mlr.press/v70/guo17a.html.

Jiawei He, Boya Zhang, Hossein Rouhizadeh, Yingjian Chen, Rui Yang, Jin Lu, Xudong Chen, Nan Liu, Irene Li, and Douglas Teodoro. Retrieval-augmented generation in biomedicine: A survey of technologies, datasets, and clinical applications. *arXiv preprint arXiv:2505.01146*, 2025a. doi: 10.48550/arXiv.2505.01146. URL https://arxiv.org/abs/2505.01146.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Inf. Fusion*, 118(C), April 2025b. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.102963. URL https://doi.org/10.1016/j.inffus.2025.102963.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1610.02136. URL https://arxiv.org/abs/1610.02136.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. 2022. URL https://arxiv.org/abs/2109.13916.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. 2017. URL https://arxiv.org/abs/1703.07737.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Ray A. Jarvis and Edward A. Patrick. Clustering using a similarity measure based on shared near neighbors, 1973. URL https://api.semanticscholar.org/CorpusID:9540064.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://dl.acm.org/doi/10.1145/3571730.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, et al. Language models (mostly) know what they know. 2022. URL https://arxiv.org/abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5684–5696, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL http://papers.neurips.cc/paper/8635-verified-uncertainty-calibration.pdf.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1:1–59, 2006. URL https://cs.nyu.edu/~yann/talks/lecun-tutorial-nips-2006.pdf.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7177, 2018. URL https://papers.neurips.cc/paper/2018/file/f4e8abd0cf5a1a336b71f4f6c6a25863-Paper.pdf. arXiv preprint arXiv:1807.03888.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Tianyu Li, Zhengbao Jiang, Vivek Srikumar, and He He. Selfcheck: Using llms to detect llm generations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. URL https://aclanthology.org/2022.emnlp-main.557.pdf.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for OOD detection. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=UXZJ3aL8vE.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://papers.neurips.cc/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf. arXiv preprint arXiv:2010.03759.

Katerina Margatina, Sebastian Ruder, Massimiliano Ciaramita, and Roi Reichart. Active retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 5431–5447, 2023. URL https://aclanthology.org/2023.findings-emnlp.436/.

Riccardo Miotto et al. Trustworthy ai in healthcare. *Nature Medicine*, 28:249–252, 2022. doi: 10.1038/s41591-022-01746-2. URL https://www.nature.com/articles/s41591-022-01746-2.

Sina Mohseni, Meghana Pitale, Jay Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. 34(4):5216–5223, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/5966.

Michael Moor, Oishi Banerjee, Zaid Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, Pranav Rajpurkar, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4. URL https://www.nature.com/articles/s41586-023-05881-4.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021. ISSN 2076-3417. doi: 10.3390/app11125456. URL https://www.mdpi.com/2076-3417/11/12/5456.

Harsha Nori, Nicholas King, Scott M. McKinney, Daniel Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023. URL https://arxiv.org/abs/2303.13375.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Jahan C. Penny-Dimri, Magdalena Bachmann, William R. Cooke, Sam Mathewlynn, Samuel Dockree, John Tolladay, Jannik Kossen, Lin Li, Yarin Gal, and Gabriel Davis Jones. Reducing large language model safety risks in women's health using semantic entropy. 2025. URL https://arxiv.org/abs/2503.00269.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, 2016. URL https://aclanthology.org/D16-1264/.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, 2018. URL https://aclanthology.org/P18-2124/.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark DePristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32, pages 14680–14691, 2019. doi: 10.48550/arXiv.1906.02845. URL https://papers.nips.cc/paper/9611-likelihood-ratios-for-out-of-distribution-detection.

Joshua David Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *Advances in Neural Information Processing Systems*, volume 34, pages 17777–17789, 2021. doi: 10.48550/arXiv.2106.11230. URL https://proceedings.neurips.cc/paper/2021/hash/27934a1f19d678a1377c257b9a780e80-Abstract.html.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. doi: 10.1109/CVPR.2015.7298682. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. 2022. URL https://arxiv.org/abs/2204.06507.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29: 1930–1940, 2023. URL https://api.semanticscholar.org/CorpusID:259947046.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. 2022. URL https://arxiv.org/abs/2203.10807.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models, 2025. URL https://arxiv.org/abs/2407.18418.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. 2018. URL https://arxiv.org/abs/1706.07567.

G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Pacific Symposium on Biocomputing*, volume 30, pages 199–214, 2025. URL https://doi.org/10.1142/9789819807024_0015.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. 2024. URL https://arxiv.org/abs/2110.11334.

A Appendices and Supplementary Material

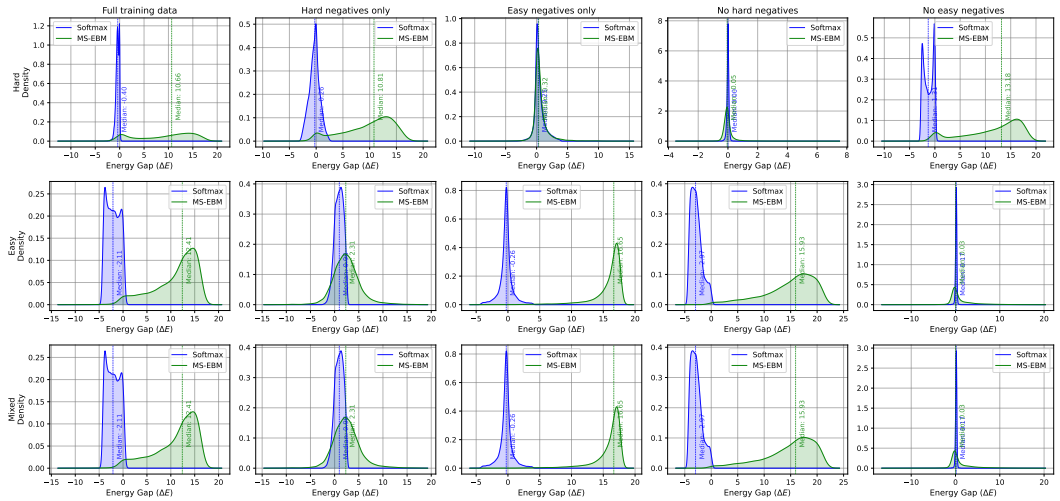

Figure 4: Energy-gap distributions across all training configurations (columns) and evaluation splits (rows). Softmax collapses near zero margins, leading to poor separation, while the MS-EBM shifts distributions rightwards and maintains clearer gaps between ID and OOD. This demonstrates how margin–structured training shapes the energy landscape, especially under hard and mixed OOD settings.

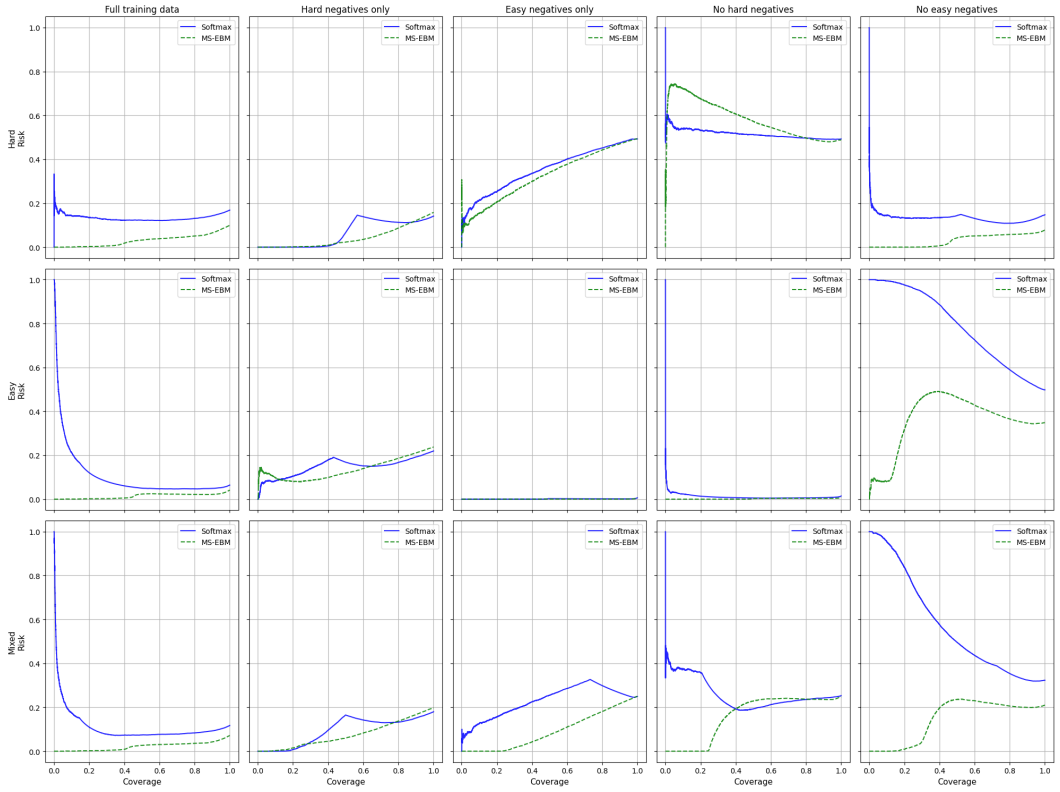

Figure 5: Risk-coverage curves across training configurations (columns) and evaluation splits (rows). Lower selective risk at higher coverage indicates better abstention. MS-EBM consistently achieves lower risk than Softmax, particularly under hard and mixed OOD conditions, while performance collapses when either easy or hard negatives are removed.

Table 2: Complete results across training configurations. We report AUROC, AUPR, FPR@95 (%), and DetErr on Hard, Easy, and Mixed OOD splits for all methods. Compared to the main paper's focused comparison (Hard/Mixed), this table expands to show every ablation setting (`full`, `hard only`, `easy only`, `no hard`, `no easy`). Results confirm that MS-EBM achieves the most reliable abstention on hard OOD, while performance collapses when either easy or hard negatives are removed.

Method	Hard				Easy				Mixed			
	AUROC	AUPR	FPR@95	DetErr	AUROC	AUPR	FPR@95	DetErr	AUROC	AUPR	FPR@95	DetErr
Full training data												
Softmax	0.895	0.922	69.4	0.147	0.937	0.963	69.7	0.064	0.916	0.944	68.9	0.110
MS-EBM	0.946	0.961	41.3	0.094	0.977	0.985	3.3	0.041	0.961	0.973	18.2	0.069
kNN (k=5)	0.855	0.832	0.6	0.429	0.998	0.999	0.6	0.015	0.926	0.936	0.6	0.221
Energy–OOD	0.895	0.922	69.5	0.169	0.937	0.963	69.3	0.064	0.916	0.944	68.7	0.117
ODIN	0.895	0.922	69.4	0.169	0.937	0.963	69.7	0.064	0.916	0.944	68.9	0.117
Mahalanobis	0.903	0.910	49.7	0.261	0.998	0.998	0.7	0.021	0.951	0.960	33.2	0.142
Hard negatives only												
Softmax	0.983	0.986	10.6	0.061	0.862	0.841	47.0	0.218	0.924	0.930	36.6	0.164
MS-EBM	0.982	0.984	10.2	0.063	0.840	0.808	50.1	0.237	0.912	0.918	40.5	0.179
kNN (k=5)	0.855	0.832	0.6	0.429	0.998	0.999	0.6	0.015	0.926	0.936	0.6	0.221
Energy–OOD	0.983	0.986	10.7	0.141	0.862	0.842	46.9	0.219	0.924	0.930	36.4	0.179
ODIN	0.958	0.963	25.8	0.144	0.887	0.868	40.8	0.195	0.923	0.925	35.0	0.170
Mahalanobis	0.812	0.759	66.5	0.232	0.586	0.603	93.5	0.434	0.702	0.690	88.7	0.330
Easy negatives only												
Softmax	0.678	0.650	82.2	0.366	1.000	1.000	0.1	0.006	0.840	0.874	70.6	0.239
MS-EBM	0.713	0.677	76.8	0.340	1.000	1.000	0.1	0.007	0.858	0.885	63.9	0.230
kNN (k=5)	0.855	0.832	0.6	0.429	0.998	0.999	0.6	0.015	0.926	0.936	0.6	0.221
Energy–OOD	0.678	0.650	82.1	0.494	1.000	1.000	0.1	0.006	0.840	0.874	70.7	0.250
ODIN	0.673	0.644	82.6	0.494	1.000	1.000	0.0	0.006	0.837	0.872	71.2	0.250
Mahalanobis	0.690	0.670	81.2	0.464	0.997	0.995	1.0	0.020	0.845	0.875	69.7	0.241
No hard negatives												
Softmax	0.489	0.515	95.9	0.486	0.992	0.995	0.1	0.014	0.743	0.818	91.8	0.252
MS-EBM	0.415	0.501	98.8	0.467	0.997	0.998	0.1	0.008	0.708	0.808	96.9	0.242
kNN (k=5)	0.855	0.832	0.6	0.429	0.998	0.999	0.6	0.015	0.926	0.936	0.6	0.221
Energy–OOD	0.489	0.515	95.9	0.493	0.992	0.995	0.1	0.014	0.743	0.818	91.8	0.253
ODIN	0.489	0.515	95.9	0.493	0.992	0.995	0.1	0.014	0.743	0.818	91.8	0.253
Mahalanobis	0.663	0.624	81.5	0.495	0.999	0.999	0.2	0.013	0.833	0.868	69.7	0.253
No easy negatives												
Softmax	0.916	0.943	67.8	0.116	0.126	0.336	100.0	0.498	0.522	0.696	100.0	0.312
MS-EBM	0.963	0.975	23.2	0.060	0.512	0.652	99.9	0.349	0.743	0.841	99.6	0.207
kNN (k=5)	0.855	0.832	0.6	0.429	0.998	0.999	0.6	0.015	0.926	0.936	0.6	0.221
Energy–OOD	0.916	0.943	68.0	0.147	0.126	0.336	100.0	0.498	0.522	0.695	100.0	0.323
ODIN	0.916	0.943	67.8	0.147	0.126	0.336	100.0	0.498	0.522	0.696	100.0	0.323
Mahalanobis	0.942	0.955	37.3	0.165	0.876	0.859	44.1	0.200	0.908	0.916	42.6	0.183

Table 3: Key hyperparameters and settings used across all experiments.

Item	Value
Encoder	BAAI/bge-m3 (frozen)
Projector	$1024 \to 512 \to 256$ (GELU), ℓ_2-normalized outputs
Seed	7
Epochs	20
Batch size	1024
Learning rate	5e-4
Weight decay	2e-4
Negative mining (k_{mine})	10
m_{sim}	0.2
m_E	1.0
m_{OOD}	6.0, $w_{\text{OOD}} = 0.8$
m_{HN}	5.0, $w_{\text{HN}} = 0.8$
Optimizer	AdamW
k-NN baseline	FAISS (cosine), $k = 5$, same BGE-M3 embeddings

Table 4: An excerpt of women's health guidelines and publications used to prepare the questions corpus in this research. Note: The full list has been submitted as the supplementary material.

Title	Synopsis	Year	Authors	Publisher
Workplace-based assessment: a new approach to existing tools	The article discusses the implementation challenges and revisions of workplace-based assessment (WPBA) tools in obstetrics and gynaecology, emphasizing the distinction between formative and summative assessments.	2014	William Parry-Smith, Ayesha Mahmud, Alex Landau, et al.	TOG
Contraceptive methods and issues around the menopause: an evidence update	The publication discusses recent advances in contraceptive methods available to perimenopausal women, issues related to menopause, and the integration of hormone replacement therapy with contraception.	2017	Shagaf H Bakour, Archana Hatti, Susan Whalen	TOG
Twin and triplet pregnancy	This guideline covers care for pregnant women and pregnant people with a twin or triplet pregnancy, aiming to reduce complications and improve outcomes.	2019	–	NICE
Management of sickle cell disease in pregnancy. A British Society for Haematology Guideline	This guideline describes the management of sickle cell disease in pregnancy, covering preconception, antenatal, intrapartum, and postnatal care, with updates on genetic diagnosis, medication review, and antenatal care recommendations.	2021	Eugene Oteng-Ntim, Sue Pavord, Richard Howard, et al.	Br J Haematol
Inducing labour	This guideline covers the circumstances for inducing labour, methods of induction, assessment, monitoring, pain relief, and managing complications to improve advice and care for pregnant women considering or undergoing induction of labour.	2021	–	NICE
Saving Lives, Improving Mothers' Care State of the Nation Report	The report presents surveillance findings and lessons learned to inform maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths, focusing on thrombosis, thromboembolism, malignancy, ectopic pregnancy, and the care of recent migrants with language difficulties from 2020–2022.	2024	MBRRACE-UK	NPEU, Univ. Oxford
Laparoscopy in urogynaecology	This article discusses the advancements and challenges in laparoscopic urogynaecological surgery, focusing on procedures for prolapse and stress incontinence, and highlights the importance of training and patient choice.	2018	Rajvinder Khasriya, Arvind Vashisht, Alfred Cutner	TOG
Failed hysteroscopy and further management strategies	This article explores various methods to overcome cervical stenosis in hysteroscopy, highlighting techniques such as pharmacological, mechanical, hygroscopic, and ultrasound-guided dilatation.	2016	Sophie Relph, Tessa Lawton, Mark Broadbent, et al.	TOG