Evaluating Foundation Models: From Biosignals and Pathology Images to LLM Confidence

Fredrik K. Gustafsson
University of Oxford
Department of Engineering Science
www.fregu856.com

CHI Lab Summer Symposium
June 10, 2026

About Me

Postdoc in the group of David Clifton at the **University of Oxford**, since July 2025.

Background:

- Dec 2023 - Jun 2025: Postdoc at **Karolinska Institutet** in Stockholm, ML for *computational pathology*, in the group of Mattias Rantalainen.
- 2018 - 2023: PhD in *Machine Learning*, **Uppsala University**.
  - Thesis: *Towards Accurate and Reliable Deep Regression Models*.
    Supervisors: Thomas Schön & Martin Danelljan.
- 2013 - 2018: BSc & MSc in *Electrical Engineering*, **Linköping University**.
  - 2016 - 2017: Graduate exchange student, **Stanford University**.

Machine learning for healthcare with a particular focus on biosignals and wearables.

More broadly, I am interested in how to build and evaluate *reliable machine learning* models, for applications within *data-driven medicine and healthcare*.

About the Presentation

I will briefly describe three papers evaluating different types of foundation models:

I: Benchmarking Pathology Foundation Models for Breast Cancer Survival Prediction
Fredrik K. Gustafsson, Constance Boissin, Johan Vallon-Christersson, David A. Clifton, Mattias Rantalainen
Preprint, 2026

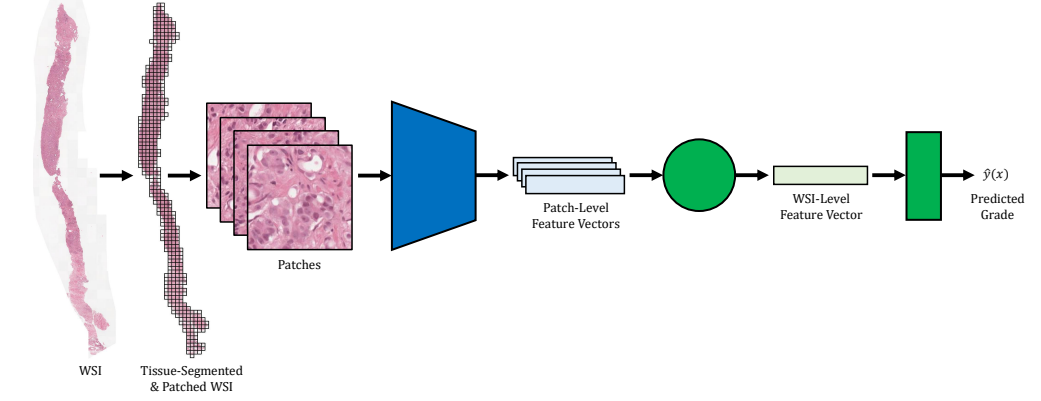
II: SignalMC-MED: A Multimodal Benchmark for Evaluating Biosignal Foundation Models on Single-Lead ECG and PPG
Fredrik K. Gustafsson, Xiao Gu, Mattia Carletti, Patitapaban Palo, David W. Eyre, David A. Clifton
Preprint, 2026

III: BAS: A Decision-Theoretic Approach to Evaluating Large Language Model Confidence
Sean Wu, Fredrik K. Gustafsson*, Edward Phillips, Boyan Gao, Anshul Thakur, David A. Clifton*
Preprint, 2026

Background: Computational Pathology

Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).

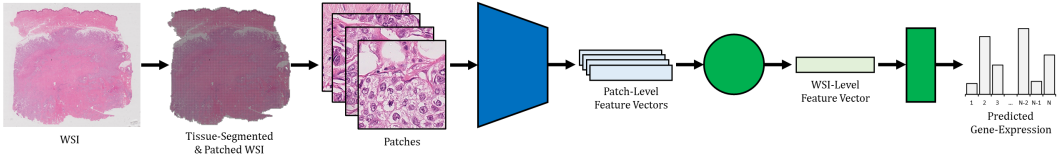
Given datasets of (WSI, label) pairs, models can be trained for applications such as *histological grading*, patient outcome prediction, and prediction of various biomarkers.



Background: Computational Pathology

Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).

Given datasets of (WSI, label) pairs, models can be trained for applications such as histological grading, patient outcome prediction, and *prediction of various biomarkers*.



Background: Pathology Foundation Models (PFMs)

Foundation models are large models trained on *large amounts of unlabeled data* using *self-supervised learning*. They are intended to be general-purpose feature extractors.

Self-supervised learning enables models to be trained on "raw" unlabeled data. Large collections of unlabeled WSIs – *WSIs without known clinical info, patient outcomes or any other type of annotations* – can thus be directly utilized in model training.

Has recently become a popular research direction within computational pathology:

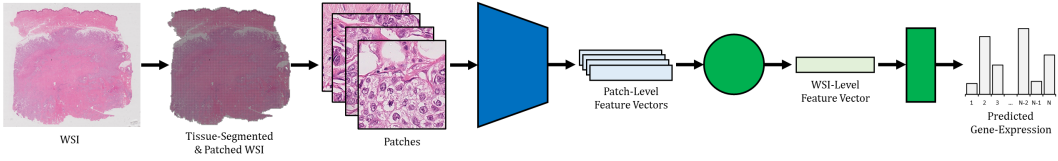
UNI: **Towards a General-Purpose Foundation Model for Computational Pathology**
Nature Medicine, 2024

Prov-GigaPath: **A Whole-Slide Foundation Model for Digital Pathology from Real-World Data**
Nature, 2024

Virchow: **A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection**
Nature Medicine, 2024

⋮

Background: Typical Workflow for Application of PFMs



- Tissue-segment each WSI and divide it into image patches (e.g. 224×224 pixels).
- Use a *frozen* foundation model to extract feature vectors for all images patches in each WSI (*typical range: 5,000 - 25,000 image patches per WSI*).
- Train a small model that, for each WSI, takes the extracted patch-level feature vectors as input and outputs a WSI-level prediction (standard supervised training).

Benchmarking Pathology Foundation Models for Breast Cancer Survival Prediction

Fredrik K. Gustafsson, Constance Boissin, Johan Vallon-Christersson, David A. Clifton, Mattias Rantalainen

Preprint, 2026

Model Name	Architecture	Size	Feature Dimension	Pretraining Data
Resnet-IN [7]	ResNet-50	25M	1024	1.3M natural images
CTransPath [22]	CNN + Swin-T	22M	768	30K WSIs
RetCCL [23]	ResNet-50	25M	2048	32K WSIs
UNI [2]	ViT-L	307M	1024	100K WSIs
UNI2-h [13]	ViT-H	682M	1536	350K WSIs
H-optimus-0 [18]	ViT-G	1.1B	1536	500K WSIs
H-optimus-1 [1]	ViT-G	1.1B	1536	1M WSIs
Prov-GigaPath [25]	ViT-G	1.1B	1536	170K WSIs
Virchow [21]	ViT-H	632M	2560	1.5M WSIs
Virchow2 [26]	ViT-H	632M	2560	3.1M WSIs
H0-mini [5]	ViT-B	86M	768	500K + 6K WSIs
CONCH [11]	ViT-B	86M	512	21K WSIs + 1.1M image-text pairs
CONCHv1.5 [3]	ViT-L	307M	768	N/A

Rank	Model Name	Model Ranks	Mean Model Rank (↓)
1	H-optimus-1	1,1,1,3	1.5
2	H0-mini	2,2,4,5	3.25
3	H-optimus-0	5,5,5,3	4.5
4	CONCHv1.5	8,8,2,1	4.75
5	UNI2-h	4,4,6,6	5
6	Virchow2	3,3,7,8	5.25
7	Virchow	10,10,2,2	6
8	CONCH	6,6,8,7	6.75
9	Prov-GigaPath	6,6,9,9	7.5
10	UNI	9,9,10,10	9.5
11	CTransPath	12,12,11,11	11.5
12	RetCCL	11,11,13,13	12
13	Resnet-IN	13,13,12,12	12.5

I: Benchmarking PFMs for Survival Prediction

We provide the first large-scale, externally validated benchmark of PFMs for breast cancer survival prediction, using three independent cohorts with over *5,400 patients*.

We evaluate 13 models: a natural-image baseline, two early pathology-specific models, seven state-of-the-art PFMs, a compact distilled PFM, and two vision-language PFMs.

Downstream survival models are trained on a dataset of *2,315* breast cancer patients and evaluated on two independent external datasets comprising *3,119* patients in total.

Models are evaluated under four settings: RFS and PFS, each assessed both for the full cohort ('All Patients') and for the clinically relevant 'ER+ & HER2-' patient subgroup.

The combined evaluation set of 3,119 patients contains 615 RFS events and 233 PFS events, with 2,524 patients (80.9% of the full set), 475 RFS events (77.2%) and 157 PFS events (67.4%) in the 'ER+ & HER2-' patient subgroup.

I: Benchmarking PFMs for Survival Prediction - Main Results

Rank	Model Name	C-index (↑)
1	H-optimus-1	0.678 (0.656 – 0.700)
2	H0-mini	0.676 (0.653 – 0.698)
3	Virchow2	0.675 (0.653 – 0.698)
4	UNI2-h	0.667 (0.644 – 0.689)
5	H-optimus-0	0.664 (0.642 – 0.686)
6	CONCH	0.663 (0.641 – 0.685)
6	Prov-GigaPath	0.663 (0.640 – 0.686)
8	CONCHv1.5	0.660 (0.637 – 0.683)
9	UNI	0.657 (0.635 – 0.680)
10	Virchow	0.656 (0.633 – 0.679)
11	RetCCL	0.645 (0.622 – 0.668)
12	CTransPath	0.632 (0.609 – 0.655)
13	Resnet-IN	0.612 (0.588 – 0.635)

(a) **RFS – All Patients**.

Rank	Model Name	C-index (↑)
1	H-optimus-1	0.670 (0.645 – 0.695)
2	H0-mini	0.668 (0.642 – 0.693)
3	Virchow2	0.665 (0.639 – 0.690)
4	UNI2-h	0.663 (0.637 – 0.688)
5	H-optimus-0	0.660 (0.633 – 0.685)
6	CONCH	0.657 (0.631 – 0.681)
6	Prov-GigaPath	0.657 (0.631 – 0.682)
8	CONCHv1.5	0.651 (0.625 – 0.676)
9	UNI	0.648 (0.622 – 0.674)
10	Virchow	0.639 (0.613 – 0.665)
11	RetCCL	0.636 (0.609 – 0.662)
12	CTransPath	0.627 (0.601 – 0.652)
13	Resnet-IN	0.600 (0.572 – 0.627)

(b) **RFS – Patient Subgroup (ER+ & HER2-)**.

Rank	Model Name	C-index (↑)
1	H-optimus-1	0.702 (0.666 – 0.737)
2	Virchow	0.695 (0.660 – 0.729)
2	CONCHv1.5	0.695 (0.660 – 0.729)
4	H0-mini	0.692 (0.655 – 0.728)
5	H-optimus-0	0.686 (0.651 – 0.722)
6	UNI2-h	0.683 (0.647 – 0.719)
7	Virchow2	0.681 (0.645 – 0.718)
8	CONCH	0.675 (0.639 – 0.710)
9	Prov-GigaPath	0.673 (0.636 – 0.710)
10	UNI	0.671 (0.634 – 0.708)
11	CTransPath	0.647 (0.609 – 0.686)
12	Resnet-IN	0.628 (0.590 – 0.666)
13	RetCCL	0.627 (0.589 – 0.666)

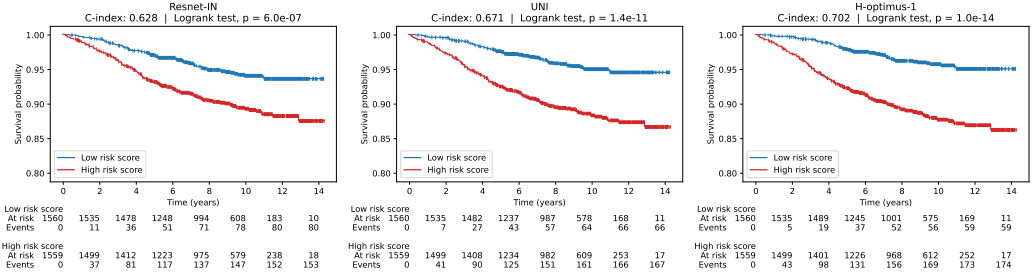
(c) **PFS – All Patients**.

Rank	Model Name	C-index (↑)
1	CONCHv1.5	0.680 (0.636 – 0.721)
2	Virchow	0.678 (0.634 – 0.720)
3	H-optimus-1	0.670 (0.626 – 0.715)
3	H-optimus-0	0.670 (0.624 – 0.714)
5	H0-mini	0.666 (0.619 – 0.711)
6	UNI2-h	0.657 (0.611 – 0.702)
7	CONCH	0.656 (0.612 – 0.700)
8	Virchow2	0.655 (0.608 – 0.700)
9	Prov-GigaPath	0.641 (0.594 – 0.687)
10	UNI	0.638 (0.592 – 0.683)
11	CTransPath	0.622 (0.573 – 0.670)
12	Resnet-IN	0.606 (0.557 – 0.655)
13	RetCCL	0.604 (0.555 – 0.653)

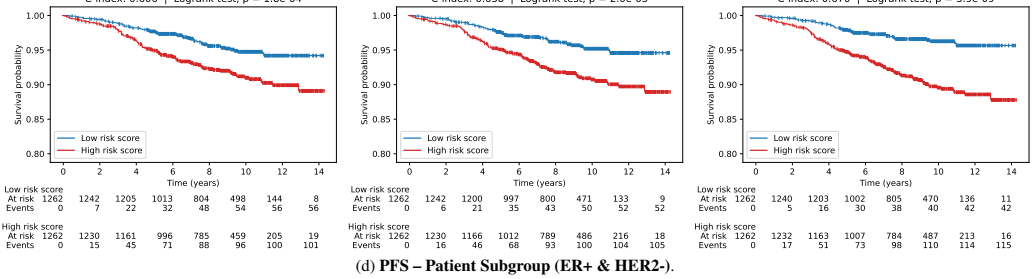
(d) **PFS – Patient Subgroup (ER+ & HER2-)**.

Rank	Model Name	Model Ranks	Mean Model Rank (↓)
1	H-optimus-1	1,1,1,3	1.5
2	H0-mini	2,2,4,5	3.25
3	H-optimus-0	5,5,5,3	4.5
4	CONCHv1.5	8,8,2,1	4.75
5	UNI2-h	4,4,6,6	5
6	Virchow2	3,3,7,8	5.25
7	Virchow	10,10,2,2	6
8	CONCH	6,6,8,7	6.75
9	Prov-GigaPath	6,6,9,9	7.5
10	UNI	9,9,10,10	9.5
11	CTransPath	12,12,11,11	11.5
12	RetCCL	11,11,13,13	12
13	Resnet-IN	13,13,12,12	12.5

I: Benchmarking PFMs for Survival Prediction - Risk Stratification

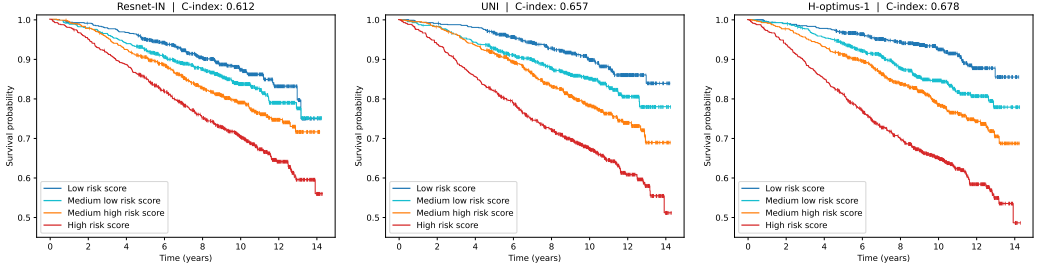


(c) **PFS – All Patients**.

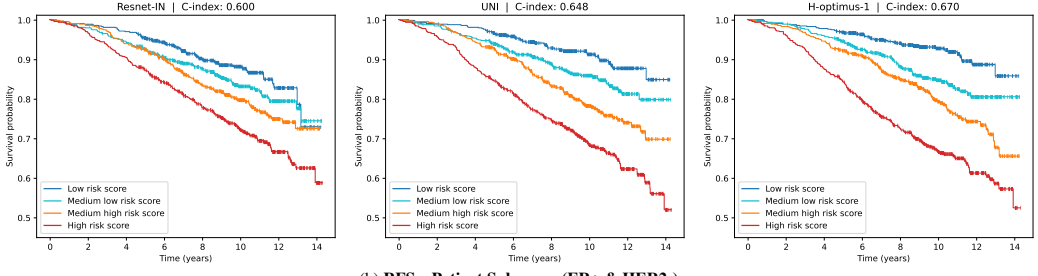


(d) **PFS – Patient Subgroup (ER+ & HER2-)**.

I: Benchmarking PFMs for Survival Prediction - Risk Stratification

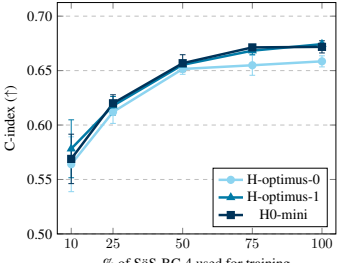


(a) **RFS – All Patients**.

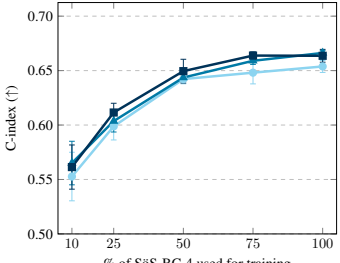


(b) **RFS – Patient Subgroup (ER+ & HER2-)**.

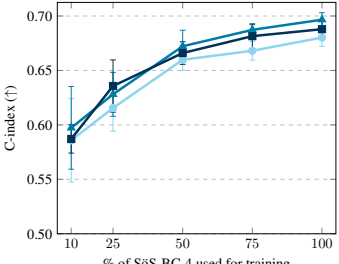
I: Benchmarking PFMs for Survival Prediction - Detailed Comparison



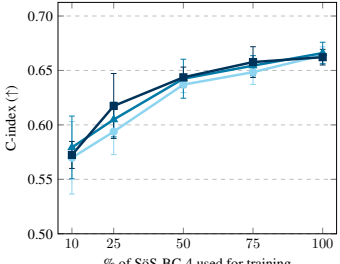
(a) **RFS – All Patients**.



(b) **RFS – Patient Subgroup (ER+ & HER2-)**.

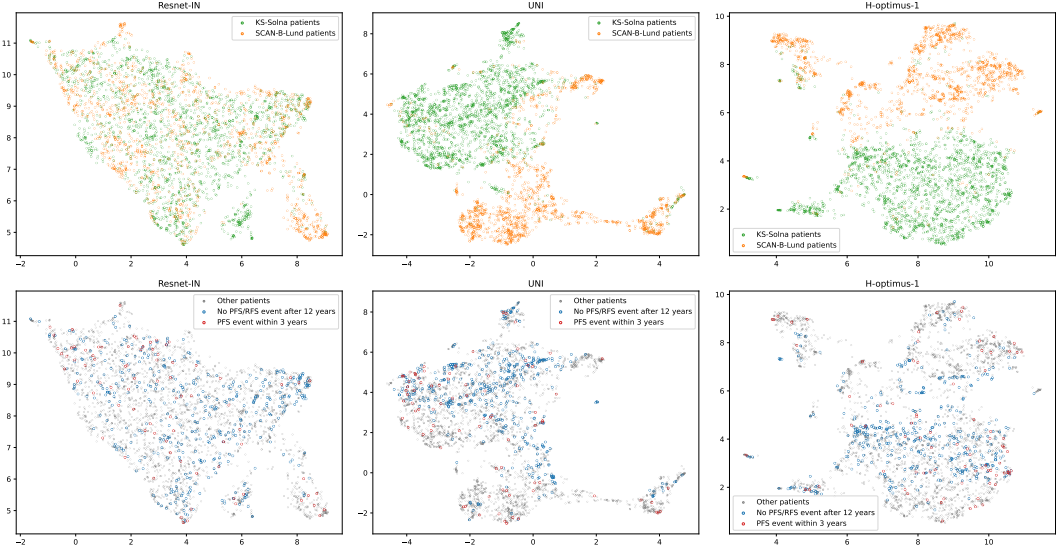


(c) **PFS – All Patients**.



(d) **PFS – Patient Subgroup (ER+ & HER2-)**.

I: Benchmarking PFMs for Survival Prediction - Feature Space Analysis



I: Benchmarking PFMs for Survival Prediction - Main Takeaways

(1/4) H-optimus-1 achieves the strongest overall performance, but absolute differences between top-performing PFMs are small and confidence intervals substantially overlap.

(2/4) Across model families, consistent generational improvements are observed, with second-generation PFMs (H-optimus-1, CONCHv1.5, UNI2-h, Virchow2) outperforming their first-generation counterparts. However, these gains remain modest despite substantial increases in pretraining data scale, suggesting diminishing returns from scaling alone.

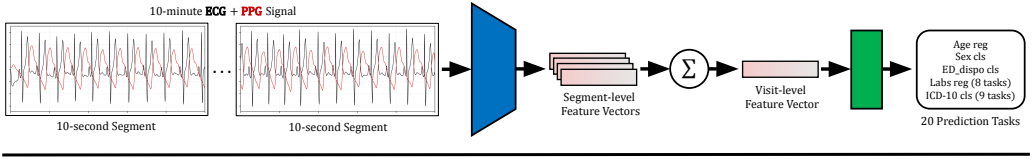
(3/4) Model size is not a reliable predictor of performance, as smaller and more efficient models can match or exceed much larger architectures, emphasizing the importance of training strategy and pretraining data quality over model scaling.

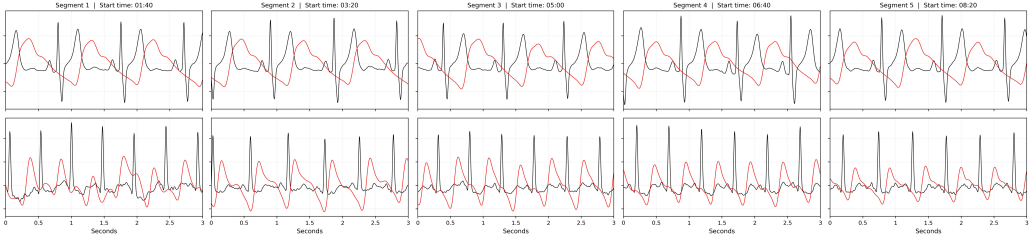
(4/4) The distilled H0-mini achieves the second-best overall ranking and slightly outperforms its teacher model H-optimus-0, while using less than 8% of the parameters and enabling much faster feature extraction. Knowledge distillation can thus yield highly efficient PFMs without sacrificing performance, making it a particularly promising approach for practical deployment.

SignalMC-MED: A Multimodal Benchmark for Evaluating Biosignal Foundation Models on Single-Lead ECG and PPG

Fredrik K. Gustafsson, Xiao Gu, Mattia Carletti, Patitapaban Palo, David W. Eyre, David A. Clifton

Preprint, 2026



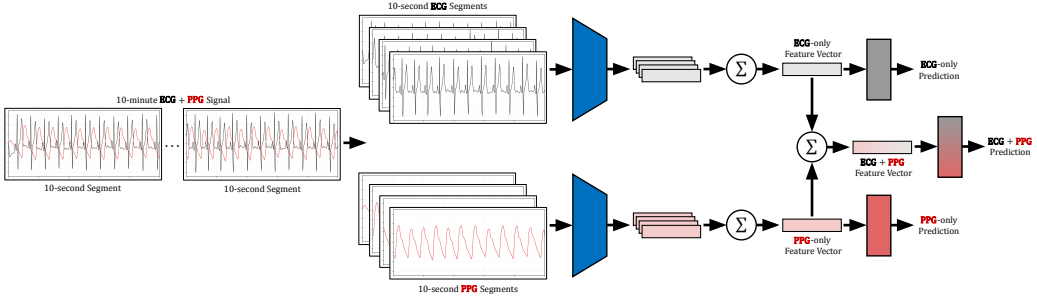


II: SignalMC-MED for Evaluating Biosignal FMs

Utilizing the MC-MED dataset, we introduce SignalMC-MED, a benchmark of *22,256* emergency department visits with synchronized 10-minute single-lead ECG and PPG.

SignalMC-MED includes 20 tasks spanning prediction of demographics, emergency department disposition, lab value regression, and detection of prior ICD-10 diagnoses.

Using this benchmark, we perform a systematic evaluation of representative time-series and biosignal FMs across ECG-only, PPG-only, and ECG + PPG settings.



II: SignalMC-MED for Evaluating Biosignal FMs - Model Rankings

Model Name	Size	Feature Dimension	Sampling Frequency	Pretraining Data
MOMENT-small [10]	35M	512	Mixed	13M time series; 1.2B timestamps
MOMENT-base	110M	768	—	—
MOMENT-large	341M	1024	—	—
Chronos-Bolt-tiny [4]	9M	256	Mixed	100B time series observations
Chronos-Bolt-mini	21M	384	—	—
Chronos-Bolt-small	48M	512	—	—
Chronos-Bolt-base	205M	768	—	—
PaPaGei [24]	6M	512	125Hz	20M 10-sec PPG segments from 13.5K subjects
D-BETA [23]	62M	768	500Hz	0.78M 10-sec, 12-lead ECGs from ∼150K subjects
ECGFounder [16]	31M	1024	500Hz	10.8M ECGs ("predominantly 10-sec, 12-lead") from 1.8M subjects
xECG [18]	57M	1024	100Hz	2.4M 7–10-sec, 12-lead ECGs from ∼1.7M subjects + 75 30-min, 12-lead ECGs
CSFM-tiny [12]	43M	768	250Hz	2.3M 10-sec, 12-lead ECGs + 0.27M
CSFM-base	109M	768	—	paired 10-sec PPGs and single-lead
CSFM-large	334M	1024	—	ECGs from 1.7M subjects

Rank	Model Name	Model Ranks	Mean Model Rank (↓)
1	CSFM-base	7,5,5,2,4	4.6
2	xECG-10min	10,3,7,4,8	6.4
3	ECGFounder	11,12,9,7,7	9.2
4	Domain Features	17,23,4,6,1	10.2
5	MOMENT-base	21,11,19,15,16	16.4
6	D-BETA	22,21,18,20,12	18.6
7	Chronos-Bolt-small	19,18,20,18,19	18.8
8	PaPaGei	24,24,24,23,24	23.8

(a) ECG-only.

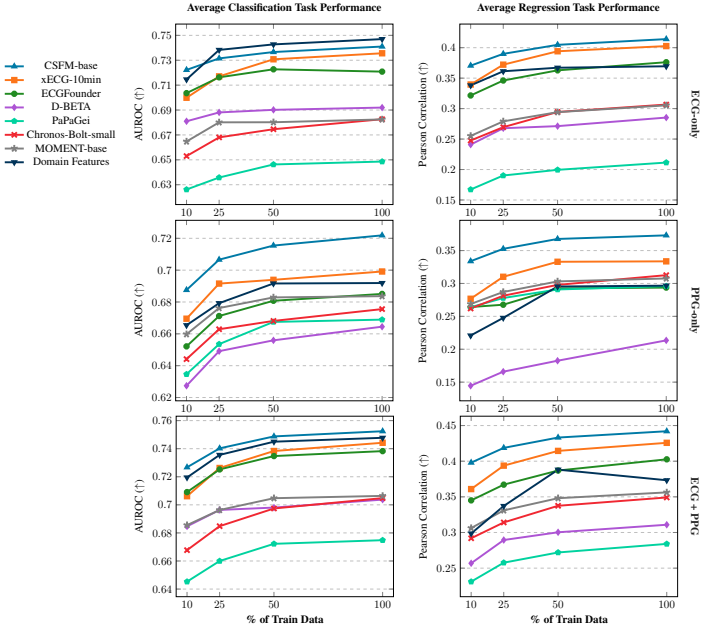
Rank	Model Name	Model Ranks	Mean Model Rank (↓)
1	CSFM-base	4,9,3,8,9	6.6
2	xECG-10min	6,10,10,12,14	10.4
3	MOMENT-base	12,17,12,13,17	14.2
4	Domain Features	15,13,13,22,15	15.6
5	Chronos-Bolt-small	13,15,17,16,22	16.6
6	ECGFounder	16,16,14,19,18	16.6
7	PaPaGei	18,22,22,17,21	20.0
8	D-BETA	23,20,21,24,23	22.2

(b) PPG-only.

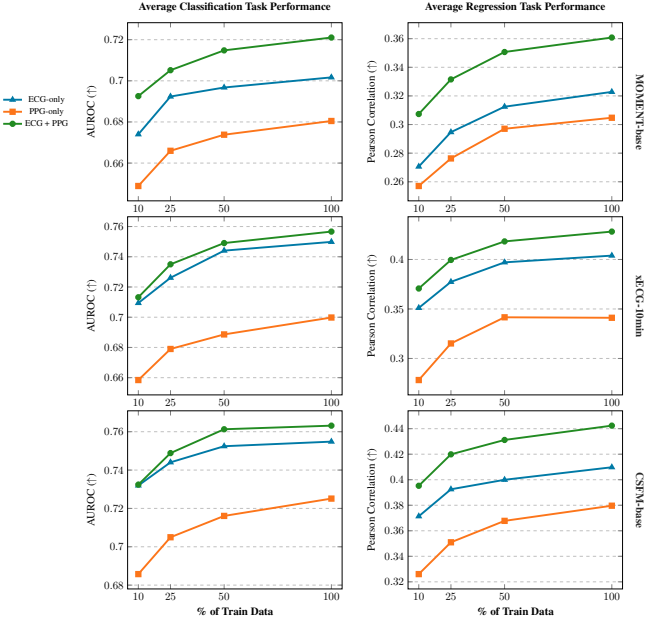
Rank	Model Name	Model Ranks	Mean Model Rank (↓)
1	CSFM-base	1,1,1,1,3	1.4
2	xECG-10min	2,2,6,3,6	3.8
3	ECGFounder	3,4,8,5,5	5.0
4	Domain Features (concat)	8,7,2,9,2	5.6
5	MOMENT-base	5,6,11,10,11	8.6
6	Chronos-Bolt-small	9,8,16,11,13	11.4
7	D-BETA	14,14,15,14,10	13.4
8	PaPaGei	20,19,23,21,20	20.6

(c) ECG + PPG.

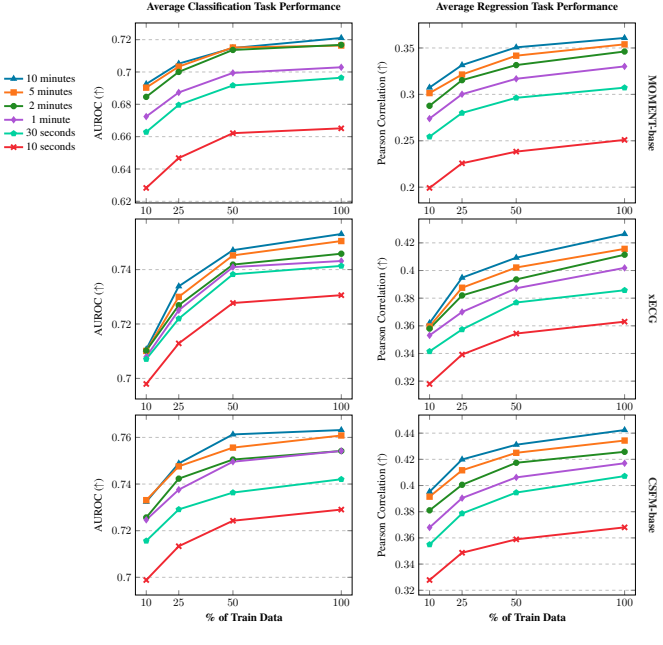
II: SignalMC-MED for Evaluating Biosignal FMs - Model Comparison



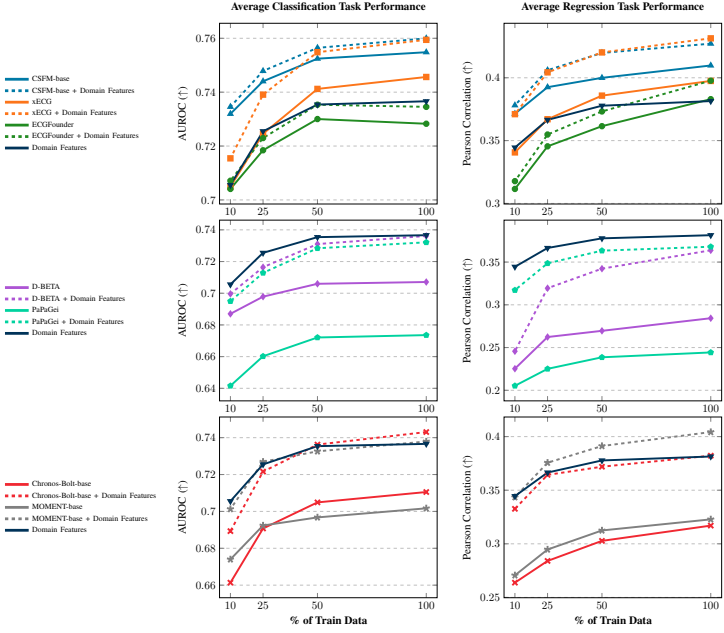
II: SignalMC-MED for Evaluating Biosignal FMs - Multimodal Fusion



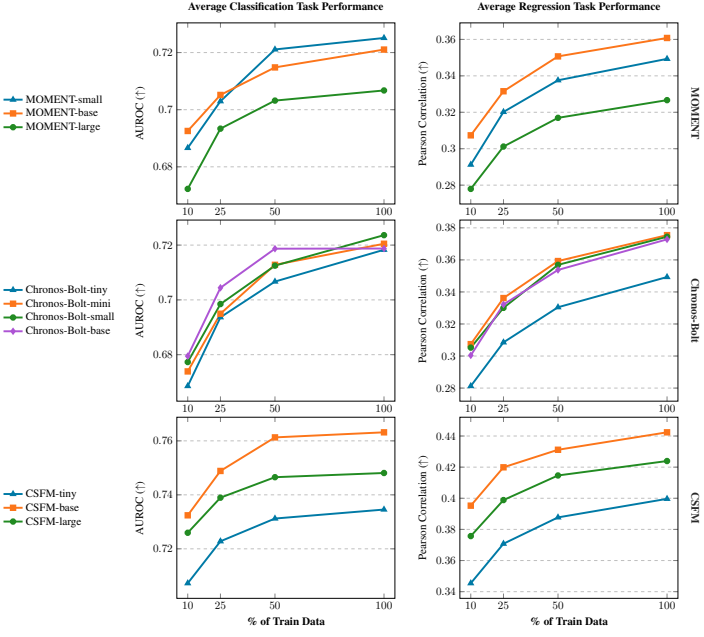
II: SignalMC-MED for Evaluating Biosignal FMs - Signal Length



II: SignalMC-MED for Evaluating Biosignal FMs - ECG Domain Features



II: SignalMC-MED for Evaluating Biosignal FMs - Model Size



II: SignalMC-MED for Evaluating Biosignal FMs - Main Takeaways

(1/5) Domain-specific biosignal FMs should be preferred over general time-series FMs for cardiovascular prediction tasks.

(2/5) Multimodal ECG and PPG data should be leveraged whenever available, as even simple late-fusion strategies yield consistent improvements over unimodal inputs.

(3/5) Longer signal segments should be prioritized, as increased temporal coverage (up to the full 10-minute signal) consistently improves performance across models and tasks.

(4/5) Larger FM variants should not be assumed to outperform smaller ones without task-specific validation.

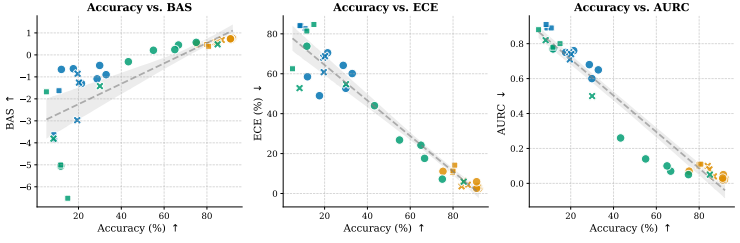
(5/5) Hand-crafted ECG domain features should be included as a simple yet strong baseline to verify that FMs provide meaningful gains, and can offer complementary benefits when combined with learned FM representations.

BAS: A Decision-Theoretic Approach to Evaluating Large Language Model Confidence

Sean Wu, Fredrik K. Gustafsson*, Edward Phillips, Boyan Gao, Anshul Thakur, David A. Clifton*

Preprint, 2026

Model	SimpleQA Acc↑	SimpleQA BAS↑	SimpleQA ECE↓	SimpleQA AURC↓	MedQA Acc↑	MedQA BAS↑	MedQA ECE↓	MedQA AURC↓	AIME Acc↑	AIME BAS↑	AIME ECE↓	AIME AURC↓
Frontier / Large Models												
DS-R1	$29.9_{2.9}$	$-0.48_{0.06}$	$52.7_{2.7}$	$0.60_{0.04}$	$92.1_{1.5}$	$0.76_{0.03}$	$2.6_{1.2}$	$0.03_{0.01}$	$66.7_{11.7}$	$0.45_{0.20}$	$17.6_{8.1}$	$0.07_{0.06}$
DS-V3.2	$17.6_{2.3}$	$-0.63_{0.06}$	$49.0_{2.9}$	$0.75_{0.04}$	$91.8_{1.5}$	$0.76_{0.03}$	$3.3_{1.3}$	$0.02_{0.01}$	$65.0_{11.7}$	$0.24_{0.28}$	$24.2_{10.0}$	$0.10_{0.07}$
GPT-4o	$21.4_{2.6}$	$-1.30_{0.06}$	$70.5_{2.5}$	$0.76_{0.04}$	$91.3_{1.6}$	$0.70_{0.05}$	$3.5_{1.5}$	$0.05_{0.01}$	$11.7_{7.5}$	$-5.06_{0.98}$	$73.8_{10.6}$	$0.77_{0.12}$
GPT-oss	$12.0_{2.0}$	$-0.66_{0.05}$	$58.5_{2.1}$	$0.77_{0.04}$	$91.0_{1.6}$	$0.74_{0.04}$	$2.6_{1.2}$	$0.02_{0.01}$	$75.0_{10.8}$	$0.57_{0.22}$	$7.2_{6.2}$	$0.05_{0.04}$
Grok-3	$32.9_{2.9}$	$-0.90_{0.08}$	$60.1_{2.8}$	$0.65_{0.04}$	$91.0_{1.6}$	$0.73_{0.03}$	$5.9_{1.4}$	$0.03_{0.01}$	$55.0_{12.5}$	$0.21_{0.22}$	$26.8_{9.9}$	$0.14_{0.08}$
Mist(L)	$28.7_{2.9}$	$-1.09_{0.09}$	$64.2_{2.7}$	$0.68_{0.04}$	$75.3_{2.3}$	$0.51_{0.05}$	$11.1_{1.9}$	$0.07_{0.01}$	$43.3_{11.7}$	$-0.31_{0.35}$	$44.0_{11.3}$	$0.26_{0.13}$
Mid-tier Models												
G3 Mini	$19.6_{2.5}$	$-0.86_{0.07}$	$60.8_{2.6}$	$0.71_{0.04}$	$86.9_{1.8}$	$0.68_{0.04}$	$4.4_{1.5}$	$0.04_{0.01}$	$85.0_{9.2}$	$0.48_{0.44}$	$5.9_{6.3}$	$0.05_{0.09}$
Llama 3.3	$19.4_{2.4}$	$-2.97_{0.24}$	$68.1_{2.6}$	$0.75_{0.04}$	$84.0_{2.0}$	$0.57_{0.04}$	$3.6_{1.5}$	$0.10_{0.02}$	$8.3_{7.5}$	$-3.81_{1.04}$	$52.8_{12.3}$	$0.82_{0.12}$
Mist(M)	$20.2_{2.5}$	$-1.26_{0.12}$	$68.8_{2.4}$	$0.74_{0.04}$	$84.8_{2.0}$	$0.57_{0.05}$	$6.4_{1.9}$	$0.08_{0.02}$	$30.0_{11.7}$	$-1.42_{0.75}$	$54.8_{10.8}$	$0.50_{0.17}$
Small / Efficient Models												
4o-mini	$8.9_{1.8}$	$-3.86_{0.21}$	$84.2_{1.8}$	$0.89_{0.03}$	$79.8_{2.2}$	$0.48_{0.05}$	$10.6_{2.1}$	$0.10_{0.02}$	$11.7_{7.5}$	$-5.01_{0.95}$	$81.4_{9.3}$	$0.78_{0.13}$
Mist(S)	$10.9_{2.0}$	$-1.63_{0.07}$	$82.6_{1.9}$	$0.89_{0.03}$	$80.0_{2.2}$	$0.47_{0.05}$	$11.4_{2.1}$	$0.11_{0.02}$	$5.0_{5.8}$	$-1.68_{0.56}$	$62.5_{11.4}$	$0.88_{0.11}$
Phi-4	$8.5_{1.7}$	$-3.63_{0.22}$	$84.1_{1.8}$	$0.91_{0.03}$	$80.8_{2.2}$	$0.39_{0.07}$	$14.2_{2.1}$	$0.11_{0.02}$	$15.0_{9.2}$	$-6.52_{0.87}$	$84.7_{9.1}$	$0.80_{0.12}$



III: BAS for Evaluating LLM Confidence

Can we trust the self-reported confidence of LLMs?

We consider the setting where only text-level access is available, and models are prompted to produce both an answer and an associated confidence estimate.

Main contributions:

A decision-theoretic metric for confidence evaluation: We introduce BAS, a confidence reliability metric derived from an explicit answer-or-abstain utility model.

A comprehensive benchmark of LLM confidence: We construct a benchmark of self-reported LLM confidence across a diverse set of models, tasks and domains, and evaluate BAS alongside widely used metrics such as ECE and AURC.

Given a prompt x, the model produces a response R together with a scalar confidence value $s \in [0, 1)$. $Z \in \{0, 1\}$ denotes the correctness of the response ($Z = 1$ if correct).

$$\mathrm{BAS} = \frac{1}{N} \sum_{i=1}^{N} U(s_i, Z_i), \qquad U(s, Z) = \begin{cases} s, & Z = 1, \\ s + \ln(1 - s), & Z = 0. \end{cases}$$

Correct predictions receive utility proportional to the model confidence s, while $\ln(1 - s)$ diverges to $-\infty$ for highly confident errors. Thus, $\mathrm{BAS} \in (-\infty, 1]$, where higher values indicate better decision-level reliability.

III: BAS for Evaluating LLM Confidence - Relation to Existing Metrics

BAS is most closely related to proper scoring rules such as log loss, sharing a sensitivity to high-confidence errors, but differs in imposing an *asymmetric* penalty that strongly prioritizes avoiding model *overconfidence*.

Property	Accuracy	ECE	AURC	Brier Score	Log Loss	**BAS**
Confidence magnitude	✗	✓	✗	✓	✓	✓
Calibration	✗	✓	✗	✓	✓	✓
Ranking	✗	✗	✓	✓	✓	✓
Proper scoring rule	✗	✗	✗	✓	✓	✓
Strongly penalizes overconfidence	✗	✗	✗	✗	✓	✓
Asymmetric overconfidence penalty	✗	✗	✗	✗	✗	✓
Decision-theoretic derivation	✗	✗	✗	✗	✗	✓

III: BAS for Evaluating LLM Confidence - Main Benchmark Results

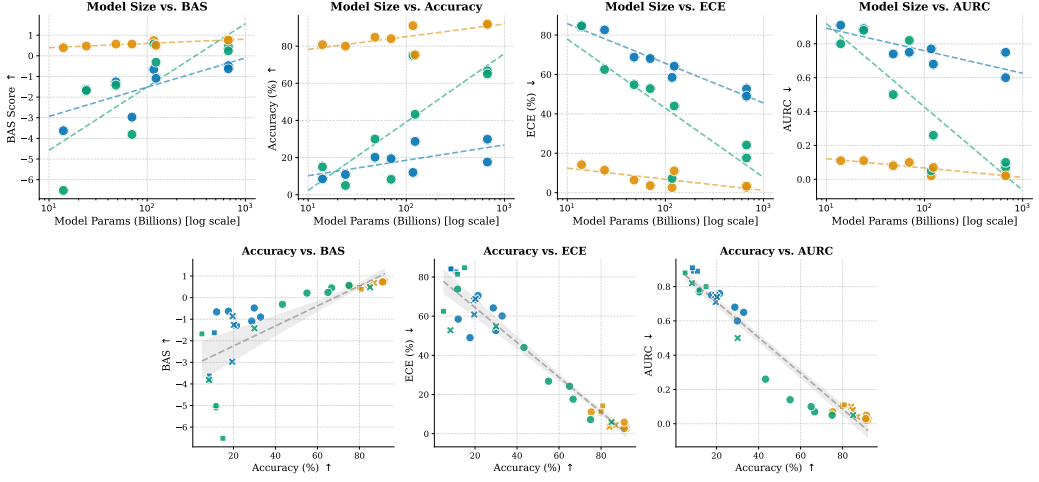
Confidence reliability varies substantially across tasks, and even frontier models can exhibit severe overconfidence in challenging settings.

Model	SimpleQA Acc↑	SimpleQA BAS↑	SimpleQA ECE↓	SimpleQA AURC↓	MedQA Acc↑	MedQA BAS↑	MedQA ECE↓	MedQA AURC↓	AIME Acc↑	AIME BAS↑	AIME ECE↓	AIME AURC↓
Frontier / Large Models												
DS-R1	$29.9_{2.9}$	$-0.48_{0.06}$	$52.7_{2.7}$	$0.60_{0.04}$	$92.1_{1.5}$	$0.76_{0.03}$	$2.61_{1.2}$	$0.03_{0.01}$	$66.7_{11.7}$	$0.45_{0.20}$	$17.6_{8.1}$	$0.07_{0.06}$
DS-V3.2	$17.6_{2.3}$	$-0.63_{0.06}$	$49.0_{2.9}$	$0.75_{0.04}$	$91.8_{1.5}$	$0.76_{0.03}$	$3.3_{1.3}$	$0.02_{0.01}$	$65.0_{11.7}$	$0.24_{0.28}$	$24.2_{10.0}$	$0.10_{0.07}$
GPT-4o	$21.4_{2.6}$	$-1.30_{0.06}$	$70.5_{2.5}$	$0.76_{0.04}$	$91.3_{1.6}$	$0.70_{0.05}$	$3.5_{1.5}$	$0.05_{0.01}$	$11.7_{7.5}$	$-5.06_{0.98}$	$73.8_{10.6}$	$0.77_{0.12}$
GPT-oss	$12.0_{2.0}$	$-0.66_{0.05}$	$58.5_{2.1}$	$0.77_{0.04}$	$91.0_{1.6}$	$0.74_{0.04}$	$2.61_{1.2}$	$0.02_{0.01}$	$75.0_{10.8}$	$0.57_{0.22}$	$7.2_{6.2}$	$0.05_{0.04}$
Grok-3	$32.9_{2.9}$	$-0.90_{0.08}$	$60.1_{2.8}$	$0.65_{0.04}$	$91.0_{1.6}$	$0.73_{0.03}$	$5.9_{1.4}$	$0.03_{0.01}$	$55.0_{12.5}$	$0.21_{0.22}$	$26.8_{9.9}$	$0.14_{0.08}$
Mist(L)	$28.7_{2.9}$	$-1.09_{0.09}$	$64.2_{2.7}$	$0.68_{0.04}$	$75.3_{2.3}$	$0.51_{0.05}$	$11.1_{1.9}$	$0.07_{0.01}$	$43.3_{11.7}$	$-0.31_{0.35}$	$44.0_{11.3}$	$0.26_{0.13}$
Mid-tier Models												
G3 Mini	$19.6_{2.5}$	$-0.86_{0.07}$	$60.8_{2.6}$	$0.71_{0.04}$	$86.9_{1.8}$	$0.68_{0.04}$	$4.4_{1.5}$	$0.04_{0.01}$	$85.0_{9.2}$	$0.48_{0.44}$	$5.9_{6.3}$	$0.05_{0.09}$
Llama 3.3	$19.4_{2.4}$	$-2.97_{0.24}$	$68.1_{2.6}$	$0.75_{0.04}$	$84.0_{2.0}$	$0.57_{0.04}$	$3.6_{1.5}$	$0.10_{0.02}$	$8.3_{7.5}$	$-3.81_{1.04}$	$52.8_{12.3}$	$0.82_{0.12}$
Mist(M)	$20.2_{2.5}$	$-1.26_{0.12}$	$68.8_{2.4}$	$0.74_{0.04}$	$84.8_{2.0}$	$0.57_{0.05}$	$6.4_{1.9}$	$0.08_{0.02}$	$30.0_{11.7}$	$-1.42_{0.75}$	$54.8_{10.8}$	$0.50_{0.17}$
Small / Efficient Models												
4o-mini	$8.9_{1.8}$	$-3.86_{0.21}$	$84.2_{1.8}$	$0.89_{0.03}$	$79.8_{2.2}$	$0.48_{0.05}$	$10.6_{2.1}$	$0.10_{0.02}$	$11.7_{7.5}$	$-5.01_{0.95}$	$81.4_{9.3}$	$0.78_{0.13}$
Mist(S)	$10.9_{2.0}$	$-1.63_{0.07}$	$82.6_{1.9}$	$0.89_{0.03}$	$80.0_{2.2}$	$0.47_{0.05}$	$11.4_{2.1}$	$0.11_{0.02}$	$5.0_{5.8}$	$-1.68_{0.56}$	$62.5_{11.4}$	$0.88_{0.11}$
Phi-4	$8.5_{1.7}$	$-3.63_{0.22}$	$84.1_{1.8}$	$0.91_{0.03}$	$80.8_{2.2}$	$0.39_{0.07}$	$14.2_{2.1}$	$0.11_{0.02}$	$15.0_{9.2}$	$-6.52_{0.87}$	$84.7_{9.1}$	$0.80_{0.12}$

III: BAS for Evaluating LLM Confidence - Scaling Trends

Larger models tend to achieve higher accuracy *and* improved confidence reliability (higher BAS, lower ECE and AURC), although substantial variation remains.

Models with higher accuracy tend to also exhibit more reliable confidence estimates.



III: BAS for Evaluating LLM Confidence - Confidence Elicitation

We evaluate four different methods to elicit confidence values $s \in [0, 1)$ from LLMs:

- *Direct Elicitation*: Produce both the answer and confidence estimate in a single response.
- *Self-Reflection*: Separate generation and confidence estimation into two distinct steps.
- *Top-k*: Produce k candidate answers with associated probabilities in a single response.
- *Top-k + Self-Reflection*: Combine both techniques.

Top-k consistently yields the highest BAS on SimpleQA across representative models.

Model	Direct Elicitation	Self-Reflection	Top-k	Top-k + Self-Reflection
GPT-4o-mini	-3.86 ± 0.21	-0.94 ± 0.09	$\mathbf{-0.42 \pm 0.02}$	-0.91 ± 0.14
GPT-oss-120B	-0.66 ± 0.04	-0.61 ± 0.05	$\mathbf{-0.22 \pm 0.02}$	-0.31 ± 0.03
Grok-3-Mini	-0.86 ± 0.07	-0.92 ± 0.08	$\mathbf{-0.50 \pm 0.05}$	-0.58 ± 0.07
Llama-3.3-70B	-2.97 ± 0.23	-2.05 ± 0.21	$\mathbf{-0.25 \pm 0.06}$	-0.54 ± 0.11
Phi-4	-3.63 ± 0.21	-3.60 ± 0.22	$\mathbf{-0.61 \pm 0.08}$	-2.23 ± 0.20

III: BAS for Evaluating LLM Confidence - Main Takeaways

(1/5) Even frontier models may fail to provide reliable confidence in challenging settings.

(2/5) Confidence reliability is task-dependent: models can accurately estimate their confidence on some tasks while exhibiting severe overconfidence on others.

(3/5) Confidence reliability generally improves both with model scale and performance: larger and more accurate models tend to also provide more reliable confidence estimates.

(4/5) Models with similar calibration (ECE) or ranking (AURC) can exhibit highly different decision-level reliability under BAS, driven by rare but highly overconfident errors.

(5/5) Confidence elicitation strongly affects reliability: simple strategies such as top-k can substantially reduce overconfidence and improve calibration in certain settings.

Questions?

Fredrik K. Gustafsson
University of Oxford
fredrik.gustafsson@eng.ox.ac.uk
www.fregu856.com