Reliable Fusion of Conflicting Experts

Pranuthi Tenali, Sahil Sidheekh, Saurabh Mathur, Vijayalakshmi Saravanan, Erik Blasch, Kristian Kersting, Sriraam Natarajan

Presented at the 6th International Conference on Dynamic Data Driven Applications Systems (DDDAS 2026)

StARLinG LAB

TECHNISCHE UNIVERSITÄT DARMSTADT

AFRL THE AIR FORCE RESEARCH LABORATORY LEAD | DISCOVER | DEVELOP | DELIVER

UTD THE UNIVERSITY OF TEXAS AT DALLAS

UTTyler THE UNIVERSITY OF TEXAS AT TYLER

MOTIVATION

A question between two specialties

THE PATIENT Myasthenia gravis (a neuromuscular disease) and a severe infection.

THE QUESTION **Which antibiotic plan should start now?**

(A) A fluoroquinolone | (B) A different broad-spectrum antibiotic | (C) Wait for cultures

Critical-care physician

BELIEF

A	**0.60**
B	0.40
C	0.00

Treat now; fluoroquinolones are standard.

Right about the urgency

conflict

Neurologist

BELIEF

A	0.05
B	0.45
C	**0.50**

Avoid: FDA boxed warning, can worsen myasthenia gravis.

Right about this patient

FUSED ANSWER

(B): treat now, with a safer antibiotic.

Routing to one expert gets it wrong; the answer needs both.

MOTIVATION

Whom to trust depends on the patient

SAME QUESTION + SAME EXPERTS • DIFFERENT PATIENT PROFILE

PATIENT A

Myasthenia gravis, severe infection

Critical-care physician	medium
Neurologist	**high**

PATIENT B

No neuromuscular disease, severe infection

Critical-care physician	**high**
Neurologist	low

WHY FUSION MUST BE CONTEXT-AWARE

Reliability of an expert is an interaction between their specialization and the context, not a fixed, global property of the expert

METHOD

Problem setup

GIVEN

N labeled questions with their answers

$$\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$$

K experts' beliefs, each over C answers

$$\mathbf{p} = (p_1, \ldots, p_K), \quad p_k \in [0,1]^C$$

Experts disagree; reliability varies by question

$$\arg\max p_i \neq \arg\max p_j$$

TO DO

Learn an aggregation model M that predicts answers based on the context z

$$M: (x, \mathbf{p}) \mapsto P(Y \mid \mathbf{p}, z), \quad z = f(x)$$

- ✓ No access to expert internals
- ✓ No retraining of the experts
- ✓ Agnostic to which experts are used

BACKGROUND

Two routes to context-aware fusion

SMOOTH

Input-dependent weighted mean

Use situational weights based on the current criteria. Weights become a function of the input, w(x): trust shifts smoothly with the input.

PIECEWISE

Expertise trees

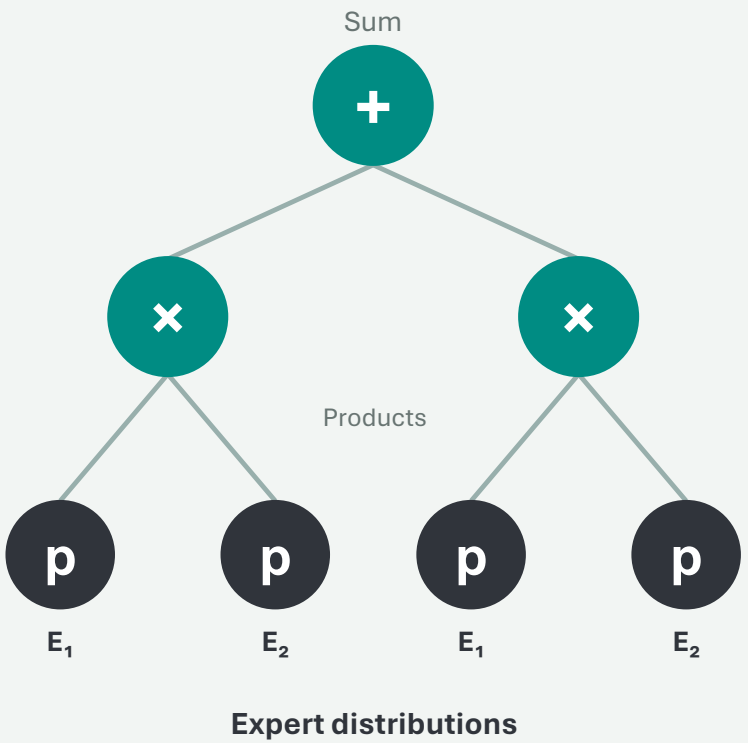
Hard splits on the context; each region gets its own fusion rule.

C²MF UNIFIES BOTH

A conditional probabilistic circuit with context-dependent weights covers both.

Abels et al. 2023, Expertise Trees Resolve Knowledge Limitations in Collective Decision-Making
Tenali et al. 2026, Context-specific Credibility-aware Multimodal Fusion with Conditional Probabilistic Circuits

Context-aware fusion with conditional PCs



Expert distributions

Model dependencies
A PC models the dependencies between the experts' answers.

Infer credibility
From those dependencies, it infers how credible each expert is.

Condition on the context
Conditioning the PC on the context makes each expert's reliability adapt based on the current situation.

Learn from data
PCs are differentiable computational graphs, like neural networks, so their parameters are learned by backpropagation.

OUR SETTING

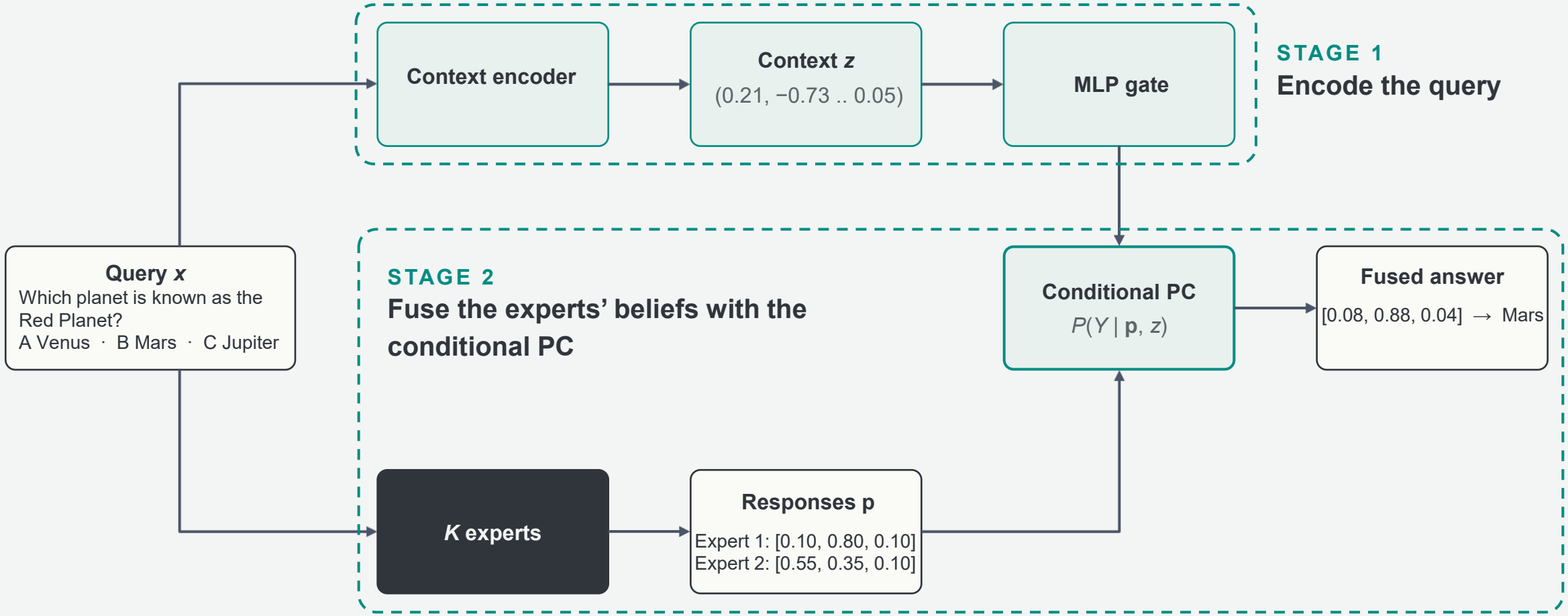
From multimodal fusion to expert fusion

	ORIGINAL SETTING **Multimodal fusion (C²MF)**	THIS WORK **Fusing expert decisions**
SOURCES	Different sensors, different information (e.g. audio, image).	Beliefs about the same question.
ACCESS	Open: internals available, models updatable.	Black boxes: outputs only, not updatable.
CONTEXT	Joint embedding of the unimodal embeddings.	Extracted from the question itself.

Our adaptation: context from the question's embedding; fuse only the experts' beliefs.

METHOD

C²MF for heterogeneous experts



METHOD

Two context-aware fusion functions

DIRECT INFERENCE

C²DPC

Answer by exact conditional inference in the CPC.

$$P(Y \mid \mathbf{p}, z) = \frac{P(Y, p_1, \ldots, p_K \mid z)}{P(p_1, \ldots, p_K \mid z)}$$

WEIGHTED MEAN

C²WM

Average the experts' responses, weighted by context-specific credibility.

$$\hat{Y} = \sum_{k=1}^{K} c_k(\mathbf{p}, z)\, p_k$$

with credibility weights:

$$c_k(\mathbf{p}, z) = D_{\mathrm{KL}}(P(Y \mid \mathbf{p}, z) \;\|\; P(Y \mid \mathbf{p} \setminus p_k, z))$$

EXPERIMENTS

Two research questions

Q1 Does context-aware fusion achieve better predictive performance?

Q2 Can we trust the fused answer's confidence?

EXPERIMENTS

MMLU with three prompted LLM experts

Data

- Massive Multitask Language Understanding (MMLU)
- 10 of its 57 tasks
- Grouped into medicine, physics and computer science.

Experts

- One expert persona per domain, merged from the top 3 expert roles an LLM proposes.

Querying

- Gemini-2.5-flash-lite
- Probability score per answer, averaged over 2 queries.

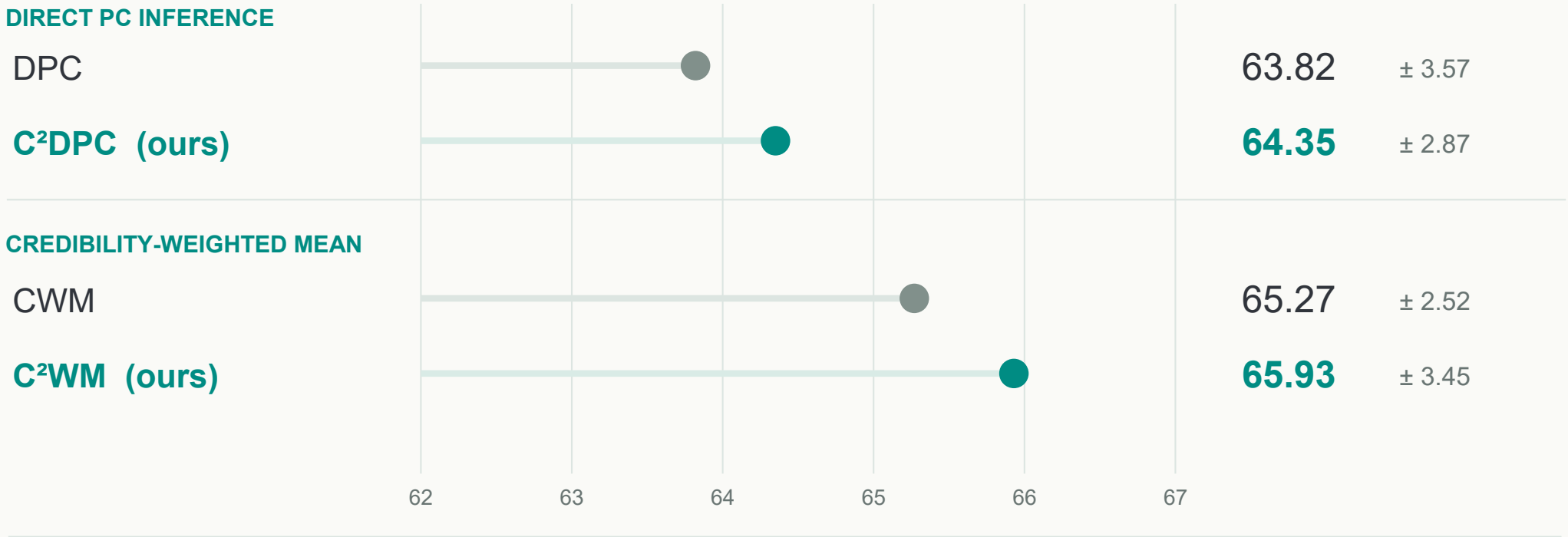
EXPERIMENTS

Methods compared

Group	Method	What the comparison tests
Individual experts	Medical · CS · Physics	Does fusing the experts beat each expert on its own?
Routing	SOE	Perfect routing to each question's domain expert
	RE	A random expert per question, the floor to beat.
Context-agnostic	WM	Static and linear: interpretable weights, but one fixed trust per expert.
	Noisy-OR	Static and non-linear: for experts whose evidence interacts non-linearly.
	DPC	C²DPC without context, isolating the effect of context.
	CWM	C²WM without context, to isolate context's effect.
Context-aware (current work)	C²DPC	A context-conditioned circuit that answers by direct inference.
	C²WM	Experts weighted by their context-specific credibility.

Q1 · DOES CONTEXT-AWARE FUSION ACHIEVE BETTER PREDICTIVE PERFORMANCE?

F1: modest gains with context

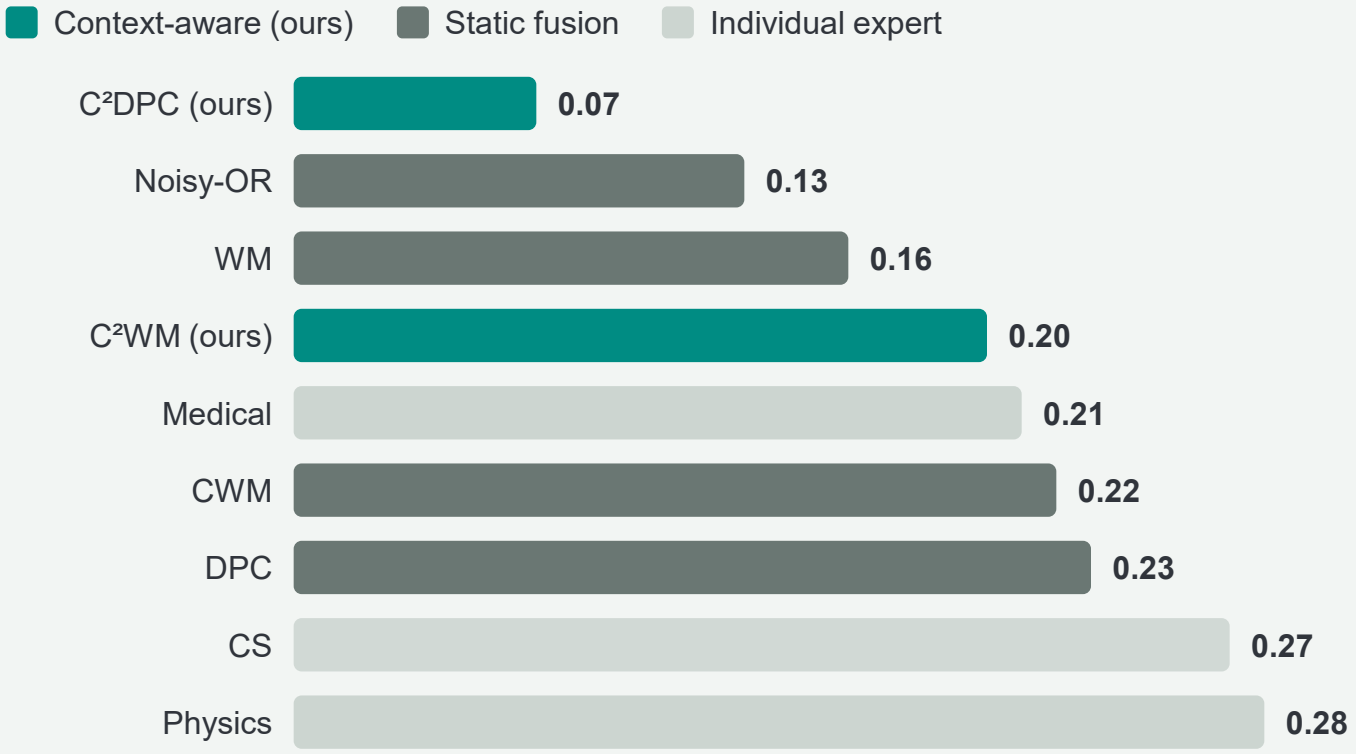


Both context-aware methods improve on their static counterparts: +0.53 and +0.66 points.

Mean ± std across 3 trials

Q2 · CAN WE TRUST THE FUSED ANSWER'S CONFIDENCE?

Expected calibration error on MMLU



ECE on the test set, mean over 3 trials

WHY IT MATTERS

Decisions follow confidence: an overconfident wrong answer does the most harm.

$$\mathrm{ECE} = \sum_{i=1}^{M} \frac{|B_i|}{N} \left|\mathrm{acc}(B_i) - \mathrm{conf}(B_i)\right|$$

Average gap between confidence and accuracy across M bins. Lower is better.

C²DPC is best calibrated: 0.07 vs 0.23 for DPC.

C²WM also beats CWM: 0.20 vs 0.22.

WRAP-UP

Conclusions and future work

WHAT WE SHOWED

- Probabilistic fusion of any experts: no internals, no retraining
- Context from the question itself sets how much to trust each expert
- On MMLU: beats every individual expert; improves on static counterparts in performance and calibration

FUTURE WORK

- Fusing human and AI experts together
- Beyond multiple choice: causal question answering and counterfactual reasoning
- Training experts from the consensus

Thank you

Questions?

PAPER

Reliable Fusion of Conflicting Experts · arxiv.org/abs/2609.27913

Supported by AFOSR award FA9550-23-1-0239, DOE FAIR award DE-SC0026194, and the DFG Cluster of Excellence “Reasonable AI” (EXC-3057).