ANCHOR: Reliable Probability Inference in Large Language Models

Abductive Network Construction with Hierarchical Orchestration

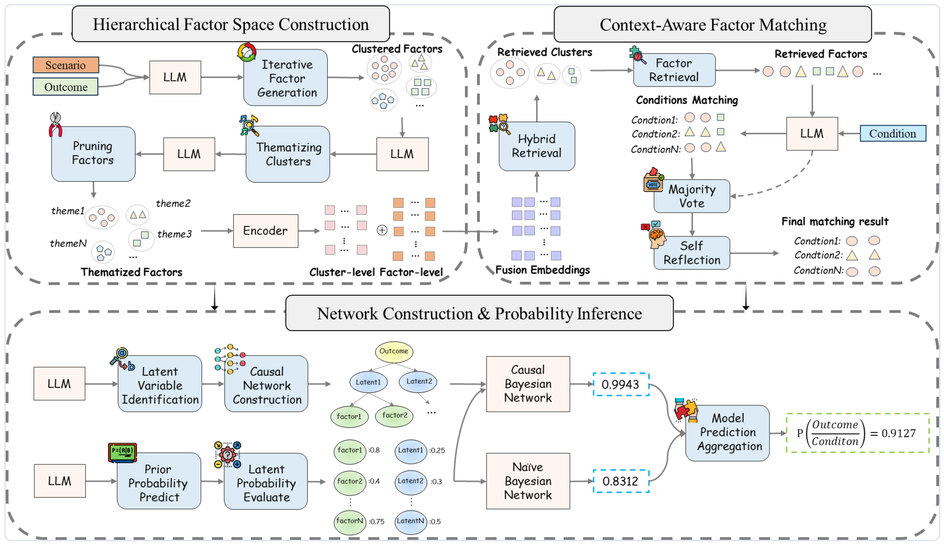
Wentao Qiu* · Guanran Luo* · Zhongquan Jian · Jingqi Gao · Meihong Wang ✉ · Qingqiang Wu ✉

School of Informatics · Institute of Artificial Intelligence · School of Film, **Xiamen University** | School of Computer & Data Science, **Minjiang University** | *Equal contribution ✉ Corresponding authors

ICML International Conference On Machine Learning

Paper & code

THE ANCHOR FRAMEWORK



(1) Factor-Space Construction builds a dense two-level factor hierarchy · **(2) Context-Aware Mapping** selects factors per condition via coarse-to-fine retrieval · **(3) Inference Orchestration** fuses Naïve Bayes + a Causal Bayesian Network into one calibrated probability.

LLMs give **overconfident, opaque** confidence. Prior abductive-Bayesian methods either leave a **sparse** factor space (many "unknown" answers) or, when enlarged, inject **noise & spurious dependencies** that break Naïve-Bayes independence. **ANCHOR** delivers **well-calibrated probabilities** — denser coverage *and* dependency-aware inference, at a fraction of the cost.

60.6%	**≈100%**	**6/6**	**0.24×**
best F1 (BIRD 52.6%)	coverage (BIRD 87.8%)	datasets won at 72B	tokens vs. BIRD

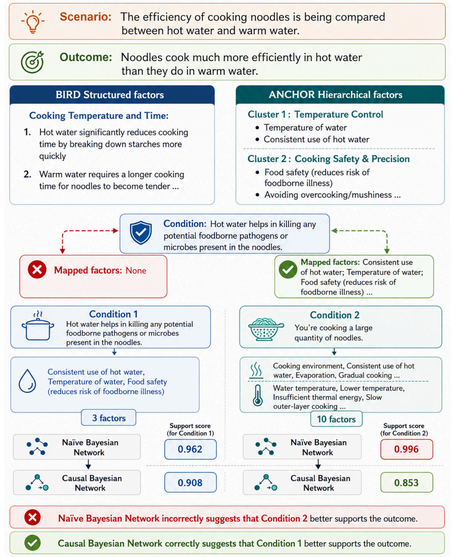
CONTRIBUTIONS

- A **multi-stage abduction pipeline** that iteratively grows a high-quality factor space, sharply cutting "unknown" predictions.
- A **two-level factor hierarchy** with coarse-to-fine, context-aware mapping — recall-first retrieval, then self-consistent filtering & reflective refinement.
- A **Causal Bayesian Network** with LLM-identified latent variables, modelling factor dependencies beyond Naïve Bayes and fused via a Linear Opinion Pool.
- **SOTA** preference alignment & decision accuracy while **cutting time and token cost**.

1 Problem & Motivation

High-stakes decisions need **calibrated** probabilities $P(O \mid C)$, but LLMs are overconfident and opaque. Existing abductive-Bayesian pipelines face a **two-horned dilemma**:

- A **sparse** factor space maps the condition to nothing → many "unknown" answers.
- **Naïvely enlarging** it injects noise & spurious dependencies that break Naïve-Bayes independence.



Worked case: BIRD maps the condition to **nothing** ("unknown"); ANCHOR's hierarchy maps **10 factors**, and its CBN — unlike Naïve Bayes — ranks the two contexts correctly.

★ Experimental Setup

- **Pairwise preference:** Common2Sense (3,822 instances).
- **Reasoning / planning:** Today, Plasma.
- **Fact-checking:** ExpertQA, XSum, CNN, COVID.
- **LLMs:** Qwen2.5-32B / 72B, DeepSeek-V3-671B, GPT-4 · K_1=3, K_2=5 · 4× RTX 4090.

2 The ANCHOR Method

Stage 1 Iterative abduction → factor space

Bottom-up abduction decouples generation from structuring: iteratively elicit diverse supporting/refuting sentences, harvest & validate factors until size N_{target}. Then build a **two-level hierarchy** — MiniLM → UMAP → HDBSCAN clusters → LLM theming (e.g. *"Economic Feasibility"*); each factor is labelled *supports* O_1/O_2/*neutral*.

$$\text{err} \le \exp(-2m(q-\tfrac{1}{2})^2)$$

self-consistency majority vote of m queries, single-vote accuracy q > ½

Stage 2 Context-aware mapping

Map condition u to a compact factor set $\mathcal{F}^*(u)$ via **coarse-to-fine retrieval** over the hierarchy (recall-first), then **self-consistent filtering** and **reflective refinement** for precision.

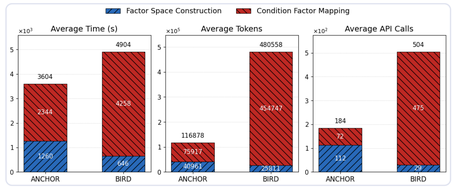
Stage 3 Inference orchestration

Build a **Naïve Bayes** model and a **Causal Bayesian Network** from the mapped factors. LLM-elicited priors initialise both; the CBN's **latent variables** capture shared causes that NB would double-count, then their posteriors fuse via a **Linear Opinion Pool**.

$$P(O_i \mid C) = \Sigma_{f \in \mathcal{V}} P(O_i \mid f) \cdot \Pi_j P(f_j \mid C)$$

deductive marginalization; the CBN relaxes the ∏ independence assumption

4 Efficiency



0.74×	**0.24×**	**0.37×**
runtime	tokens	API calls

Per scenario vs. BIRD. Up-front factor-space construction is repaid by far cheaper hierarchical mapping downstream.

3 Results

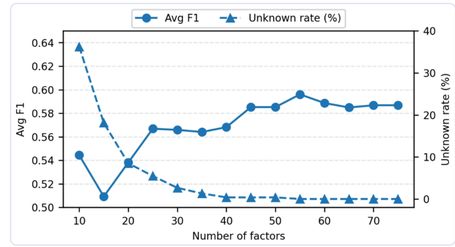
Preference F1 on Common2Sense (micro-avg)

Method	Model	Avg F1	Cov.
Vanilla	Qwen-72B	.470	–
Compare★	GPT-4	.556	–
BIRD‡	Qwen-72B	.526	99.99%
ANCHOR	Qwen-32B	.563	99.79%
ANCHOR	**Qwen-72B**	**.606**	**99.95%**

Decision accuracy (6 datasets, Qwen-72B)

Dataset	BIRD	ANCHOR
ExpertQA	**.582**	**.611**
XSum	**.511**	**.564**
CNN	**.540**	**.621**
Today	**.754**	**.817**
COVID	.664	.726
Plasma	.718	.764

5 Coverage ↔ Accuracy



As the factor space grows, the **unknown rate falls 36%** → **≈0%** while average F1 keeps rising — denser hierarchies buy coverage *and* accuracy.

6 Ablation

- **−.137** w/o LLM-elicited params — elicitation matters most.
- **−.115** w/o clustering — the two-level hierarchy is essential.
- **−.093** w/o Naïve Bayes · **−.014** w/o CBN — both branches help.
- LOP ≈ BMA pooling (Δ ≤ .004); coverage stays **≈100%**.

Avg-F1 change vs. the full model (= .572), Common2Sense / Qwen-32B.

Idea — A **dense, hierarchical** factor space plus **causal** modeling beats sparse abduction — coverage and calibration together.

Method — Bottom-up abduction → coarse-to-fine mapping → **Naïve Bayes** ⊕ **Causal Bayesian Network** fused by a Linear Opinion Pool.

Result — **60.6% F1**, **≈100% coverage**, and wins on **all 6** decision datasets at its own scale — also beating a 671B baseline on the hard ones.

Practical — Reliable probabilities at **0.24× tokens**, 0.74× time and 0.37× API calls of BIRD — deployable at scale.

Code: **github.com/ChowRunFa/ANCHOR** · Contact: wtqiu@stu.xmu.edu.cn · wangmh@xmu.edu.cn · wuqq@xmu.edu.cn