FUTUREMAP: OMNIMODAL FUTURE PREDICTION WITH TEMPORAL HYPERGRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Future prediction asks what will happen, and when, from only the evidence visible at a vantage point, and large language model (LLM) agents now make it feasible over archives whose precursors span five modalities. Forecasters have progressed from agentic search to temporal knowledge graphs and hypergraphs, and recently to omnimodal inputs. However, they leave omnimodal evidence unstructured, guess the regularities linking precursors to outcomes, and date an event on whatever deadline they are shown. We formalise Omnimodal Future Prediction (OmniFP) and release OMNIFP-BENCH, 1,530 questions over eight domains and 7,422 frozen resources in five modalities. We propose FUTUREMAP, a training-free forecaster that builds an omnimodal temporal hypergraph, backtests hyperrules with time and magnitude kernels, and reads one deadline-blind deduction out in code. Against twelve baselines, FUTUREMAP leads on all three backbones by $+4.1$ to $+5.3$ points. Our software and data are publicly available.[1]

1 INTRODUCTION

Future prediction asks a system to judge an event that has not yet happened — whether, when and how much — from only the evidence visible at a vantage point (Xu et al., 2026b). LLM agents now make this feasible end to end, from searching an archive to committing to an answer (Halawi et al., 2024). Yet precursors are rarely text alone: a news report, a photograph, a filing, a TV clip and a podcast often point to the same event (Figure 1), so a forecaster has to read an *omnimodal* archive.

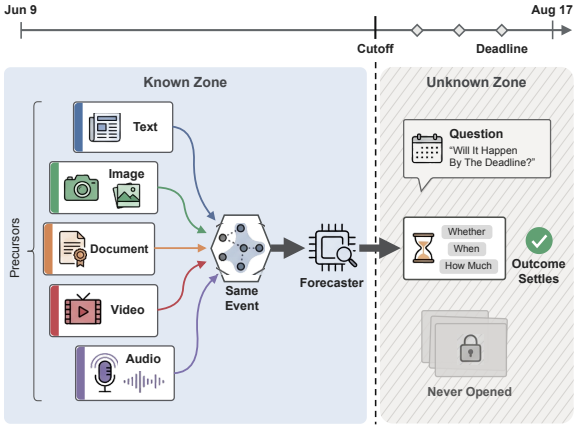

Figure 1: **The OmniFP task:** forecast from omnimodal precursors what settles after the cutoff.

Two lines of work pursue this goal. *Agentic forecasters* evolved from multi-turn search in REACT (Yao et al., 2023) and MIROFLOW (Su et al., 2026) to experience reuse in EXPEL (Zhao et al., 2024) and calibrated belief updating in LEAP (Chen et al., 2026d). *Structured forecasters* organise evidence as temporal knowledge graphs (Lee et al., 2023; Tan et al., 2026) and hypergraphs (Luo et al., 2025; Chen et al., 2026b), and FUTUREOMNI (Chen et al., 2026a) extends forecasting to omnimodal inputs.

However, three challenges remain (Figure 2). **(C1) Omnimodal evidence.** One event is often reported in several media, yet evidence stores are text- or image-only, or flat as in FUTUREOMNI, so cross-media corroboration is lost. **(C2) Unmeasured regularities.** Retrieval returns precursors but not how likely, how soon and how large an outcome follows; agents leave this to the model's guess, and rule learners (Xiong et al., 2023) need curated graphs. **(C3) Deadline-anchored timing.** A model shown a deadline dates the event on it, so its yes/no answers reflect a fixed bias rather than a forecast.

[1] Anonymous GitHub link: `https://anonymous.4open.science/r/FutureMap`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

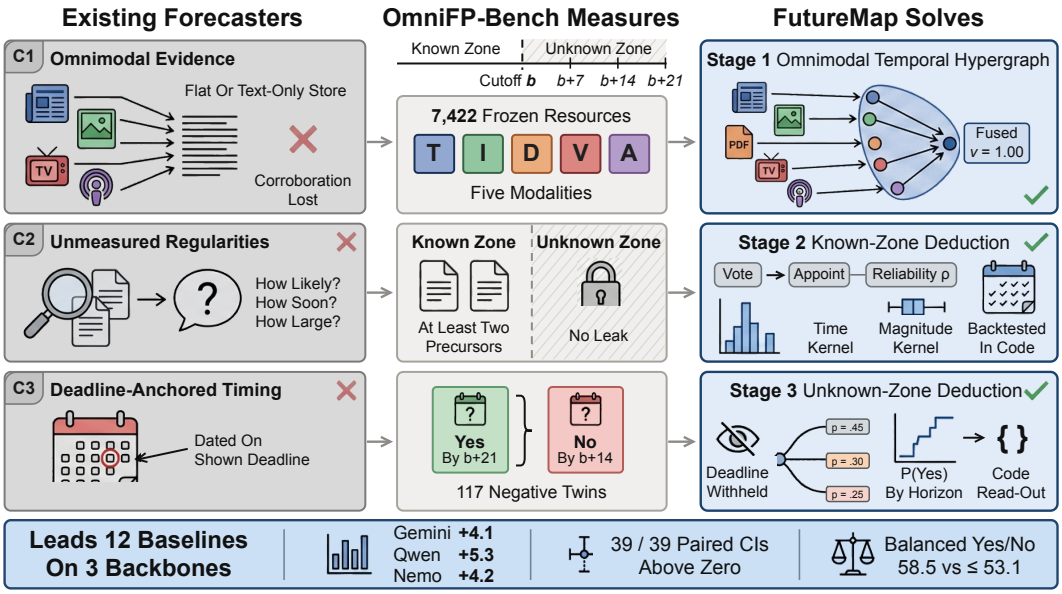

Figure 2: **Overview.** Existing forecasters leave C1–C3 open; OMNIFP-BENCH measures them, and the three stages of FUTUREMAP address them, leading every baseline on all three backbones.

As no benchmark measures these with frozen omnimodal evidence, we formalise **Omnimodal Future Prediction (OmniFP)** and release **OMNIFP-BENCH**. A system reads a frozen, five-modality known zone up to a cutoff b, and the answer settles after b in resources it never opens (Figure 1); every item has at least two precursors and no leak. The benchmark holds 1,530 items of seven answer types over eight domains and 7,422 real resources, including 117 *negative twins* that re-ask a *Yes* question at an earlier deadline to test *when* (§3).

We propose **FUTUREMAP** (§4), a training-free forecaster whose three stages answer C1–C3. **(i) Omnimodal Temporal Hypergraph** reads each resource in its own medium and fuses cross-media reports of one event into a hyperedge weighted by temporal validity. **(ii) Known-Zone Deduction** backtests typed hyper-rules inside the archive, measuring rule reliability, time kernels and magnitude kernels in code. **(iii) Unknown-Zone Deduction** asks the backbone once for dated branches that must fire these rules, withholding the deadline until a code read-out.

Against twelve training-free baselines on three frozen backbones, FUTUREMAP is the most accurate on every backbone, by $+4.1$ to $+5.3$ points, with all 39 paired confidence intervals above zero. It alone tells *when*: it gets both questions of a twin pair right about twice as often as the best baseline, and its balanced yes/no accuracy (58.5–59.9) exceeds every baseline's (≤ 53.1) (§5).

2 RELATED WORK

We review the two closest lines of work; Appendix G gives an extended discussion.

LLM-based Future Prediction. LLM forecasting began as agentic search in REACT (Yao et al., 2023), MIROFLOW (Su et al., 2026) and TIMESEEK (Mostafa et al., 2026), then reused experience in EXPEL (Zhao et al., 2024) and updated beliefs in mini-prophet (Murphy, 2026) and LEAP (Chen et al., 2026d), while benchmarks widened from text (Ye et al., 2024; Karger et al., 2025; Zeng et al., 2026) to omnimodal inputs (Chen et al., 2026a). Yet all leave the evidence–outcome link implicit.

Graph-based Retrieval and Reasoning. Structured retrieval makes that link explicit: TKG-ICL (Lee et al., 2023), MEMOTIME (Tan et al., 2026) and ATOM (Lairgi et al., 2026a) keep temporal knowledge graphs, from which TILP (Xiong et al., 2023) mines rules, and HYPER-GRAPHRAG (Luo et al., 2025) and MMHRAG (Chen et al., 2026b) build hyperedges over text and images. None is omnimodal or backtests its rules; FUTUREMAP does both.

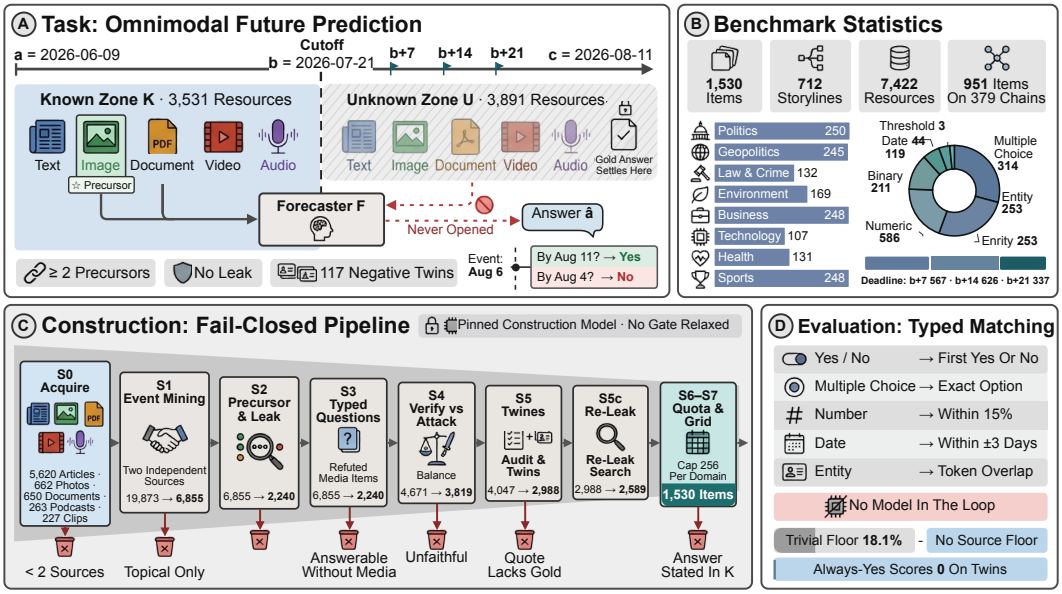

Figure 3: **OmniFP and OMNIFP-BENCH.** (A) Task: forecast from the known zone $\mathcal{K}$, read in five modalities, an answer that settles in the unknown zone $\mathcal{U}$. (B) Statistics: 1,530 items over eight domains and seven answer types, with 7,422 resources. (C) Construction: a fail-closed S0–S7 pipeline. (D) Evaluation: typed matching, accuracy and floors.

3 THE OMNIFP TASK AND BENCHMARK

Figure 3 summarises the task and the benchmark; Appendix D gives full details.

Task Definition. An instance $(q, \tau, \mathcal{R})$ pairs a question q with a protocol $\tau = (a, b, c)$ of three dates and a frozen pool $\mathcal{R}$ of resources, each with a modality $m(r) \in \{\mathrm{T}, \mathrm{I}, \mathrm{D}, \mathrm{V}, \mathrm{A}\}$ (text, image, document, video, audio) and a metadata date t_r. The dates split the pool,

$$\mathcal{K} = \{r \in \mathcal{R} : a \le t_r \le b\}, \qquad \mathcal{U} = \{r \in \mathcal{R} : b < t_r \le c\}, \qquad F : (q, \mathcal{K}) \longmapsto \hat{a}, \tag{1}$$

and a system F reads the known zone $\mathcal{K}$ in all five modalities, while the gold a^* settles in the unknown zone $\mathcal{U}$, in resources it never opens. An item has at least two precursors in $\mathcal{K}$ and no resource there states its outcome (*no leak*), so it is neither a guess nor a lookup; every model a system runs, its encoder included (t_{embed}), is released by b, and nothing is fetched at run time.

OmniFP-Bench. With $\tau = (2026\text{-}06\text{-}09, 2026\text{-}07\text{-}21, 2026\text{-}08\text{-}11)$, deadlines lie on the grid $\{b{+}7, b{+}14, b{+}21\}$. The release holds 1,530 items over 712 storylines in eight domains (951 on 379 cross-domain chains) and seven answer types, from numeric value (586) to numeric threshold (3). 117 *negative twins* re-ask a *Yes* binary at the last grid deadline before its event, so only placing the event in time answers both. Of the 7,422 resources, 3,531 lie in $\mathcal{K}$ and 3,891 in $\mathcal{U}$; items a closed-book probe answers are flagged, and per-domain counts are in Table 7.

Benchmark Construction. One pinned construction model, `gpt-4o-mini`, of no evaluation backbone's family, drives a fail-closed pipeline (Figure 3C): **S0** acquires articles, photographs, documents, podcasts and clips; **S1** keeps events stated by two independent sources; **S2** enforces the precursor and no-leak conditions in all five modalities; **S3** writes typed questions with a quoted gold, labelling an item non-text only if text cannot answer it; **S4** judges quote-settled items under opposed VERIFY and ATTACK; **S5–S5c** audit dates, quotes and leaks and add the twins; **S6–S7** cap each domain at 256 items and grid the deadlines. Table 6 traces 19,873 events to 1,530 items.

Evaluation. Answers are graded by typed matching with no model in the loop (Figure 3D, Eq. 25): yes/no on the first *yes* or *no*, multiple choice on the gold option, numbers within 15%, dates within ±3 days and named entities by token overlap, with tolerances fixed in advance. We report accuracy per domain and **Overall** against two floors: each backbone's *No source* row and a trivial guesser at 18.1% (*Yes* on yes/no, uniform on multiple choice) that scores 0 on the twins (Appendix D.3).

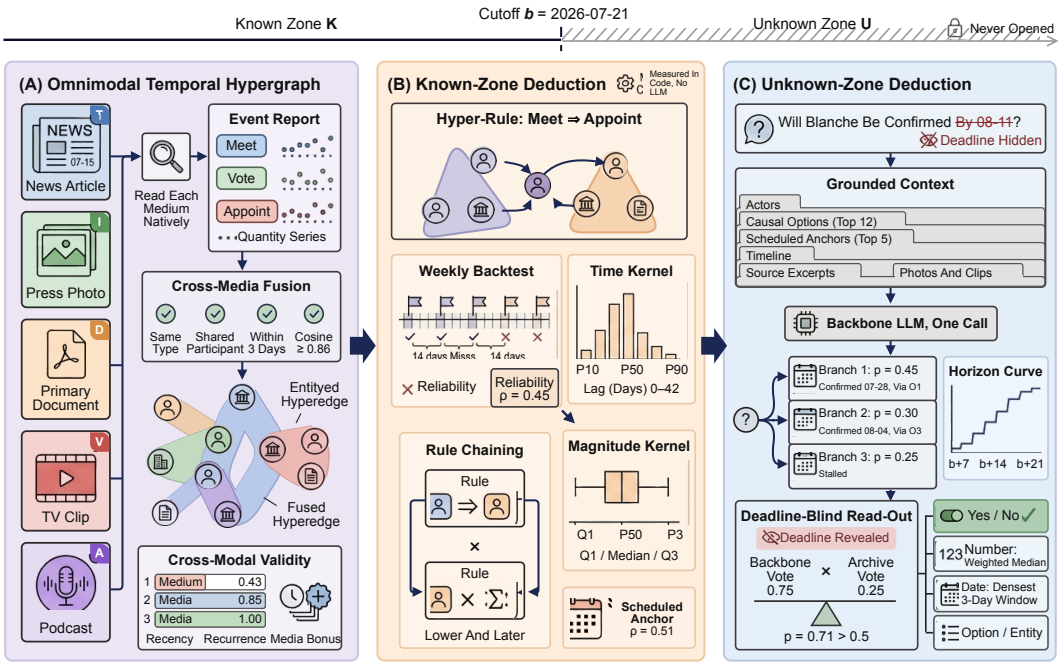

Figure 4: **FUTUREMAP.** (A) Omnimodal Temporal Hypergraph fuses media. (B) Known-Zone Deduction backtests rules. (C) Unknown-Zone Deduction reads out dated branches, deadline-blind.

4 METHODOLOGY: FUTUREMAP

FUTUREMAP (Figure 4) answers C1–C3 of §1 in three training-free stages: a hypergraph built once from $\mathcal{K}$, hyper-rules and kernels mined once, in code, and one grounded deduction query per item q.

4.1 OMNIMODAL TEMPORAL HYPERGRAPH

Reading each resource in its own medium. For C1, stage 1 builds a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, v)$ of entities, event hyperedges and validities. A builder model Read (Gemini 2.5 Flash-Lite) reads each $r \in \mathcal{K}$ natively, stills and clips included, into ≤ 8 event reports $\tilde{e}$ and ≤ 8 quantity points π:

$$\mathsf{Read}(r) = \big(\{\tilde{e} = (\theta_{\tilde{e}}, \mathcal{V}_{\tilde{e}}, r, t^{\mathrm{sch}}_{\tilde{e}})\},\ \{\pi = (u, k, y, t_\pi)\}\big), \qquad \theta_{\tilde{e}} \in \Theta,\ \ \mathcal{V}_{\tilde{e}} \subseteq \mathcal{V},\ \ t_\pi \leq t_r, \tag{2}$$

where a report has one of 30 types $\theta_{\tilde{e}} \in \Theta$, participants $\mathcal{V}_{\tilde{e}}$ and the date $t^{\mathrm{sch}}_{\tilde{e}}$ of any event it announces, and a point is value y of indicator k for entity u at t_π; the points sharing (u, k) form one series $Q_{u,k}$.

Omnimodal hyperedges. With encoder $\mathbf{z} = \mathrm{enc}(\cdot)$, reports $\tilde{e}$ of r and $\tilde{e}'$ of $r' \neq r$ are linked if they share type and participant, lie within $\delta = 3$ days and match in meaning at cosine $\gamma = 0.86$:

$$\tilde{e} \leftrightarrow \tilde{e}' \iff \theta_{\tilde{e}} = \theta_{\tilde{e}'} \ \wedge\ \mathcal{V}_{\tilde{e}} \cap \mathcal{V}_{\tilde{e}'} \neq \emptyset \ \wedge\ |t_r - t_{r'}| \leq \delta \ \wedge\ \cos(\mathbf{z}_{\tilde{e}}, \mathbf{z}_{\tilde{e}'}) \geq \gamma. \tag{3}$$

Each component of $\leftrightarrow$ is a hyperedge $e = (\theta_e, \mathcal{V}_e, \mathcal{S}_e, t_e, t^{\mathrm{sch}}_e)$ with participants $\mathcal{V}_e$, support $\mathcal{S}_e \subseteq \mathcal{K}$ in modalities $\Omega_e = \{m(r)\}_{r \in \mathcal{S}_e}$ and date $t_e = \min_{\mathcal{S}_e} t_r$; a press release and its clip yield $|\Omega_e| = 2$.

Cross-modal temporal validity. After INFER (Li et al., 2025), a fact's weight at b decays from its last attestation $\bar{t}_e = \max_{\mathcal{S}_e} t_r$ and grows with its $f_e = |\mathcal{S}_e|$ attestations and, we add, its media:

$$v_e = \min\Big\{1,\ \big[(b - \bar{t}_e + 1)^{-1/2} + \lambda(f_e)\, f_e\big]\big(1 + \beta\,(|\Omega_e| - 1)\big)\Big\}, \qquad \lambda(f) = \sigma\big(\sqrt{f}/\eta\big) - \tfrac{1}{2}, \tag{4}$$

with σ the logistic function, $\eta = 2$ and per-medium bonus $\beta = 0.25$; v_e weights every later use of e.

Proposition 1. *Fusion never lowers validity: a fused event is at least as valid as its reports.*

Proof. We provide experimental results in Sections 5.3–5.4 and proofs in Appendix B.1. □

4.2 KNOWN-ZONE DEDUCTION: BACKTESTED HYPER-RULES AND KERNELS

Hyper-rules and time kernels. For C2, stage 2 measures regularities in code. Let $e \bowtie e'$ if hyperedges share a participant in $\mathcal{V}^\circ$, $\mathcal{V}$ without hubs (>400 hyperedges). A hyper-rule $\phi = (\theta_\phi^{\mathrm{pre}} \Rightarrow \theta_\phi^{\mathrm{post}})$ says a $\theta_\phi^{\mathrm{pre}}$ precursor is followed on a shared participant by a $\theta_\phi^{\mathrm{post}}$ consequence; its time kernel κ_ϕ is the density of first-passage lags up to $L = 42$ days, smoothed with width $\zeta = 2$:

$$\ell_\phi(e) = \min\{\, t_{e'} - t_e :\ e' \bowtie e,\ \theta_{e'} = \theta_\phi^{\mathrm{post}},\ 1 \le t_{e'} - t_e \le L \,\}, \qquad K_\phi(h) = \textstyle\sum_{\ell=0}^{h} \kappa_\phi(\ell). \quad (5)$$

Reliability by backtesting. A rule is scored as a forecaster: at weekly pseudo-cutoffs $b' \in \mathcal{B}' = \{a{+}7, a{+}14, \dots\} \cap (a, b - H]$, each precursor $e \in \mathcal{F}_\phi(b')$ of type $\theta_\phi^{\mathrm{pre}}$ from the week up to b' predicts a consequence within $H = 14$ days; over its $n_\phi = \sum_{b'} |\mathcal{F}_\phi(b')|$ forecasts, with indicator $\mathbb{I}$,

$$\rho_\phi = \frac{n_\phi^+ + 1}{n_\phi + 2}, \qquad n_\phi^+ = \sum\nolimits_{b' \in \mathcal{B}'} \sum\nolimits_{e \in \mathcal{F}_\phi(b')} \mathbb{I}\big[\exists\, e' \bowtie e :\ \theta_{e'} = \theta_\phi^{\mathrm{post}},\ t_{e'} \in (b', b'{+}H]\big]. \quad (6)$$

Rules with $n_\phi \ge 4$ above base rate form Φ, and scheduled hyperedges (*anchors*) attain $\rho_{\mathcal{A}} = 0.51$.

Composing rules and magnitude kernels. Chaining rules with $\theta_{\phi_1}^{\mathrm{post}} = \theta_{\phi_2}^{\mathrm{pre}}$ multiplies reliabilities and adds lags, giving $\Phi^{\le 2}$; a numeric q takes the quartiles μ_q of its ten latest comparable points $\mathcal{X}_q$:

$$\rho_{\phi_1 \circ \phi_2} = \rho_{\phi_1} \rho_{\phi_2}, \quad \kappa_{\phi_1 \circ \phi_2}(\ell) = \textstyle\sum_{i=0}^{\ell} \kappa_{\phi_1}(i)\, \kappa_{\phi_2}(\ell - i), \quad \mu_q = \big(y_{(1/4)}, y_{(1/2)}, y_{(3/4)}\big). \quad (7)$$

Proposition 2. *Chaining rules can only lower and delay a forecast, never raise or advance it.*

Proof. We provide experimental results in Sections 5.3–5.5 and proofs in Appendix B.2. □

4.3 UNKNOWN-ZONE DEDUCTION: GROUNDED BRANCHES, DEADLINE-BLIND READ-OUT

Grounding the question. For C3, stage 3 grounds q in code. With deadline d_q and horizon $\Delta_q = d_q - b$, the target $\tilde{q}$ is q without "by d_q" if yes/no, else q; it fixes the nearest type θ_q, entities $\mathcal{V}_q \subseteq \mathcal{V}^\circ$ and relevance $s_q(e) = \max\{0, \cos(\mathbf{z}_e, \mathbf{z}_{\tilde{q}})\}$. *Options* pair recent precursors with rules into θ_q:

$$\mathcal{O}_q = \mathrm{top}_{12}\{(e, \phi) \in \mathcal{E} \times \Phi^{\le 2} :\ \mathcal{V}_e \cap \mathcal{V}_q \ne \emptyset,\ 0 \le b - t_e \le L,\ \theta_\phi^{\mathrm{pre}} = \theta_e,\ \theta_\phi^{\mathrm{post}} = \theta_q\}, \quad (8)$$

ranked by $\rho_\phi v_e s_q(e)$, and every item grounds all twelve; the anchors $\mathcal{A}_q$ are the top five scheduled hyperedges still due around the cutoff, $t_e^{\mathrm{sch}} \ge b - 2$, with $s_q(e) \ge 0.55$, ranked alike by $\rho_{\mathcal{A}} v_e s_q(e)$.

Branches that must fire rules. The backbone LM reads a context $\mathbf{x}_q$ (the latest hyperedges of $\mathcal{V}_q$, $\mathcal{O}_q$ with ρ_ϕ and lag percentiles, $\mathcal{A}_q$, a timeline $\mathcal{N}_q$, $\mathcal{X}_q$, μ_q for numbers, excerpts Ξ_q) and returns $J \le 4$ exclusive branches ψ_j of probability p_j, date t_j and answer y_j, each firing what via_j cites:

$$\mathsf{LM}(\mathbf{x}_q) = \big(\{\psi_j\}_{j=1}^{J},\, C_q\big), \qquad \psi_j = (p_j,\, t_j,\, y_j,\, \mathrm{via}_j), \qquad \mathrm{via}_j \subseteq \mathcal{O}_q \cup \mathcal{A}_q. \quad (9)$$

Deadline-blind timing. Models date the event on a shown deadline, so $\tilde{q}$ omits d_q and $C_q(h)$ is the event's probability by $b{+}h$, $h \in \{7, 14, 21\}$. With running maximum $\bar{C}_q$ ($\bar{C}_q(0) = 0$), the backbone decides *whether* it falls in the window and half of *when*; rules and anchors due by d_q, $\mathcal{A}_q^d$, vote:

$$p_q^{\mathrm{LM}} = \tfrac{1}{2}\bar{C}_q(\Delta_q) + \tfrac{\Delta_q}{42}\bar{C}_q(21), \quad p_q^{\mathrm{arc}} = 1 - \prod\nolimits_{(e,\phi) \in \mathcal{O}_q} (1 - \rho_\phi K_\phi(\Delta_q)) \prod\nolimits_{e \in \mathcal{A}_q^d} (1 - \rho_{\mathcal{A}} s_q(e)), \quad (10)$$

and the answer is *Yes* iff $\hat{p}_q = (1 - \alpha) p_q^{\mathrm{LM}} + \alpha p_q^{\mathrm{arc}} > \frac{1}{2}$, with weight $\alpha = \frac{1}{4}$ for the archive's vote.

Typed read-out. A number is the weighted median of branch answers and the last series value y_q^{last} ($\omega = 0.75$), a date the densest ± 3-day window, an option or entity the likeliest stated one:

$$\hat{a}_q^{\mathrm{num}} = \mathrm{wmed}\big(\{(y_j, p_j)\}_{j=1}^{J} \cup \{(y_q^{\mathrm{last}}, \omega)\}\big), \qquad \hat{a}_q^{\mathrm{date}} = \arg\max_t \textstyle\sum_{j=1}^{J} p_j\, \mathbb{I}\big[|t_j - t| \le 3\big]. \quad (11)$$

Proposition 3. *For a fixed deduction, the probability of* Yes *never falls as the deadline moves later.*

Proof. We provide experimental results in Sections 5.3–5.6 and proofs in Appendix B.3. □

Table 1: **Main results on OMNIFP-BENCH**: accuracy per domain and **Overall** (1,530 items; §5.1); every row reads all five modalities. Chips: modality T text, M image, O omnimodal; index chunks, binary edges, hyperedges. Subscripts: gap to the block's **No source** row, green above, red below. **Bold** / underline: best / runner-up retrieval row. Shaded: evidence served via a parser.

Model	Method	Pol	Geo	Law	Env	Biz	Tech	Hlth	Spt	Overall
Mistral Nemo	No source	20.8	17.6	18.2	20.1	19.4	20.6	25.2	22.2	20.3
T → T	ReAct	20.8$_{\pm 0.0}$	19.2$_{+1.6}$	18.2$_{\pm 0.0}$	15.4$_{-4.7}$	20.6$_{+1.2}$	27.1$_{+6.5}$	26.7$_{+1.5}$	27.0$_{+4.8}$	21.6$_{+1.3}$
$0.019 / $0.030 per 1M	TKG-ICL	17.2$_{-3.6}$	15.5$_{-2.1}$	15.9$_{-2.3}$	14.2$_{-5.9}$	17.7$_{-1.7}$	22.4$_{+1.8}$	22.1$_{-3.1}$	20.2$_{-2.0}$	17.8$_{-2.5}$
128K ctx · Jul 2024	ExpeL	18.4$_{-2.4}$	18.8$_{+1.2}$	16.7$_{-1.5}$	15.4$_{-4.7}$	23.8$_{+4.4}$	28.0$_{+7.4}$	28.2$_{+3.0}$	26.6$_{+4.4}$	21.7$_{+1.4}$
Text T → T native	HyperGraphRAG	23.6$_{+2.8}$	20.8$_{+3.2}$	20.5$_{+2.3}$	21.9$_{+1.8}$	23.0$_{+3.6}$	26.2$_{+5.6}$	25.2$_{\pm 0.0}$	**33.5**$_{+11.3}$	24.5$_{+4.2}$
Image → T OCR	MemoTime	13.2$_{-7.6}$	21.2$_{+3.6}$	18.2$_{\pm 0.0}$	13.0$_{-7.1}$	15.7$_{-3.7}$	17.8$_{-2.8}$	22.9$_{-2.3}$	19.4$_{-2.8}$	17.5$_{-2.8}$
Doc → T text layer	ATOM	20.4$_{-0.4}$	19.6$_{+2.0}$	18.9$_{+0.7}$	16.6$_{-3.5}$	20.2$_{+0.8}$	23.4$_{+2.8}$	29.8$_{+4.6}$	26.2$_{+4.0}$	21.6$_{+1.3}$
Video → T ASR+OCR	FutureOmni	25.6$_{+4.8}$	20.4$_{+2.8}$	18.9$_{+0.7}$	16.6$_{-3.5}$	20.6$_{+1.2}$	22.4$_{+1.8}$	**32.1**$_{+6.9}$	30.6$_{+8.4}$	23.5$_{+3.2}$
Audio → T ASR	MiroFlow	17.2$_{-3.6}$	15.9$_{-1.7}$	15.2$_{-3.0}$	13.6$_{-6.5}$	19.0$_{-0.4}$	17.8$_{-2.8}$	22.1$_{-3.1}$	20.6$_{-1.6}$	17.7$_{-2.6}$
	TimeSeek	21.6$_{+0.8}$	19.2$_{+1.6}$	19.7$_{+1.5}$	16.6$_{-3.5}$	20.6$_{+1.2}$	21.5$_{+0.9}$	21.4$_{-3.8}$	26.6$_{+4.4}$	21.1$_{+0.8}$
	mini-prophet	22.0$_{+1.2}$	22.0$_{+4.4}$	17.4$_{-0.8}$	22.5$_{+2.4}$	21.0$_{+1.6}$	18.7$_{-1.9}$	31.3$_{+6.1}$	28.2$_{+6.0}$	23.1$_{+2.8}$
	MMHRAG	22.4$_{+1.6}$	22.0$_{+4.4}$	22.0$_{+3.8}$	20.1$_{\pm 0.0}$	23.8$_{+4.4}$	19.6$_{-1.0}$	29.8$_{+4.6}$	27.8$_{+5.6}$	23.6$_{+3.3}$
	LEAP	23.6$_{+2.8}$	25.3$_{+7.7}$	22.0$_{+3.8}$	21.3$_{+1.2}$	23.0$_{+3.6}$	25.2$_{+4.6}$	30.5$_{+5.3}$	32.3$_{+10.1}$	25.5$_{+5.2}$
	FutureMap	**29.6**$_{+8.8}$	**32.2**$_{+14.6}$	**28.0**$_{+9.8}$	**24.9**$_{+4.8}$	**26.2**$_{+6.8}$	**33.6**$_{+13.0}$	29.8$_{+4.6}$	**33.5**$_{+11.3}$	**29.7**$_{+9.4}$
Qwen3.5 Flash	No source	23.2	22.4	18.9	20.1	22.2	28.0	26.7	22.6	22.7
T → T	ReAct	20.4$_{-2.8}$	23.7$_{+1.3}$	25.0$_{+6.1}$	18.9$_{-1.2}$	25.8$_{+3.6}$	**38.3**$_{+10.3}$	35.9$_{+9.2}$	29.4$_{+6.8}$	26.1$_{+3.4}$
$0.065 / $0.260 per 1M	TKG-ICL	22.4$_{-0.8}$	24.5$_{+2.1}$	20.5$_{+1.6}$	20.7$_{+0.6}$	25.0$_{+2.8}$	26.2$_{-1.8}$	30.5$_{+3.8}$	22.6$_{\pm 0.0}$	23.8$_{+1.1}$
1M ctx · Feb 2026	ExpeL	22.8$_{-0.4}$	23.7$_{+1.3}$	21.2$_{+2.3}$	22.5$_{+2.4}$	22.2$_{\pm 0.0}$	37.4$_{+9.4}$	29.8$_{+3.1}$	24.2$_{+1.6}$	24.5$_{+1.8}$
Text T → T native	HyperGraphRAG	18.0$_{-5.2}$	24.1$_{+1.7}$	17.4$_{-1.5}$	19.5$_{-0.6}$	23.0$_{+0.8}$	31.8$_{+3.8}$	33.6$_{+6.9}$	31.5$_{+8.9}$	24.4$_{+1.7}$
Image → native	MemoTime	19.6$_{-3.6}$	20.4$_{-2.0}$	18.9$_{\pm 0.0}$	18.3$_{-1.8}$	21.8$_{-0.4}$	21.5$_{-6.5}$	29.0$_{+2.3}$	19.4$_{-3.2}$	20.8$_{-1.9}$
Doc → T text layer	ATOM	22.8$_{-0.4}$	24.5$_{+2.1}$	25.0$_{+6.1}$	20.7$_{+0.6}$	25.4$_{+3.2}$	29.9$_{+1.9}$	36.6$_{+9.9}$	28.6$_{+6.0}$	26.1$_{+3.4}$
Video → native	FutureOmni	23.2$_{\pm 0.0}$	28.2$_{+5.8}$	24.2$_{+5.3}$	19.5$_{-0.6}$	25.0$_{+2.8}$	32.7$_{+4.7}$	35.1$_{+8.4}$	32.3$_{+9.7}$	27.1$_{+4.4}$
Audio → T ASR	MiroFlow	21.6$_{-1.6}$	24.9$_{+2.5}$	25.0$_{+6.1}$	23.7$_{+3.6}$	**31.0**$_{+8.8}$	37.4$_{+9.4}$	35.1$_{+8.4}$	33.1$_{+10.5}$	28.3$_{+5.6}$
	TimeSeek	25.2$_{+2.0}$	24.9$_{+2.5}$	**27.3**$_{+8.4}$	22.5$_{+2.4}$	24.6$_{+2.4}$	30.8$_{+2.8}$	35.1$_{+8.4}$	27.8$_{+5.2}$	26.6$_{+3.9}$
	mini-prophet	22.8$_{-0.4}$	21.2$_{-1.2}$	18.2$_{-0.7}$	19.5$_{-0.6}$	25.8$_{+3.6}$	28.0$_{\pm 0.0}$	32.8$_{+6.1}$	28.6$_{+6.0}$	24.4$_{+1.7}$
	MMHRAG	18.4$_{-4.8}$	22.0$_{-0.4}$	18.9$_{\pm 0.0}$	17.8$_{-2.3}$	23.4$_{+1.2}$	33.6$_{+5.6}$	36.6$_{+9.9}$	28.6$_{+6.0}$	24.1$_{+1.4}$
	LEAP	24.8$_{+1.6}$	23.7$_{+1.3}$	24.2$_{+5.3}$	23.1$_{+3.0}$	25.0$_{+2.8}$	26.2$_{-1.8}$	35.1$_{+8.4}$	31.9$_{+9.3}$	26.5$_{+3.8}$
	FutureMap	**34.0**$_{+10.8}$	**36.7**$_{+14.3}$	26.5$_{+7.6}$	**27.8**$_{+7.7}$	27.8$_{+5.6}$	33.6$_{+5.6}$	**39.7**$_{+13.0}$	**40.3**$_{+17.7}$	**33.6**$_{+10.9}$
Gemini 2.5 Flash Lite	No source	20.0	21.6	17.4	20.1	19.4	21.5	28.2	21.0	20.9
T → T	ReAct	19.6$_{-0.4}$	24.1$_{+2.5}$	22.0$_{+4.6}$	13.6$_{-6.5}$	20.2$_{+0.8}$	29.9$_{+8.4}$	26.7$_{-1.5}$	27.4$_{+6.4}$	22.5$_{+1.6}$
$0.100 / $0.400 per 1M	TKG-ICL	19.6$_{-0.4}$	21.2$_{-0.4}$	20.5$_{+3.1}$	14.8$_{-5.3}$	18.5$_{-0.9}$	22.4$_{+0.9}$	22.1$_{-6.1}$	19.8$_{-1.2}$	19.7$_{-1.2}$
1M ctx · Jul 2025	ExpeL	18.8$_{-1.2}$	18.4$_{-3.2}$	17.4$_{\pm 0.0}$	16.6$_{-3.5}$	19.4$_{\pm 0.0}$	29.9$_{+8.4}$	29.8$_{+1.6}$	25.4$_{+4.4}$	21.2$_{+0.3}$
Text T → T native	HyperGraphRAG	21.2$_{+1.2}$	22.0$_{+0.4}$	20.5$_{+3.1}$	17.8$_{-2.3}$	21.8$_{+2.4}$	33.6$_{+12.1}$	30.5$_{+2.3}$	30.2$_{+9.2}$	24.1$_{+3.2}$
Image → native	MemoTime	18.8$_{-1.2}$	21.2$_{-0.4}$	17.4$_{\pm 0.0}$	17.2$_{-2.9}$	19.8$_{+0.4}$	25.2$_{+3.7}$	28.2$_{\pm 0.0}$	23.0$_{+2.0}$	21.0$_{+0.1}$
Doc → T text layer	ATOM	20.4$_{+0.4}$	22.9$_{+1.3}$	17.4$_{\pm 0.0}$	21.9$_{+1.8}$	21.0$_{+1.6}$	32.7$_{+11.2}$	34.4$_{+6.2}$	27.0$_{+6.0}$	23.9$_{+3.0}$
Video → native	FutureOmni	22.8$_{+2.8}$	23.7$_{+2.1}$	22.7$_{+5.3}$	18.3$_{-1.8}$	22.6$_{+3.2}$	30.8$_{+9.3}$	32.1$_{+3.9}$	31.0$_{+10.0}$	25.1$_{+4.2}$
Audio → T ASR	MiroFlow	22.0$_{+2.0}$	22.4$_{+0.8}$	21.2$_{+3.8}$	14.2$_{-5.9}$	16.9$_{-2.5}$	32.7$_{+11.2}$	31.3$_{+3.1}$	25.8$_{+4.8}$	22.5$_{+1.6}$
	TimeSeek	24.8$_{+4.8}$	22.4$_{+0.8}$	22.7$_{+5.3}$	20.7$_{+0.6}$	**27.8**$_{+8.4}$	**37.4**$_{+15.9}$	34.4$_{+6.2}$	29.8$_{+8.8}$	26.8$_{+5.9}$
	mini-prophet	21.2$_{+1.2}$	20.0$_{-1.6}$	21.2$_{+3.8}$	16.6$_{-3.5}$	20.2$_{+0.8}$	24.3$_{+2.8}$	35.1$_{+6.9}$	30.6$_{+9.6}$	23.3$_{+2.4}$
	MMHRAG	23.6$_{+3.6}$	22.0$_{+0.4}$	23.5$_{+6.1}$	18.3$_{-1.8}$	21.8$_{+2.4}$	30.8$_{+9.3}$	31.3$_{+3.1}$	28.2$_{+7.2}$	24.4$_{+3.5}$
	LEAP	24.8$_{+4.8}$	24.5$_{+2.9}$	**25.8**$_{+8.4}$	21.3$_{+1.2}$	27.0$_{+7.6}$	30.8$_{+9.3}$	35.1$_{+6.9}$	31.5$_{+10.5}$	27.2$_{+6.3}$
	FutureMap	**28.4**$_{+8.4}$	**33.1**$_{+11.5}$	**25.8**$_{+8.4}$	**24.9**$_{+4.8}$	27.0$_{+7.6}$	32.7$_{+11.2}$	**40.5**$_{+12.3}$	**38.3**$_{+17.3}$	**31.2**$_{+10.3}$

5 EXPERIMENTS

We answer eight research questions (RQs). **RQ1:** Does FUTUREMAP outperform prior forecasters on OMNIFP-BENCH? **RQ2:** Does each component contribute? **RQ3:** Is the omnimodal temporal hypergraph a better evidence store? **RQ4:** Are the backtested hyper-rules and kernels reliable? **RQ5:** Does deadline-blind branching separate *whether* from *when*? **RQ6:** What does a forecast cost? **RQ7:** How robust is the margin? **RQ8:** How does FUTUREMAP forecast a single case?

5.1 EXPERIMENTAL SETUP

Datasets and Evaluation Metrics. We evaluate on OMNIFP-BENCH (§3, Appendix D): 1,530 items over 712 storylines, eight domains and seven answer types, with 117 negative twins that reward placing an event in time. A system reads only the known zone $\mathcal{K}$, 3,531 resources in five modalities up to the cutoff b, and forecasts outcomes due one to three weeks later. Typed matching grades every answer with no model in the loop; we report accuracy (%) per domain and **Overall**.

Baselines and Ablations. We compare twelve training-free forecasters run from their released code (Appendix E): the agentic REACT (Yao et al., 2023), EXPEL (Zhao et al., 2024), MIROFLOW (Su et al., 2026), TIMESEEK (Mostafa et al., 2026), mini-prophet (Murphy, 2026) and LEAP (Chen et al., 2026d); the temporal-KG TKG-ICL (Lee et al., 2023), MEMOTIME (Tan et al., 2026) and ATOM (Lairgi et al., 2026a); the hypergraph HYPERGRAPHRAG (Luo et al., 2025) and

Table 2: **Ablations of FUTUREMAP** (1,530 items). Each row removes one component, grouped by the stage of §4 it belongs to; subscripts: gap to the full method; —: acts on yes/no only.

Stage	Variant	Overall accuracy Gemini	Qwen	Nemo	Mean	Mean over backbones Yes/No	Numeric
FUTUREMAP (full)		**31.2**	**33.6**	**29.7**	**31.5**	**63.7**	**14.4**
1	− hypergraph (source excerpts only)	27.6$_{-3.6}$	30.8$_{-2.8}$	29.6$_{-0.1}$	29.3$_{-2.2}$	57.7$_{-6.0}$	13.2$_{-1.2}$
2	− causal hyper-rules and anchors	30.1$_{-1.1}$	33.3$_{-0.3}$	28.5$_{-1.2}$	30.6$_{-0.9}$	61.7$_{-2.0}$	13.9$_{-0.5}$
	− magnitude kernels	30.5$_{-0.7}$	33.5$_{-0.1}$	29.1$_{-0.6}$	31.0$_{-0.5}$	63.6$_{-0.1}$	13.2$_{-1.2}$
	− first-passage vote	30.4$_{-0.8}$	32.6$_{-1.0}$	29.0$_{-0.7}$	30.7$_{-0.8}$	59.8$_{-3.9}$	—
3	− horizon curve	30.5$_{-0.7}$	32.9$_{-0.7}$	29.2$_{-0.5}$	30.9$_{-0.6}$	60.8$_{-2.9}$	—
	− persistence branch	30.3$_{-0.9}$	33.0$_{-0.6}$	29.2$_{-0.5}$	30.8$_{-0.7}$	63.6$_{-0.1}$	12.7$_{-1.7}$

MMHRAG (Chen et al., 2026b); and the omnimodal FUTUREOMNI (Chen et al., 2026a). *No source* answers from the backbone's weights alone, and six ablations each drop one component (Table 2).

Implementation Details. Every row runs on three frozen backbones via OpenRouter: Mistral Nemo (text), Qwen3.5 Flash (+image, +video) and Gemini 2.5 Flash-Lite (+audio, +file), all released before t_{embed}, at temperature 0. All methods search one day-sharded index over $\mathcal{K}$ that never loads a post-cutoff resource; media a backbone cannot read pass through fixed parsers (shaded in Table 1). Every store is built once for all backbones, with a single hyperparameter setting (Appendix C).

5.2 MAIN RESULTS (RQ1)

As shown in Table 1, FUTUREMAP is first on **Overall** on all three backbones, ahead of the strongest baseline by +4.1, +5.3 and +4.2 points on Gemini, Qwen and Nemo, and first or tied-first in 18 of 24 domain cells. It gains +9.4 to +10.9 over *No source*, against at most +6.3 for any baseline, and is the only row above that floor in all 24 cells. Averaged over the backbones (Figure 5), it leads the best baseline of every domain, by +1.5 on Technology up to +9.5 on Geopolitics.

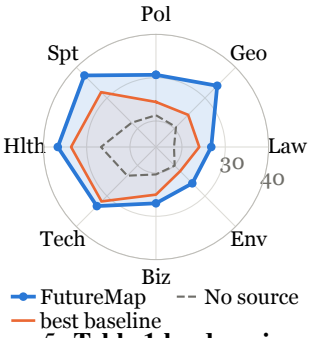

Figure 5: **Table 1 by domain**, averaged over the backbones.

5.3 ABLATION STUDY (RQ2)

Every removal in Table 2 lowers **Overall** on all three backbones and on their mean, so each component contributes. Dropping the hypergraph, stage 1, costs the most, 3.6 points on Gemini and 6.0 on yes/no; each other component lifts the type it targets: on yes/no, hyper-rules and anchors +2.0, the first-passage vote +3.9 and the horizon curve +2.9; on numeric items, the magnitude kernels +1.2 and the persistence branch +1.7.

5.4 ANALYSIS OF THE OMNIMODAL TEMPORAL HYPERGRAPH (RQ3)

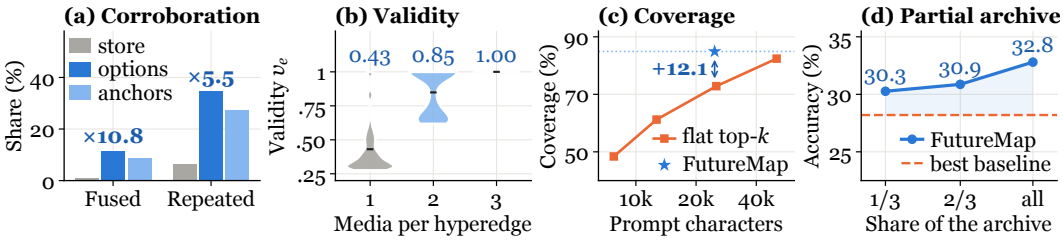

Figure 6: **Stage 1 as an evidence store**: corroboration, validity by media, coverage, data efficiency.

Corroboration Is Promoted. As shown in Figure 6, fused hyperedges are 1.1% of the store but 11.4% of rule options (×10.8), repeated ones rise from 6.3% to 34.7% (a), and mean validity v_e grows from 0.43 in one medium to 0.85 in two and 1.00 in three, in line with Proposition 1 (b).

Dense and Data-Efficient. The prompt covers 84.9% of an item's gold precursors, 12.1 points above flat retrieval of equal length (c); built from a third of $\mathcal{K}$, FUTUREMAP already beats the best baseline on the full archive, 30.3 vs. 28.2, and 32.8 with all of it (d; backbone mean, 500 items).

5.5 ANALYSIS OF THE BACKTESTED HYPER-RULES AND KERNELS (RQ4)

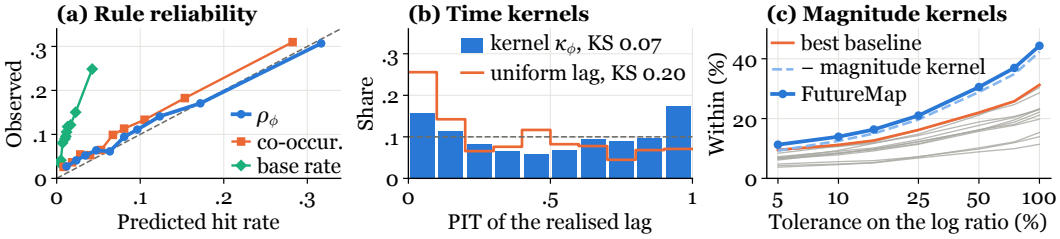

Figure 7: **Stage 2 re-mined up to** $b-14$ and scored on the next 14 days (a, b); numeric accuracy (c).

Calibrated Out of Sample. As shown in Figure 7, on unseen days the backtested ρ_ϕ tracks the observed hit rate (ECE 0.014) better than co-occurrence (0.023) or the base rate (0.093; a), and the realised lags have near-uniform PIT under the time kernels (KS 0.074 vs. 0.203 for a uniform lag; b).

Magnitude Kernels Help. On 500 positive numeric items, FUTUREMAP leads every baseline at every tolerance (+1.8 to +13.0), and removing the magnitude kernel lowers the whole curve (c).

5.6 ANALYSIS OF DEADLINE-BLIND BRANCHING AND READ-OUT (RQ5)

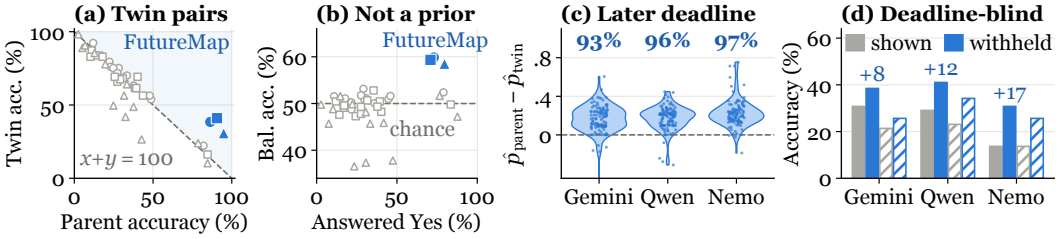

Figure 8: **Separating *whether* from *when*** on 117 pairs: *Yes* by one deadline, *No* by an earlier one.

Only FUTUREMAP Tells When. As shown in Figure 8, the other rows (one point per row and backbone) cluster near $x+y=100$, where answers ignore the deadline, while FUTUREMAP sits well above it on all three backbones (a) and leads balanced yes/no accuracy, 58.5–59.9 vs. $\leq$53.1 (b).

Deadline-Blind by Design. A parent's read-out tops its twin's in 93–97% of pairs and never trails it on a shared deduction (Proposition 3; c); hiding the deadline lifts twin accuracy 7.7–17.1 (d).

5.7 ANALYSIS OF COST AND ROBUSTNESS (RQ6, RQ7)

On the Cost–Accuracy Frontier. As shown in Figure 9a, FUTUREMAP makes one deduction per item for 1.58, 2.09 and 0.17 thousandths of a dollar on Gemini, Qwen and Nemo, below 9, 9 and 8 of the 12 baselines; none is more accurate at any cost, and each cheaper one trails by $\geq$ 6.1 points.

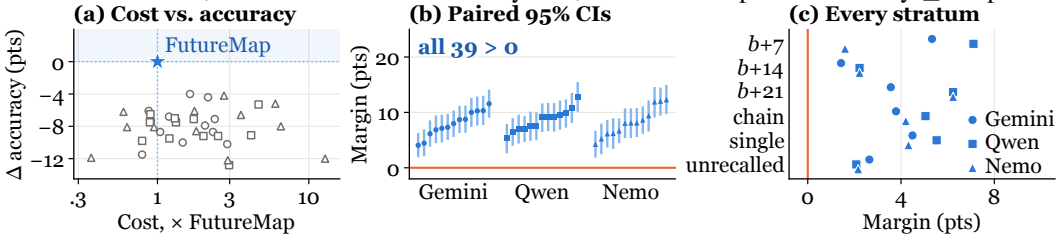

Figure 9: **Cost and robustness**: (a) baselines relative to FUTUREMAP; (b, c) its margin.

Significant in Every Stratum. All 39 paired 95% CIs exclude zero (lowest bound +1.8; b), and the margin over each stratum's best baseline is positive in all 18 cells, even on the 1,249 items not answerable from memory (c); FUTUREMAP alone solves 42, 48 and 31 items, 1.6–2× any baseline.

Table 3: **Case study (RQ8)**: the known-zone evidence every row can read, and each row's answer on Gemini / Qwen / Nemo; quotes are verbatim from the boxed run, decisive words in red.

Query (Politics, 21-day horizon). *Will Todd Blanche be confirmed with a Senate vote margin of less than 3 by 2026-08-11?* Cutoff b = 2026-07-21. **Answer:** *Yes*, confirmed 50–49 (reported 2026-08-11).

V C-SPAN2 clip, 2026-07-16



"Attorney General Nominee Todd Blanche Testifies at Confirmation Hearing", 6 keyframes

A NPR Up First podcast, 2026-07-15



0:00 "...Todd Blanche faces a Senate confirmation hearing today"
4:46 "only a couple of GOP defections could kill..."

I LA Times photo, 07-16



"...Blanche must meet Epstein's accusers to earn his vote for attorney general"

T Semafor, 06-10: "just one GOP rebel could stop the whole thing in the Judiciary Committee"

Family	Method	G	Q	N	What decides its answer
—	No source	✗	✗	✗	"He was confirmed by the Senate on February 4, 2025, … margin of 53–47"
Agentic	ReAct	✗	✗	✗	"it is highly unlikely that Todd Blanche will be confirmed"
	ExpeL	✓	✗	✗	"I must default to 'No' … the absence of affirmative evidence"
	MiroFlow	✗	✗	✗	"multiple obstacles makes a larger margin more likely if confirmed"
	TimeSeek	✓	✓	✗	"it's unlikely that Todd Blanche will be confirmed"
	mini-prophet	✗	✓	✗	"Submitted forecast: {"Yes": 0.45, "No": 0.55}"
	LEAP	✗	✗	✗	"LEAP posterior over options: – Yes: 0.000 – No: 1.000"
Temporal KG	TKG-ICL	✗	✓	✓	"no information about the outcome of the vote or the margin"
	MemoTime	✗	✗	✗	"do not contain information regarding the Senate vote margin"
	ATOM	✗	✗	✗	"no direct evidence in the facts to predict such a tight margin"
Hypergraph	HyperGraphRAG	✗	✓	✗	"It is more likely that he will not be confirmed"
	MMHRAG	✗	✗	✓	"if any one of them votes against him, the nomination could fail"
Omnimodal	FutureOmni	✗	✓	✗	"it's difficult to definitively predict a margin of *less than 3*"
Ours	**FUTUREMAP**	✓	✓	✓	two grounded chains date a confirmation before 08-11 (below)

Stage 1: hyperedges	Stage 2: rules	Stage 3: branches (Qwen)	Read-out
T T v = .75 07-15 vote: committee decides on sending it to the Senate	vote → appoint lag p10 / p50 / p90 3 / 11 / 33 d	p = **.45** 07-23 committee advances it 07-28 confirmed, 51–49	$\hat{p}$ at 08-11 **Qwen .71** Gemini .60 Nemo .75 **Yes ✓** actual 50–49
A T T V **v = 1.00 (fused)** 07-15 meet: Blanche faces a Senate confirmation hearing	meet → meet → appoint lag p10 / p50 / p90 7 / 18 / 35 d	p = **.30** 07-22 Tillis blocks the vote 08-04 confirmed after delay	
same-day single reports: v = .50		p = .25: withdrawn or stalled	

5.8 CASE STUDY (RQ8)

Baselines Stop Where the Archive Stops. As shown in Table 3, without a reported margin, 27 of the 36 baseline runs miss the answer: agents read the gap as evidence against, graph retrievers find no stored fact to return, and *No source* on Qwen recalls a 2025 Senate vote that never happened.

FutureMap Deduces the Vote. Fused from four reports in three media, the hearing hyperedge reaches v=1.00; mined rules carry it and a committee report to two confirmations dated before the deadline, and the read-out answers *Yes* on all three backbones, with $\hat{p}_q$ from .60 to .75 by 08-11.

6 CONCLUSION

We introduced OmniFP, forecasting from an omnimodal archive closed at a cutoff, with OMNIFP-BENCH, 1,530 dated items over five media, and FUTUREMAP, which answers it in three stages: an omnimodal temporal hypergraph fuses the reports of one event across media, backtested hyper-rules and kernels measure what follows and when, and deadline-blind branches are read out at any deadline. It is the most accurate forecaster on all three backbones, 4.1 to 5.3 points above the strongest baseline with every paired interval above zero, separates *whether* from *when* on negative twins, and costs one grounded call per item. Streaming archives are a natural next step.

AI USE STATEMENT

We used generative AI tools only to assist with language editing of the manuscript and with code implementation, while the research questions, method design, experiments and conclusions are the authors' own; the LLMs that appear inside FUTUREMAP and the benchmark pipeline are components of the studied system and are fully specified in §3, §4 and Appendix A. All AI-assisted text and code were reviewed, verified and tested by the authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

OMNIFP-BENCH is built entirely from publicly available sources (a public news crawl, official public records, public podcast feeds and a public broadcast archive; Appendix D.2), and the study involves no human subjects, crowdsourced annotation or private personal data. FUTUREMAP is a training-free research system whose every forecast is grounded in cited, dated evidence, and we release it to support transparent and auditable research on forecasting in accordance with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The source code of FUTUREMAP and of the evaluation harness is available at the anonymous link in the abstract, and every model call runs on frozen, pinned backbones at temperature 0 with a fixed seed and a disk cache (§5.1, Appendix E). The appendix provides all prompts (Appendix A), complete proofs (Appendix B), the full algorithm with hyperparameters (Appendix C), the benchmark construction and evaluation pipeline (Appendix D), the pinned commit of every baseline (Appendix E) and the per-experiment protocol (Appendix F).

REFERENCES

Kian Ahrabian, Eric Boxer, and Jay Pujara. Toward better temporal structures for geopolitical events forecasting, 2026. URL https://arxiv.org/abs/2601.00430.

Zafar Ali, Asad Khan, Aalia Malik, and Pavlos Kefalas. Graphloom: Reliability-calibrated graph evidence routing for multimodal kg-rag, 2026. URL https://arxiv.org/abs/2608.15056.

Federico Barrera-Lemarchand, Mariano Sigman, and Joaquin Navajas. The wisdom of artificial deliberative crowds, 2026. URL https://arxiv.org/abs/2609.22497.

Wanxu Cai, Zhengyu Chen, Huaisheng Zhu, Wei Wang, Jingang Wang, and Qiang Xu. Distilling temporal search and reasoning: Evolving llms for future prediction via harness-assisted efficient data synthesis, 2026. URL https://arxiv.org/abs/2607.25554.

Yuanpu Cao, Yongkang Du, Yurui Chang, Lu Lin, and Jinghui Chen. From narrative to auditable forecasts: A structured scaffold for agentic forecasting, 2026. URL https://arxiv.org/abs/2609.05905.

Jingyi Chai, Bingyang Zheng, Xiangrui Liu, Hao Lu, Zihang Zhou, Tianchen Wang, Kemeng Zhang, and Siheng Chen. Agentic time machine as an infrastructure for future-event forecasting, 2026. URL https://arxiv.org/abs/2606.21013.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling open-ended reasoning to predict the future, 2026. URL https://arxiv.org/abs/2512.25070.

He Chang, Zhulin Tao, Lifang Yang, Xianglin Huang, and Yunshan Ma. Scattered hypothesis generation for open-ended event forecasting. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17288–17304, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.855. URL https://aclanthology.org/2026.findings-acl.855/.

Yurui Chang, Yongkang Du, Yuanpu Cao, Jinghui Chen, and Lu Lin. Forecastcompass: Guiding agentic forecasting with adaptive factor memory, 2026b. URL https://arxiv.org/abs/2605.30858.

Qian Chen, Jinlan Fu, Changsong Li, Min Zhang, See-Kiong Ng, and Xipeng Qiu. Futureomni: Evaluating future forecasting from omni-modal context for multimodal LLMs. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=76zKCPBKXP.

Shenao Chen, Yidan Xu, Xiangmin Han, Rundong Xue, Duanpo Wu, Yuhan Gao, Chenggang Yan, and Yue Gao. Hypergraph-based multimodal retrieval-augmented generation with incremental refinement, 2026b. URL https://arxiv.org/abs/2608.16628.

Yongqiang Chen, Guangyi Chen, Yuewen Sun, and Kun Zhang. Analogical deep research: Retrieving and integrating historical analogies for foresight analysis, 2026c. URL https://arxiv.org/abs/2607.13602.

Yufei Chen, Yiran Zhao, Xiaogang Xu, Qipeng Xie, Jiafei Wu, and Zhe Liu. Leap: Likelihood elicitation and aggregation for llm-based probabilistic forecasting, 2026d. URL https://arxiv.org/abs/2609.01337.

Nirupam Chetlapalli, Yiming Liao, Min-Chun Chen, and Keke Chen. Diverse by reasoning: Harnessing the wisdom of llm crowds for future prediction, 2026. URL https://arxiv.org/abs/2608.24001.

Hailong Chu, Hongbing Li, Yunlong Chu, Shutai Huang, Xingyue Zhang, Tinghe Yan, Jinsong Zhang, Shuo Zhang, and Lei Li. Echo: Event-centric hypergraph operations via multi-agent collaboration for multimedia event extraction, 2026. URL https://arxiv.org/abs/2603.06683.

Yuzhuo Cui, Zongye Zhang, and Qingjie Liu. Livingrag: Augmenting graph rag with experience, 2026a. URL https://arxiv.org/abs/2608.25960.

Zeyang Cui, Jiannong Cao, Zhiyuan Wen, Bo Yuan, Junlan Feng, and Shengyuan Chen. Edgemem: Llm-free agent memory construction and retrieval via evidence-preserving multi-anchor hypergraph, 2026b. URL https://arxiv.org/abs/2609.05553.

Hui Dai, Ryan Teehan, Parsa Torabian, and Mengye Ren. Aligning llms with human uncertainty: A beta-bernoulli calibrator for llm forecasting, 2026. URL https://arxiv.org/abs/2605.27668.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Yichao Feng, Ran Zhang, Haoran Luo, Zhenghong Lin, Carl Yang, and Anh Tuan Luu. Diachronic hypergraphs for orchestrated multi-agent multimodal memory curation, 2026a. URL https://arxiv.org/abs/2608.29678.

Yifan Feng, Hao Hu, Shihui Ying, Xingliang Hou, Shiquan Liu, Mingyuan Yang, Junchang Li, Shaoyi Du, Nanning Zheng, Han Hu, and Yue Gao. Hyper-rag: combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *Nature Communications*, 17(1):5778, Apr 2026b. ISSN 2041-1723. doi: 10.1038/s41467-026-71411-1. URL https://doi.org/10.1038/s41467-026-71411-1.

Julia Gastinger, Christian Meilicke, and Heiner Stuckenschmidt. Counttrucola: Rule learning for interpretable temporal knowledge graph forecasting, 2026. URL https://arxiv.org/abs/2509.09474.

Xiaokun Guo, Zhen Xu, Dongdong Huo, Yanqiu Zhang, Dongjin Yu, and Yu Wang. Scop: Structured constraint parsing for evidence-space control in temporal knowledge graph question answering, 2026. URL https://arxiv.org/abs/2609.22213.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 50426–50468. Curran Associates, Inc., 2024. doi: 10.52202/079017-1598. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5a5acfd0876c940d81619c1dc60e7748-Paper-Conference.pdf.

Zhixin Han, Yanzhi Zhang, Chuyang Wei, Maohang Gao, Xiawei Yue, Kefei Chen, Yu Zhuang, Haoxiang Guan, Jiyan He, Jian Li, Yitong Duan, Yu Shi, Mengting Hu, and Shuxin Zheng. Futureworld: A live reinforcement learning environment for predictive agents with real-world outcome rewards, 2026. URL https://arxiv.org/abs/2604.26733.

Zihao Jiang, Miao Peng, Zhenyan Shan, Wenjie Xu, Ben Liu, Gong Chen, Ziqi Gao, and Min Peng. Tkg-thinker: Towards dynamic reasoning over temporal knowledge graphs via agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2602.05818.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93943–93980, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea74e45a229dac70b5b63b28d8934db6-Paper-Conference.pdf.

Houda Khrouf, Pedro Fillastre, and Sebastiao Correia. Optimizing hypergraph-based rag: Toward better fact extraction and chunk retrieval, 2026. URL https://arxiv.org/abs/2607.20506.

Yassir Lairgi, Ludovic Moncla, Khalid Benabdeslem, Rémy Cazabet, and Pierre Cléau. ATOM: AdapTive and OptiMized dynamic temporal knowledge graph construction using LLMs. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 950–966, Rabat, Morocco, March 2026a. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.49. URL https://aclanthology.org/2026.findings-eacl.49/.

Yassir Lairgi, Ludovic Moncla, Khalid Benabdeslem, Rémy Cazabet, and Pierre Cléau. C-unseen: Weak signal detection in dynamic temporal knowledge graphs via llm reasoning, 2026b. URL https://arxiv.org/abs/2608.26870.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. Temporal knowledge graph forecasting without knowledge using in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 544–557, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.36. URL https://aclanthology.org/2023.emnlp-main.36/.

Ningyuan Li, Haihong E, Tianyu Yao, Tianyi Hu, Yuhan Li, Haoran Luo, Meina Song, and Yifan Zhu. Infer: A neural-symbolic model for extrapolation reasoning on temporal knowledge graph. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 91993–92008, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e5a71ba556c84fef542aaace56b6cfe9-Paper-Conference.pdf.

Zehan Li, Yuxuan Wang, Ali El Lahib, Ying-Jieh Xia, and Xinyu Pi. Simulated ignorance fails: A systematic study of llm behaviors on forecasting problems before model knowledge cutoff. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 5837–5845. International Joint Conferences on Artificial Intelligence Organization, 8 2026. doi: 10.24963/ijcai.2026/650. URL https://doi.org/10.24963/ijcai.2026/650. Main Track.

Wen-Sheng Lien, Yu-Kai Chan, Hao-Lung Hsiao, Bo-Kai Ruan, Meng-Fen Chiang, Chien-An Chen, Yi-Ren Yeh, and Hong-Han Shuai. Hyperrag: Reasoning n-ary facts over hypergraphs for retrieval augmented generation. In *Proceedings of the ACM Web Conference 2026*, pp. 2465–2476.

ACM, April 2026. doi: 10.1145/3774904.3792710. URL http://dx.doi.org/10.1145/3774904.3792710.

Jiashi Lin, Changhong Jiang, Xiangru Lin, Ruifei Zhang, Xinyi Zhu, Jiyao Liu, Cheng Tang, Ye Du, Shujian Gao, Junzhi Ning, Lihao Liu, Ziyan Huang, Tianbin Li, Jin Ye, and Junjun He. Evograph-r1: Self-evolving multimodal knowledge hypergraphs for agentic retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 756–765, June 2026.

Jiashuo Liu, Siyuan Chen, Zaiyuan Wang, Zhiyuan Zeng, Jiacheng Guo, Liang Hu, Lingyue Yin, Suozhi Huang, Wenxin Hao, Yang Yang, Zerui Cheng, Zixin Yao, Lingyue Yin, Haoxin Liu, Jiayi Cheng, Yuzhen Li, Zezhong Ma, Bingjie Wang, Bingsen Qiu, Xiao Liu, Zeyang Zhang, Zijian Liu, Jinpeng Wang, Mingren Yin, Tianci He, Yali Liao, Yixiao Tian, Zhenwei Zhu, Anqi Dai, Ge Zhang, Jingkai Liu, Kaiyuan Zhang, Wenlong Wu, Xiang Gao, Xinjie Chen, Zhixin Yao, Zhoufutu Wen, B. Aditya Prakash, Jose Blanchet, Mengdi Wang, Nian Si, and Wenhao Huang. Futurex-pro: Extending future prediction to high-value vertical domains, 2026a. URL https://arxiv.org/abs/2601.12259.

Wenjin Liu, Shen Pang, Chenxi Wang, Tiesunlong Shen, Jiajie He, Zhe Cui, Xiaobao Wu, Anh Tuan Luu, and Haoran Luo. Tiem: Temporal integration of hypergraph evidence and skill memory for event-driven financial forecasting, 2026b. URL https://arxiv.org/abs/2608.13024.

Yuan Lu and Yukuan Zhang. The corroboration illusion: When more news makes llm forecasts less true, 2026. URL https://arxiv.org/abs/2609.22246.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 152206–152234. Curran Associates, Inc., 2025. doi: 10.52202/085713-5089. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/df55ee6e59f8ac4a625219e11fe9ddba-Paper-Conference.pdf.

Haoran Luo, Haihong E, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, and Anh Tuan Luu. Graph-r1: Towards agentic graphRAG framework via end-to-end reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YXnFGsSkCC.

Hamza Mostafa, Om Shastri, and Dennis Lee. Timeseek: Temporal reliability of agentic forecasters, 2026. URL https://arxiv.org/abs/2604.04220.

Kevin Murphy. Agentic forecasting using sequential bayesian updating of linguistic beliefs, 2026. URL https://arxiv.org/abs/2604.18576.

Luyao Niu, Zepu Wang, Shuyi Guan, Yang Liu, and Peng Sun. Event-causnet: Unlocking causal knowledge from text with large language models for reliable spatio-temporal forecasting, 2025. URL https://arxiv.org/abs/2511.12769.

An Nguyen Phu, Dung Nguyen Quang, Luu Hieu An, Linh Ngo Van, Trung Le, and Thien Huu Nguyen. Hyperprove: Answer-guided hypergraph expansion for multi-hop question answering, 2026. URL https://arxiv.org/abs/2609.13768.

Manuel Serna-Aguilera, Raegan Anderes, Page Dobbs, and Khoa Luu. Nico-rag: Multimodal hypergraph retrieval-augmented generation for understanding the nicotine public health crisis, 2026a. URL https://arxiv.org/abs/2603.02047.

Manuel Serna-Aguilera, Raegan Anderes, Page Dobbs, and Khoa Luu. Prism-rag: Multimodal hypergraph retrieval-augmented generation for tobacco product and legislative policy reasoning, 2026b. URL https://arxiv.org/abs/2609.23769.

Shiqian Su, Sen Xing, Xuan Dong, Muyan Zhong, Bin Wang, Xizhou Zhu, Yuntao Chen, Wenhai Wang, Yue Deng, Pengxiang Zhu, Ziyuan Liu, Tiantong Li, Jiaheng Yu, Zhe Chen, Lidong Bing, and Jifeng Dai. Miroflow: Towards high-performance and robust open-source agent framework for general deep research tasks, 2026. URL https://arxiv.org/abs/2602.22808.

Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. Memotime: Memory-augmented temporal knowledge graph enhanced large language model reasoning. In *Proceedings of the ACM Web Conference 2026*, pp. 4220–4231. ACM, April 2026. doi: 10.1145/3774904.3792581. URL http://dx.doi.org/10.1145/3774904.3792581.

Aditi Tiwari, Aashrith Bandaru, and Heng Ji. Competence-gated pooling of language models and priors for event forecasting, 2026. URL https://arxiv.org/abs/2609.12101.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=bbhdeL8EUX.

Benjamin Turtel, Paul Wilczewski, Kris Skotheim, Ville A. Satopää, and Philip E. Tetlock. How proper scoring rules shape llm forecasting, 2026. URL https://arxiv.org/abs/2608.28482.

Zhenran Wang, Zhonghan Bian, Jinsong Li, and Zhangyang Qi. Societybench: Forecasting counterfactual social-world evolution, 2026. URL https://arxiv.org/abs/2608.04009.

Chuyang Wei, Maohang Gao, Zhixin Han, Kefei Chen, Yu Zhuang, Haoxiang Guan, Yanzhi Zhang, Yilin Cheng, Xiren Zhou, Huanhuan Chen, Jian Li, Jiyan He, Yu Shi, Yitong Duan, and Shuxin Zheng. Harnessing pre-resolution signals for future prediction agents, 2026. URL https://arxiv.org/abs/2604.15719.

Keshu Wu, Chenchen Kuai, Zihao Li, Jiwan Jiang, Shiyu Shen, Shian Wang, Chan-Wei Hu, Zhengzhong Tu, and Yang Zhou. Knowledge is not static: Order-aware hypergraph rag for language models, 2026. URL https://arxiv.org/abs/2604.12185.

Siheng Xiong, Yuan Yang, Faramarz Fekri, and James Clayton Kerce. TILP: Differentiable learning of temporal logical rules on knowledge graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_X12NmQKvX.

Ruiyao Xu, Tiankai Yang, and Wei-Chieh Huang. Hyperskill: Self-evolving llm agents via hypergraph-structured skill memory, 2026a. URL https://arxiv.org/abs/2608.16114.

Xiaogang Xu, Jiaqi Tang, Jianmin Chen, Yingying Yan, Zhenchao Tang, Xiangxin Zhou, Xiaobin Hu, Wei Wei, Jinfeng Wu, Qifeng Chen, Lu Zhou, Jiafei Wu, Zhe Liu, Jianwei Yin, and Weimin Zheng. Llm-based agents for forecasting and prediction: Methods, training, evaluation, and applications, 2026b. URL https://arxiv.org/abs/2608.23058.

Neeraj Yadav. Temporal validity in retrieval memory: Eliminating stale-fact errors for ai agents over evolving knowledge, 2026. URL https://arxiv.org/abs/2606.26511.

Peizheng Yan, Yu Zhao, Liang Xie, Juntong Qi, Mingming Wang, and Erwei Yin. Event-causal rag: A retrieval-augmented generation framework for long video reasoning in complex scenarios, 2026. URL https://arxiv.org/abs/2605.06185.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. Mirai: Evaluating llm agents for event forecasting, 2024. URL https://arxiv.org/abs/2407.01231.

Xiao Ye, Jacob Dineen, Evan Zhu, Shijie Lu, Kevin Song, and Ben Zhou. Hindcast: Replaying prediction markets to evaluate llm forecasters, 2026. URL https://arxiv.org/abs/2607.14051.

Minwoo Yu and Young guk Ha. From explanations to interventions: Execution-guided counterfactual synthesis in temporal graphs, 2026. URL https://arxiv.org/abs/2609.11170.

Xuanqing Yu, Wangtao Sun, Jingwei Li, Kang Liu, Chengbao Liu, and Jie Tan. ONSEP: A novel online neural-symbolic framework for event prediction based on large language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6335–6350, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.378. URL https://aclanthology.org/2024.findings-acl.378/.

Xiangjun Zai, Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, and Wenjie Zhang. Proh: Dynamic planning and reasoning over knowledge hypergraphs for retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026*, pp. 4256–4267, 2026.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Xinjie Chen, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, Xiangsheng Li, Jingkai Liu, Shuang Ni, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xin Zhou, Jose Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. Futurex: An advanced live benchmark for llm agents in future prediction. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 19522–19563, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/21315abf210cef71317e887e9cda78b3-Paper-Conference.pdf.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i17.29936. URL http://dx.doi.org/10.1609/aaai.v38i17.29936.

Feng Zhao, Kangzheng Liu, Teng Peng, Yu Yang, and Guandong Xu. Dymrl: Dynamic multi-space representation learning for multimodal event forecasting in knowledge graph, 2026. URL https://arxiv.org/abs/2603.24636.

Ziwei Zheng, Peiqiong Chen, and Bang Wang. Abstract event causal rules: Induction and application, 2026. URL https://arxiv.org/abs/2608.05205.

Linhao Zhong, Zongze Du, Linyu Wu, Yu Bo, Hourong Li, Chenchen Jing, Hao Chen, Yuling Xi, and Chunhua Shen. Foredreamer: A self-evolving dual-agent memory architecture for future event prediction, 2026. URL https://arxiv.org/abs/2608.20920.

APPENDIX

A PROMPTS USED IN FUTUREMAP

FUTUREMAP calls a model at two points: the builder Read reads each resource once (§4.1) and the backbone LM deduces once per item (§4.3), with the same prompt on every backbone; stage 2 (§4.2) has no prompt, its rules, kernels and anchors being measured in code. Red marks what is filled per resource or per item, blue what depends on the answer type.

A.1 HYPERGRAPH CONSTRUCTION PROMPT

As shown in Figure 10, each call reads one resource r in its own medium: an article, a document or a transcript as text (its first 7,000 characters), a photograph or a TV clip attached natively, its body line naming only the article a photograph accompanied or a clip's channel and air date. The reply is $\mathsf{Read}(r)$ of Eq. (2): an event is a report $\tilde{e}$ with type $\theta_{\tilde{e}}$ (`type`), participants $\mathcal{V}_{\tilde{e}}$ (`subj`, `obj`) and announced date $t^{\mathrm{sch}}_{\tilde{e}}$ (`on`); a quantity is a point $\pi = (u, k, y, t_\pi)$. The 30 glosses of Θ also give each question its type θ_q in stage 3. The builder is Gemini 2.5 Flash-Lite, run once over the archive, so every backbone reads the same hypergraph.

```
You are building a temporal event hypergraph from a news archive. Read the resource below ({kind}) and extract the events it attests and the quantities it reports.
Resource: "{title}" | source: {source} | published: {date}
{body}
Event types (use the key):
- announce: announces, discloses or confirms something
- schedule: sets a date, deadline or plan for a future event
- meet: meets, visits, talks or negotiates with
- agree: signs, agrees, concludes a deal or treaty
- propose: proposes, introduces, files or tables a bill, plan or bid
- approve: approves, passes, authorises, confirms or ratifies
- reject: rejects, blocks, vetoes, cancels or suspends
- vote: holds a vote or an election
- appoint: appoints, nominates, elects, hires or promotes someone
- resign: resigns, is dismissed, fired or leaves a post
- charge: sues, charges, indicts, investigates or accuses
- arrest: arrests, detains, raids or seizes
- rule: a court or regulator rules, sentences or issues a verdict or order
- sanction: imposes sanctions, tariffs, bans, restrictions or penalties
- attack: attacks, strikes, bombs, fights or launches military action
- deescalate: agrees a ceasefire, truce, withdrawal, release of prisoners or aid
- protest: protests, strikes, demonstrates or boycotts
- report: publishes data, statistics, a report, survey or poll
- market: a price, rate, index, currency or stock moves
- earnings: reports earnings, revenue, profit, guidance or financial results
- launch: launches, releases, unveils or ships a product, service or mission
- invest: invests, acquires, merges, funds, raises capital or buys
- disaster: a natural disaster, fire, flood, storm, heatwave or accident occurs
- health: an outbreak, case count, drug approval or health measure
- casualty: people are killed, injured, displaced or evacuated
- sport_result: a match, race or tournament result, score or standing
- sport_move: a player signs, transfers, is injured or a contract is agreed
- statement: says, warns, threatens, criticises or comments
- forecast: predicts, expects or projects a future outcome
- other: anything else

Return JSON only:
{"summary": "<one sentence: what this resource shows or reports>",
  "events": [
    {"s": "<one self-contained sentence: who did / said / announced / scheduled what; full names, no pronouns; every number and date the resource gives for it>",
      "type": "<event type key>",
      "subj": ["<acting participant>", "..."],
      "obj": ["<participant acted upon, or the thing concerned (a bill, product, place, team)>", "..."],
      "kind": "occurred" | "scheduled" | "expected",
      "on": "<YYYY-MM-DD the event is scheduled or expected for, if the resource states a future date; else null>"}
  ],
  "quantities": [
    {"entity": "<who or what the number is about>", "indicator": "<what is measured, e.g. 'death toll', 'share price', 'votes in favour', 'evacuees'>",
      "value": <number>, "unit": "<unit or null>", "as_of": "<YYYY-MM-DD the value refers to, or null>"}
  ]}
At most 8 events (the most forecast-relevant first: decisions, votes, results, deadlines, schedules, plans, conflicts, rulings) and at most 8 quantities. Use the resource's own facts only.
```

Figure 10: Prompt for reading one resource, Read in Equation (2). {kind} is *a news article*, *a primary document*, *a podcast transcript*, *a news photograph, shown below* or *a TV news clip with its audio, shown below*; the type list is the ontology Θ.

A.2 DEDUCTION PROMPT

As shown in Figure 11, the one call per item gives LM of Eq. (9) its context $\mathbf{x}_q$ in labelled blocks: ACTORS (up to four entities of $\mathcal{V}_q$, five latest hyperedges each), CAUSAL OPTIONS ($\mathcal{O}_q$, with ρ_ϕ and the 10th, 50th and 90th lag percentiles of κ_ϕ), SCHEDULED ANCHORS ($\mathcal{A}_q$, with $\rho_{\mathcal{A}}$), TIMELINE ($\mathcal{N}_q$, with media, source and validity v_e) and SOURCE EXCERPTS (Ξ_q), then up to two photos and one clip. A reply branch is $\psi_j = (p_j, t_j, y_j, \mathrm{via}_j)$. A yes/no target omits its deadline, and `by` states C_q at $b{+}7$, $b{+}14$ and $b{+}21$ days; Table 4 gives what other types use instead.

You are the deduction stage of a forecasting system. Today is 2026-07-21; the archive ends today and nothing after it is known. The actors below can only move the story through the CAUSAL OPTIONS -- rules mined from the archive, each with its backtested reliability (rho) and the observed lag from precursor to consequence -- and the SCHEDULED ANCHORS, announcements whose date the archive gives.

TARGET: Will Anthony Joshua defeat Kristian Prenga?

ACTORS (actor: recent memory):
- Anthony Joshua: 2026-07-17 Anthony Joshua makes his return against Kristian Prenga in July. / 2026-07-21 Barry Hearn predicts Anthony Joshua will deliver a very, very painful night for Kristian Prenga. / 2026-07-21 Anthony Joshua will return to the ring for the first time this year in Jeddah, Saudi Arabia on Saturday. [...]

CAUSAL OPTIONS (id | precursor already in the archive | rule chain | rho | lag days p10/p50/p90):
O1 | 2026-07-17 Daniel Dubois brutally knocked out Anthony Joshua in September 2024. | sport_result => sport_result | 0.57 | 1/6/17 [...]
O3 | 2026-07-17 Anthony Joshua makes his return against Kristian Prenga in July. | sport_move => sport_result | 0.38 | 1/6/18 [...]
O8 | 2026-07-17 Anthony Joshua makes his return against Kristian Prenga in July. | sport_move => sport_move => sport_result | 0.13 | 6/16/35 [...]
O10 | 2026-07-17 Anthony Joshua is scheduled to take on Kristian Prenga on Saturday, July 25, 2026. | schedule => sport_result | 0.10 | 1/6/22 [...]

SCHEDULED ANCHORS (id | announced event | scheduled date | how often such announcements came true on time in the archive):
[...] A4 | Anthony Joshua is scheduled to take on Kristian Prenga on Saturday, July 25, 2026. | 2026-07-25 | 0.51
A5 | The main card for Anthony Joshua vs Kristian Prenga begins at 1 p.m. EST / 6 p.m. BST on July 25, 2026. | 2026-07-25 | 0.51

TIMELINE (oldest first; id | date | type | media, source | validity | event):
[...] E20 | 2026-07-17 | schedule | article, si.com | v=0.57 | Anthony Joshua is scheduled to take on Kristian Prenga on Saturday, July 25, 2026. [scheduled for 2026-07-25] [...]
E25 | 2026-07-17 | sport_result | article, si.com | v=0.57 | Daniel Dubois brutally knocked out Anthony Joshua in September 2024. [...]
SOURCE EXCERPTS: [...]
[T] Barry Hearn predicts Anthony Joshua to deliver 'very, very painful night' for Kristian Prenga (africa.espn.com, 2026-07-21)
Barry Hearn has predicted Kristian Prenga will endure a "very, very painful night" against a fired-up and focused Anthony Joshua on Saturday. [...]

Deduce forward from today. Keep 2-4 mutually exclusive branches of how the story unfolds. Each branch fires the options and anchors its actors' situation makes likely -- cite them by id in "via" -- and dates each step with the cited option's lag (precursor date + p10..p90 days) or the anchor's date; only where no option or anchor applies, use your own well-founded expectation (via "none"). Date each branch where the causal evidence puts it, early or late. Also give, in "by", the probability that the target has happened by each of these dates (cumulative, so never decreasing), whatever the question's own deadline is.
Return JSON only:
{"branches": [
  {"date": "<YYYY-MM-DD on which the target happens (or is first reported to hold) in this branch, or null if it does not happen within 60 days>",
   "p": <probability of this branch; all branches sum to 1>,
   "via": ["<option / anchor ids this branch fires, or none>"],
   "path": ["<YYYY-MM-DD: speculative event, under 20 words>", "... at most 4 steps"]}
 ],
 "by": {"2026-07-28": <probability>, "2026-08-04": <probability>, "2026-08-11": <probability>},
 "rationale": "<two sentences>"}

PHOTOS AND CLIPS from the archive that match the target: [Image 1] Barry Hearn predicts Anthony Joshua to deliver 'very, very painful night' for Kristian Prenga (africa.espn.com, 2026-07-21) <photo> [...]

{"branches": [
  {"date": "2026-07-25", "p": 0.9, "via": ["O3", "O10", "O11", "O12", "A4", "A5", "E20", "E21", "E22", "E23", "E28", "E31"],
   "path": ["2026-07-17: Anthony Joshua is scheduled to take on Kristian Prenga on Saturday, July 25, 2026.", "2026-07-21: Anthony Joshua will return to the ring for the first time this year in Jeddah, Saudi Arabia on Saturday.", "2026-07-25: Anthony Joshua defeats Kristian Prenga."]},
  {"date": null, "p": 0.1, "via": ["O1", "O2", "E25", "E24"],
   "path": ["2026-07-17: Daniel Dubois brutally knocked out Anthony Joshua in September 2024.", "2026-07-17: Anthony Joshua dismantled Jake Paul inside six rounds and broke his jaw in two places.", "The fight between Anthony Joshua and Kristian Prenga is cancelled or postponed indefinitely."]}
 ],
 "by": {"2026-07-28": 0.9, "2026-08-04": 0.9, "2026-08-11": 0.9},
 "rationale": "The fight is scheduled for July 25, 2026, and multiple sources confirm this. Barry Hearn's prediction strongly suggests Joshua will win, and there's no indication of cancellation or a loss for Joshua in the provided data. The probability of Joshua winning is therefore very high."}

Figure 11: Prompt for deduction, $\mathsf{LM}(\mathbf{x}_q)$ in Equation (9), for *"Will Anthony Joshua defeat Kristian Prenga by 2026-07-28?"* (gold *Yes*), abridged ([...]: left-out lines), with the reply of Gemini 2.5 Flash-Lite below the line. Eq. (10) gives $p_q^{\mathrm{LM}} = \frac{1}{2}(0.9) + \frac{7}{42}(0.9) = 0.60$, $p_q^{\mathrm{arc}} = 0.98$ and $\hat{p}_q = 0.69 > \frac{1}{2}$: *Yes*, as on the other two backbones.

Table 4: How the prompt of Figure 11 changes for the other answer types; prompt text is set in the prompt's font.

Every type	TARGET is q itself, with its options for multiple choice. Next line: Today is 2026-07-21; the outcome is not yet known and must be forecast. Your answer must be {form}. Then the type's note below. Branch rule: Each branch gives the answer the question would have in it, in the required answer form. Date rule: when the question-relevant outcome occurs in this branch, or null. Each branch opens with "answer": "<the answer in this branch>", and "by" is dropped.
Choice	{form}: exactly one of the options, copied verbatim. Note: Each branch's answer must be one of the options, verbatim; option_probs gives every option's probability (summing to 1) and must agree with the branches. Before "rationale": "option_probs": {<every option, verbatim>: <probability>, ...},
Entity	{form}: the name only. Note: Prefer one of the candidate answers below when one fits the question; name someone else only if the timeline supports it better. answer_probs weighs every candidate you consider possible, so that a name the branches did not tell a story about can still be the answer. Before SOURCE EXCERPTS: CANDIDATE ANSWERS (entities of this storyline, most relevant first): and the twelve entities of $\mathcal{N}_q$ most relevant to q. Before "rationale": "answer_probs": {<every candidate answer you consider possible, verbatim, plus any name of your own>: <probability; all sum to 1>, ...},
Number	{form}: a single number (unit: {unit}). Note: Anchor on the most recent value of this quantity in the series, timeline or excerpts, on its trend, and on the scale the comparable quantities are reported at; each branch's answer is a bare number in the unit asked. Before SOURCE EXCERPTS: QUANTITY SERIES (entity / indicator: date: value): and up to six series of $\mathcal{V}_q$ (last five points each); COMPARABLE QUANTITIES (the same kind of number reported elsewhere in the archive; the scale a quantity of this kind is reported at): and $\mathcal{X}_q$; and the magnitude kernel μ_q as Reported values of this kind sit between $\{y_{(1/4)}\}$ and $\{y_{(3/4)}\}$, middle $\{y_{(1/2)}\}$. Put each branch's number where its own evidence puts it, but a number far outside that band has to be earned by the path.
Date	{form}: a single date in YYYY-MM-DD format.

B THEORETICAL PROOF

The three propositions of §4 are exact properties of Eqs. (4), (7) and (10). They hold for every item and every backbone, whatever the backbone replies, and they need no assumption on the data beyond the definitions in §4.

B.1 PROOF OF PROPOSITION 1

Proposition 1. *Fusion never lowers validity: a fused event is at least as valid as its reports.*

Proof. Validity depends on a hyperedge only through its support. For any non-empty $\mathcal{S} \subseteq \mathcal{K}$, with last date $\bar{t}_{\mathcal{S}} = \max_{r\in\mathcal{S}} t_r \le b$ and modalities $\Omega_{\mathcal{S}} = \{m(r) : r \in \mathcal{S}\}$, Eq. (4) reads

$$v(\mathcal{S}) = \min\{1,\ \big[g_{\mathrm{t}}(\bar{t}_{\mathcal{S}}) + g_{\mathrm{f}}(|\mathcal{S}|)\big]\, g_{\mathrm{m}}(|\Omega_{\mathcal{S}}|)\}, \tag{12}$$

with recency $g_{\mathrm{t}}(t) = (b-t+1)^{-1/2}$, recurrence $g_{\mathrm{f}}(f) = \lambda(f)f$ and modality $g_{\mathrm{m}}(n) = 1+\beta(n-1)$. Connected components do not depend on the order in which links are found, so $\mathcal{S}_e$, and with it $v_e = v(\mathcal{S}_e)$, is well defined. A report $\tilde{e}$ of r, were it left unfused, is the hyperedge with support $\{r\}$: $f = 1$, $|\Omega| = 1$, $\bar{t} = t_r$. We show the stronger claim that v is monotone under inclusion, $\mathcal{S} \subseteq \mathcal{S}' \Rightarrow v(\mathcal{S}) \le v(\mathcal{S}')$, factor by factor. *(i) Recency.* $\bar{t}_{\mathcal{S}'} \ge \bar{t}_{\mathcal{S}}$, and $g'_{\mathrm{t}}(t) = \frac{1}{2}(b-t+1)^{-3/2} > 0$ on $(-\infty, b]$, so $g_{\mathrm{t}}(\bar{t}_{\mathcal{S}'}) \ge g_{\mathrm{t}}(\bar{t}_{\mathcal{S}})$. *(ii) Recurrence.* For $f > 0$, $\sigma(\sqrt{f}/\eta) > \frac{1}{2}$, so $\lambda(f) > 0$, and

$$\lambda'(f) = \frac{\sigma(\sqrt{f}/\eta)\big(1-\sigma(\sqrt{f}/\eta)\big)}{2\eta\sqrt{f}} > 0, \qquad g'_{\mathrm{f}}(f) = \lambda(f) + f\lambda'(f) > 0, \tag{13}$$

so g_{f} is strictly increasing and $|\mathcal{S}'| \ge |\mathcal{S}|$ gives $g_{\mathrm{f}}(|\mathcal{S}'|) \ge g_{\mathrm{f}}(|\mathcal{S}|)$. *(iii) Modality.* $\Omega_{\mathcal{S}} \subseteq \Omega_{\mathcal{S}'}$ and $\beta > 0$ give $g_{\mathrm{m}}(|\Omega_{\mathcal{S}'}|) \ge g_{\mathrm{m}}(|\Omega_{\mathcal{S}}|) \ge 1$. Since $g_{\mathrm{t}}, g_{\mathrm{f}} > 0$, the bracket is positive; a product of positive non-decreasing terms is non-decreasing, and so is $x \mapsto \min\{1, x\}$. Hence $v(\mathcal{S}) \le v(\mathcal{S}')$. Taking $\mathcal{S} = \{r_i\}$ and $\mathcal{S}' = \mathcal{S}_e \ni r_i$ for each report $\tilde{e}_i$ fused into e,

$$v_e = v(\mathcal{S}_e) \ \ge\ v(\{r_i\}) = v_{\tilde{e}_i} \qquad \text{for every report } \tilde{e}_i \text{ of } e. \tag{14}$$

The inequality is strict whenever e has two or more attestations and $v_{\tilde{e}_i} < 1$: then $g_{\mathrm{f}}(|\mathcal{S}_e|) > g_{\mathrm{f}}(1)$ by (ii), the bracket grows strictly, and $\min\{1, x'\} > x$ for $x' > x$, $x < 1$. A second medium multiplies the uncapped value by $1 + \beta = 1.25$ on top of this, so cross-modal corroboration is rewarded beyond recurrence alone.

Consequence. Every later use of e is monotone in v_e: the ranking score $\rho_\phi v_e s_q(e)$ of an option in Eq. (8) and the score $\rho_{\mathcal{A}} v_e s_q(e)$ of an anchor. By Eq. (14) an event corroborated by further resources, or in a further medium, therefore ranks at least as high among the options and anchors of $\mathbf{x}_q$ as any of its single reports would. □

B.2 PROOF OF PROPOSITION 2

Proposition 2. *Chaining rules can only lower and delay a forecast, never raise or advance it.*

Proof. Let $\phi = \phi_1 \circ \phi_2$ with $\theta^{\mathrm{post}}_{\phi_1} = \theta^{\mathrm{pre}}_{\phi_2}$, and let $\ell_1 \sim \kappa_{\phi_1}$ and $\ell_2 \sim \kappa_{\phi_2}$ be independent lags, so that by Eq. (7) the chain's lag is $\ell_\phi = \ell_1 + \ell_2 \sim \kappa_\phi$ on $\{0, \dots, 2L\}$. We show that the chain's contribution at every horizon is no larger than either step's,

$$\rho_\phi\, K_\phi(h) \ \le\ \min\big\{\rho_{\phi_1} K_{\phi_1}(h),\ \rho_{\phi_2} K_{\phi_2}(h)\big\} \qquad \text{for every } h \ge 0. \tag{15}$$

Lower. In Eq. (6), $0 \le n^+_\phi \le n_\phi$ gives $\frac{1}{n_\phi+2} \le \rho_\phi \le \frac{n_\phi+1}{n_\phi+2}$, so every reliability lies strictly inside $(0, 1)$, and therefore $\rho_\phi = \rho_{\phi_1}\rho_{\phi_2} < \rho_{\phi_k}$ for $k = 1, 2$. *Later.* κ_{ϕ_1} and κ_{ϕ_2} are probability mass functions on non-negative lags, so exchanging the order of summation in Eq. (7) gives

$$K_\phi(h) = \sum_{\ell=0}^{h}\sum_{i=0}^{\ell} \kappa_{\phi_1}(i)\,\kappa_{\phi_2}(\ell - i) = \sum_{i=0}^{h} \kappa_{\phi_1}(i)\, K_{\phi_2}(h-i) \ \le\ \sum_{i=0}^{h} \kappa_{\phi_1}(i) = K_{\phi_1}(h), \tag{16}$$

since $K_{\phi_2} \le 1$; convolution is symmetric, so the same argument gives $K_\phi(h) \le K_{\phi_2}(h)$. Equivalently, $\Pr[\ell_1 + \ell_2 \le h] \le \Pr[\ell_k \le h]$ because $\ell_{3-k} \ge 0$: the chain's lag first-order stochastically

dominates each step's. Multiplying the two non-negative inequalities proves Eq. (15). Because $\rho_\phi < \rho_{\phi_k}$, it is moreover strict at every horizon h with $K_\phi(h) > 0$, i.e. wherever the chain can fire at all.

Dominance also orders every statistic of timing. For each quantile level $u \in (0, 1)$,

$$Q_\phi(u) = \min\{h : K_\phi(h) \geq u\} \;\geq\; \max_k Q_{\phi_k}(u), \qquad \mathbb{E}[\ell_\phi] = \mathbb{E}[\ell_1] + \mathbb{E}[\ell_2] > \max_k \mathbb{E}[\ell_k], \quad (17)$$

the first because $K_\phi \leq K_{\phi_k}$ pointwise, the second because each kernel puts positive mass on lags ≥ 1. In particular, the lag percentiles $p10/p50/p90$ that the context $\mathbf{x}_q$ shows for a chained option are never earlier than those of either of its steps.

Consequence. (a) The ranking score $\rho_\phi v_e s_q(e)$ in Eq. (8) is linear in ρ_ϕ, so from the same precursor a chain never outranks a single rule as reliable as either of its steps. (b) In Eq. (10) the factor $1 - \rho_\phi K_\phi(\Delta_q)$ of a chain exceeds each step's, so a chain never adds more to p_q^{arc} than either step would. (c) By induction on the number of steps, Eqs. (15)–(17) extend to chains of any length: each composition multiplies in one more factor $\rho < 1$ and convolves in one more non-negative lag. □

B.3 Proof of Proposition 3

Proposition 3. *For a fixed deduction, the probability of* Yes *never falls as the deadline moves later.*

Proof. Fix a yes/no item, its deduction $(\{\psi_j\}_{j=1}^J, C_q)$ and its grounding. The grounding (θ_q, $\mathcal{V}_q$, s_q, $\mathcal{O}_q$, $\mathcal{A}_q$) is computed from the target $\tilde{q}$, which omits d_q, so in Eq. (10) the deadline enters only through $\Delta = d - b$ and the anchor set $\mathcal{A}_q^d$. Every deadline lies in $[b, b{+}21]$ (§3); write $\hat{p}_q(d)$ for the read-out at deadline d and take $b \leq d \leq d' \leq b{+}21$.

(i) The backbone term. The running maximum $M_k = \max_{i \leq k} C_q(h_i)$ is non-decreasing in k and $M_1 \geq 0 = \bar{C}_q(0)$; linear interpolation between non-decreasing knots is a convex combination of them, so $\bar{C}_q$ is non-decreasing on $[0, 21]$ with values in $[0, 1]$. The second term $\frac{\Delta}{42}\bar{C}_q(21)$ is linear in Δ with slope $\bar{C}_q(21)/42 \geq 0$. Hence

$$p_q^{\mathrm{LM}}(d') - p_q^{\mathrm{LM}}(d) = \tfrac{1}{2}\big[\bar{C}_q(\Delta') - \bar{C}_q(\Delta)\big] + \tfrac{\Delta'-\Delta}{42}\,\bar{C}_q(21) \;\geq\; \tfrac{\Delta'-\Delta}{42}\,\bar{C}_q(21) \;\geq\; 0, \qquad (18)$$

and $0 \leq p_q^{\mathrm{LM}} \leq \frac{1}{2} + \frac{21}{42} = 1$. Without a curve, $p_q^{\mathrm{LM}}(d) = \sum_j p_j\, \mathbb{I}[t_j \in [b, d]]$ is non-decreasing because $[b, d] \subseteq [b, d']$.

(ii) The archive term. Write $p_q^{\mathrm{arc}}(d) = 1 - F(\Delta)\, G(d)$ with $F(\Delta) = \prod_{(e,\phi)\in\mathcal{O}_q}(1 - \rho_\phi K_\phi(\Delta))$ and $G(d) = \prod_{e\in\mathcal{A}_q^d}(1 - \rho_{\mathcal{A}} s_q(e))$. Every factor lies in $[0, 1]$, since $\rho_\phi, \rho_{\mathcal{A}} \in (0, 1)$, $K_\phi \in [0, 1]$ and $s_q \in [0, 1]$. Each K_ϕ is a cumulative sum of a non-negative kernel, so each factor of F is non-increasing in Δ; and $\mathcal{A}_q^d \subseteq \mathcal{A}_q^{d'}$, so a later deadline only adds factors of G, each at most 1:

$$F(\Delta')\, G(d') \;\leq\; F(\Delta)\, G(d') \;\leq\; F(\Delta)\, G(d), \qquad \text{i.e.} \qquad p_q^{\mathrm{arc}}(d') \;\geq\; p_q^{\mathrm{arc}}(d). \qquad (19)$$

Thus p_q^{arc} is non-decreasing in d; capping each term ρK or $\rho_{\mathcal{A}} s$ at a constant below 1, as a numerical guard, keeps every factor in $[0, 1]$ and non-increasing, so it changes nothing.

(iii) The read-out. $\hat{p}_q = (1-\alpha)p_q^{\mathrm{LM}} + \alpha\, p_q^{\mathrm{arc}}$ with $\alpha = \frac{1}{4}$ is a convex combination of non-decreasing functions of d, and the archive term is present at every deadline because $\mathcal{O}_q$ is non-empty for every item (§4.3). Combining Eqs. (18) and (19),

$$\hat{p}_q(d') - \hat{p}_q(d) \;\geq\; (1-\alpha)\,\frac{d'-d}{42}\,\bar{C}_q(21) \;\geq\; 0, \qquad \hat{p}_q(d) \in [0, 1], \qquad (20)$$

so $\hat{p}_q$ is non-decreasing in d, and strictly increasing whenever the backbone gives the event any probability of happening within the window.

Consequence. The set of deadlines answered *Yes*, $\{d : \hat{p}_q(d) > \frac{1}{2}\}$, is an up-set of $[b, b{+}21]$: there is a single crossing date $d_q^\star$, possibly an endpoint, with *Yes* for every $d > d_q^\star$ and *No* for every $d < d_q^\star$. Read from one deduction, questions about the same event that differ only in their deadline are answered consistently: the read-out never says the event will have happened by an earlier deadline but not by a later one. □

C FUTUREMAP ALGORITHM DETAILS

Algorithm 1 gives the full workflow of FUTUREMAP in the notation of §4, in three phases. **(1) Omnimodal Temporal Hypergraph.** Read parses each resource once in its own medium; reports of one event across media fuse into typed hyperedges weighted by cross-modal validity v_e. **(2) Known-Zone Deduction.** In code, type-pair hyper-rules are mined with first-passage time kernels, scored as forecasters at weekly pseudo-cutoffs and chained; anchors are backtested alike. **(3) Unknown-Zone Deduction.** Per item, the question is grounded in $\mathcal{G}$ without a model call, the backbone is asked once for dated branches that fire the grounded rules and anchors, and the typed answer is read off in code.

Algorithm 1 FUTUREMAP: Deduction over an Omnimodal Temporal Hypergraph

Require: Archive $\mathcal{K}$ closed at vantage point b, type ontology Θ, backbone LM, item q with answer type τ_q
Ensure: Answer $\hat{a}_q$

// 1: Omnimodal Temporal Hypergraph ▷ built once, shared by all backbones
$\big(\{\tilde{e}\}, \{\pi\}\big) \leftarrow \mathsf{Read}(r)$ for each $r \in \mathcal{K}$ ▷ ≤ 8 typed reports, ≤ 8 quantity points, Eq. 2
Link $\tilde{e} \leftrightarrow \tilde{e}'$ by type, shared participant, $|t_r - t_{r'}| \le \delta$ and $\cos(\mathbf{z}_{\tilde{e}}, \mathbf{z}_{\tilde{e}'}) \ge \gamma$ ▷ Eq. 3
for each connected component of $\leftrightarrow$ **do**
  Form $e = (\theta_e, \mathcal{V}_e, \mathcal{S}_e, t_e, t_e^{\mathrm{sch}})$, $\Omega_e \leftarrow \{m(r) : r \in \mathcal{S}_e\}$; $\mathcal{E} \leftarrow \mathcal{E} \cup \{e\}$
  $v_e \leftarrow \min\{1, [(b - \bar{t}_e + 1)^{-1/2} + \lambda(f_e) f_e](1 + \beta(|\Omega_e| - 1))\}$ ▷ Eq. 4
end for
Group points into series $Q_{u,k}$ by entity u and indicator k, ordered by t_π
// 2: Known-Zone Deduction ▷ mined once, in code, inside $\mathcal{K}$
$\mathcal{V}^\circ \leftarrow \mathcal{V}$ without hubs; for each $e \in \mathcal{E}$ and $\theta' \in \Theta$, record the first-passage lag $\ell_{(\theta_e \Rightarrow \theta')}(e)$ ▷ Eq. 5
Backtest at $b' \in \mathcal{B}'$: each $e \in \mathcal{F}_\phi(b')$ adds 1 to n_ϕ and $\mathbb{I}[\exists\, e' \bowtie e : \theta_{e'} = \theta_\phi^{\mathrm{post}},\ t_{e'} \in (b', b'+H]]$ to n_ϕ^+
$\rho_\phi \leftarrow (n_\phi^+ + 1)/(n_\phi + 2)$; $\Phi \leftarrow \{\phi : n_\phi \ge 4,\ \rho_\phi >$ base rate of $\theta_\phi^{\mathrm{post}}\}$ ▷ Eq. 6
$\kappa_\phi \leftarrow$ Gaussian-smoothed (ζ) lag density on $\{0, \dots, L\}$; $K_\phi(h) \leftarrow \sum_{\ell \le h} \kappa_\phi(\ell)$
$\Phi^{\le 2} \leftarrow \Phi \cup \{\phi_1 \circ \phi_2\}$ with $\rho_{\phi_1 \circ \phi_2} = \rho_{\phi_1}\rho_{\phi_2}$, $\kappa_{\phi_1 \circ \phi_2} = \kappa_{\phi_1} * \kappa_{\phi_2}$ ▷ Eq. 7
$\rho_{\mathcal{A}} \leftarrow$ backtested hit rate of scheduled hyperedges dated inside $\mathcal{K}$
// 3: Unknown-Zone Deduction ▷ per item, one backbone call
$\Delta_q \leftarrow d_q - b$; $\tilde{q} \leftarrow q$ without "by d_q" if τ_q is yes/no, else $\tilde{q} \leftarrow q$
Ground $\tilde{q}$: type θ_q, entities $\mathcal{V}_q \subseteq \mathcal{V}^\circ$, relevance $s_q(e) = \max\{0, \cos(\mathbf{z}_e, \mathbf{z}_{\tilde{q}})\}$
$\mathcal{O}_q \leftarrow \mathrm{top}_{12}$ precursor–rule pairs $(e, \phi) \in \mathcal{E} \times \Phi^{\le 2}$ into θ_q, ranked by $\rho_\phi v_e s_q(e)$ ▷ Eq. 8
$\mathcal{A}_q \leftarrow \mathrm{top}_5$ scheduled e with $t_e^{\mathrm{sch}} \ge b - 2$, $s_q(e) \ge 0.55$, ranked by $\rho_{\mathcal{A}} v_e s_q(e)$
$\mathbf{x}_q \leftarrow (\mathcal{V}_q,\ \mathcal{O}_q$ with ρ_ϕ and lag percentiles, $\mathcal{A}_q,\ \mathcal{N}_q,\ Q_{u,k}, \mathcal{X}_q, \mu_q$ if numeric, $\Xi_q)$
$\big(\{\psi_j = (p_j, t_j, y_j, \mathrm{via}_j)\}_{j=1}^{J},\ C_q\big) \leftarrow \mathsf{LM}(\mathbf{x}_q),\ \mathrm{via}_j \subseteq \mathcal{O}_q \cup \mathcal{A}_q$ ▷ Eq. 9
if τ_q is yes/no **then**
  $p_q^{\mathrm{LM}} \leftarrow \frac{1}{2}\bar{C}_q(\Delta_q) + \frac{\Delta_q}{42}\bar{C}_q(21)$ ▷ running max of C_q, Eq. 10
  $p_q^{\mathrm{arc}} \leftarrow 1 - \prod_{(e,\phi) \in \mathcal{O}_q}(1 - \rho_\phi K_\phi(\Delta_q)) \prod_{e \in \mathcal{A}_q^d}(1 - \rho_{\mathcal{A}} s_q(e))$
  return $\hat{a}_q \leftarrow$ *Yes* **if** $(1-\alpha)p_q^{\mathrm{LM}} + \alpha p_q^{\mathrm{arc}} > \frac{1}{2}$ **else** *No*
else if τ_q is numeric **then**
  return $\hat{a}_q \leftarrow \mathrm{wmed}(\{(y_j, p_j)\}_{j=1}^{J} \cup \{(y_q^{\mathrm{last}}, \omega)\})$ ▷ Eq. 11
else if τ_q is a date **then**
  return $\hat{a}_q \leftarrow \arg\max_t \sum_j p_j\, \mathbb{I}[|t_j - t| \le 3]$
else ▷ option or named entity
  return $\hat{a}_q \leftarrow \arg\max$ of the stated distribution over options or candidates, else of branch mass
end if

Complexity Analysis. First, construction makes one builder call per resource, $O(|\mathcal{K}|)$, and fusion compares only reports sharing a type and a participant within δ days, $O(Rd)$ for R reports of dimension d. Second, mining scans each participant's time-sorted hyperedges within L days, the backtest costs $O(|\mathcal{B}'|\,|\mathcal{E}|\,|\Theta|)$, and a chained kernel is an $O(L^2)$ convolution. Third, per item, grounding is $O(|\mathcal{E}|d)$ retrieval plus a scan of the recent precursors of $\mathcal{V}_q$ against $\Phi^{\le 2}$, then one backbone call and a read-out linear in $J \le 4$ branches. Phases (1)–(2) run once, shared by every backbone and item, so a query pays for one grounded deduction (§5.7).

Hyperparameters. One setting serves all backbones and domains: fusion δ=3, γ=0.86; validity η=2, β=0.25; hub size 400; kernels L=42, ζ=2; backtest H=14, $n_\phi \ge$4; per item, 12 options, 5 anchors ($s_q \ge$0.55), $\le$50 timeline hyperedges, 2–4 branches, $h \in \{7, 14, 21\}$, α=$\frac{1}{4}$, ω=0.75, a $\pm$3-day date window.

D OMNIFP-BENCH DETAIL

This appendix expands §3 into a self-contained reference for OMNIFP-BENCH v4.1: the task each item poses (§D.1), the pipeline that turns a frozen crawl into released items (§D.2), and the scorer that grades an answer (§D.3). Throughout, $\mathcal{M} = \{\mathrm{T}, \mathrm{I}, \mathrm{D}, \mathrm{V}, \mathrm{A}\}$ is the modality alphabet, $\tau = (a, b, c) =$ (2026-06-09, 2026-07-21, 2026-08-11) the protocol, $\mathcal{G} = \{b{+}7, b{+}14, b{+}21\}$ the horizon grid, and κ_q the answer type of item q.

D.1 TASK DESIGN

Instance and interface. An item is $q = (\mathrm{text}_q, \kappa_q, d_q, e_q, y_q^*, \mathcal{P}_q, \mathcal{S}_q)$: a question, its answer type, a deadline $d_q \in \mathcal{G}$, the event e_q it asks about, dated $t_e \in (b, c]$, the gold answer, its precursors and its settling resources. A system is a map

$$F : \big(\mathrm{text}_q,\ \mathrm{contract}(\kappa_q),\ \mathcal{K}\big) \longmapsto \hat{y}_q, \qquad \mathcal{K} = \{r \in \mathcal{R} : a \le t_r \le b\}, \tag{21}$$

which sees the question, its answer form (§D.3) and the known zone in all five modalities, and nothing else: $\mathcal{S}_q$, e_q, t_e and all other fields are withheld, nothing is searched or fetched at run time, and one F answers every item.

Resolution. The gold is fixed by the unknown zone alone. Every item is settled by $|\mathcal{S}_q| \ge 2$ resources in $\mathcal{U} = \{r : b < t_r \le c\}$ (2–7, mean 3.4); its settling modality $m^*(q) \in \mathcal{M}$ is that of the resource the gold is read from; and its deadline is the first grid point on or after the event, the question always being posed at b:

$$d_q = \min\{g \in \mathcal{G} : g \ge t_e\}, \qquad y_q^* = \mathrm{read}\big(\mathcal{S}_q\big), \quad \mathcal{S}_q \subseteq \mathcal{U}, \quad b < t_e \le d_q \le c. \tag{22}$$

A question names its deadline ("… by 2026-08-04?") but never the event's date; the three deadlines hold 567, 626 and 337 items.

Forecastability constraints. Two conditions, enforced against the frozen resources (§D.2), separate a forecast from a lookup and from a guess:

$$\underbrace{|\mathcal{P}_q| \ge 2,\ \ \mathcal{P}_q \subseteq \mathcal{K},\ \ \mathrm{signal}(r, e_q) = 2\ \ \forall r \in \mathcal{P}_q}_{\text{precursor}} \qquad \text{and} \qquad \underbrace{\neg\exists\, r \in \mathcal{K} :\ \mathrm{states}(r, y_q^*)}_{\text{no leak}}, \tag{23}$$

where $\mathrm{signal} \in \{0, 1, 2\}$ grades a resource as unrelated, topical, or carrying a signal a careful forecaster could reason from. Released items cite 3.6 precursors on average (2–11).

Answer types. Table 5 lists the seven types. A binary asks whether an event occurs by d_q; a numeric threshold is a binary about a quantity; a perturbed binary is *No* by construction. 117 of the 119 perturbed binaries are *negative twins*: for a released *Yes* binary q whose event falls after the first grid point, the twin q^- re-asks the identical question at the last grid point before the event,

$$d_{q^-} = \max\{g \in \mathcal{G} : g < t_e\}, \qquad y_{q^-}^* = \mathrm{No}, \tag{24}$$

once a corpus check confirms that the event did not already occur inside $(b, d_{q^-}]$; the other two swap one quantity of a true claim. A twin differs from its original in the deadline alone, so answering both correctly requires placing the event *in time*, not merely judging it plausible.

Settling modality and chains. $m^*(q)$ is earned by refutation: a non-text item survives only if the construction model, shown the event's articles and the resource's caption or title, *fails* it, and a clip's question must also defeat a single middle frame. A storyline is a component of a domain's events linked by a shared process or two shared actors, and a chain when it bears on a second domain: 951 items lie on 379 chains, with up to six linked domains per item.

D.2 BENCHMARK CONSTRUCTION

Design principles. *(i)* One pinned construction model, `gpt-4o-mini` (2024-07-18), of no evaluation backbone's family. *(ii)* Judging under opposed incentives (VERIFY against ATTACK), a claim counting only if its quoted span is in the named resource. *(iii)* A rule-based audit over the frozen files. *(iv)* Fail closed: no gate is relaxed to fill a domain.

Table 5: **Answer types**, each with one released item: domain, settling modality m^* and deadline d_q. The twin re-asks the binary above it; both are posed at $b =$ 2026-07-21.

Type	n	Space	Example question	Dom./m^*/d_q	Gold
Numeric value	586	$\mathbb{R}$, unit	What will Airbnb, Inc.'s Q2 2026 revenue be by 2026-08-11?	Bus./D/$b+21$	3600000000 (USD)
Multiple choice	314	4 options	What action will Nikol Pashinyan take regarding Russian restrictions on Armenian exports by 2026-07-28? {Seek assistance from the EU; Ignore the restrictions; Ask Vladimir Putin for help; Impose counter-restrictions}	Geo./T/$b+7$	Ask Vladimir Putin for help
Named entity	253	free text	Which authorities will order the evacuation of the Cap Ferret peninsula by 2026-07-28?	Env./T/$b+7$	French authorities
Binary	211	Yes/No	Will Mitsubishi Electric announce plans to build facilities for the Global Combat Air Programme by 2026-08-11?	Bus./T/$b+21$	Yes
Perturbed binary	119	Yes/No	… the same question by **2026-08-04**? (negative twin; the event is on 2026-08-06)	Bus./T/$b+14$	No
Date	44	day	On what date will SpaceX launch its 13th Super Heavy-Starship rocket, by 2026-07-28?	Tech./T/$b+7$	July 25, 2026
Numeric threshold	3	Yes/No	Will SK Hynix report an operating profit of at least 60 trillion won by 2026-08-04?	Bus./T/$b+14$	Yes

Table 6: **Construction funnel**, read from the build log. *In*/*out* count event cards for S1–S2 and items after; S3 also draws on 4,168 record and media candidates.

Stage	Gate	In	Out	Largest rejection causes
S1 Events	2 indep. sources	19,873	6,855	$<$2 independent sources (6,683); quote unfound (1,542); event in $\mathcal{K}$ (1,408)
S2 Precursor, leak	Eq. (23)	6,855	2,240	topical evidence only (3,900); outcome stated in $\mathcal{K}$ (507); one precursor (142)
S3/S3d/S3m	typed, refuted	2,240	4,899	illegible or incidental media detail (2,297); off-event (563); answerable without the medium (499)
S4 Judge	verify $\wedge$ $\neg$attack	4,671	3,819	unfaithful to its source (492); attack proves the gold wrong (356)
S5 Audit	deterministic	4,047	2,988	quote lacks the gold (624); quote not in its resource (421); gold in $\mathcal{K}$ (263)
S5c Re-leak	grounded quote	2,988	2,589	answer stated in $\mathcal{K}$ (297), and its twin (102)
S6–S7 Release	quota, grid	reserve	1,530	domain cap of 256; twins left without a valid grid deadline (50)

S0 Acquisition. Text: 5,620 articles from 409 outlets in a frozen CC-NEWS crawl, dated by crawl time, syndicated copies sharing one story id. Image: 662 photographs embedded in those articles. Document: 650 primary records from 311 issuers (Federal Register, federal appellate and Supreme Court opinions, 8-K exhibits, WHO outbreak notices, arXiv, governing bodies), kept as the issuer's PDF or page. Audio: 263 episodes of 49 shows from 52 podcast feeds, cut to at most 20 minutes (mean 16) and transcribed locally. Video: 227 one-minute clips from five TV News Archive channels. Media occupy 11 GB.

S1–S7. Table 6 traces the pipeline. **S1** clusters unknown-zone articles on shared rare terms around a tier-1 report, and anchors further events on a primary document or clip that a news report independently confirms. **S2** searches $\mathcal{K}$ lexically in all five modalities and densely for articles and images, and applies Eq. (23). **S3** writes up to three questions of distinct types per event under a rotating type plan, with the deadline of Eq. (22) and a gold quoted from $\mathcal{S}_q$; **S3d** and **S3m** add document-, recording-, photograph- and clip-settled items by refutation. **S4** runs VERIFY over six criteria (faithful, verifiable, temporal, no leak, matcher-gradable, well-posed), ATTACK, and a closed-book PROBE that records, but never rejects, the items a model answers from memory. **S5** checks the date order of Eq. (22), that the quote occurs in its resource and carries the gold, that the gold is absent from the event's known-zone resources, two registered domains per event and no near-duplicates, then adds the twins of Eq. (24). **S5c** re-searches all of $\mathcal{K}$ in each item's own wording and drops it, with its twin, when a grounded quote states the answer. **S6** fills each domain up to 256 items, preferring non-text modalities, chains and storylines under eight items, and builds the pool; **S7** shuffles the options, retypes numeric golds and places every deadline on $\mathcal{G}$.

Release and verification. An item record adds to its question the gold, type, unit, m^*, deadline, event, storyline, chain links, precursor ids and a `resolution_spec` (settling ids, quote); a pool record holds its content, date, source, media path and role. A verifier re-checks from the shipped files alone, with no failure on v4.1, that every resource lies in its zone, every item has two or more precursors in the known pool and none in the unknown, every quote occurs in its cited unknown-zone resource, every image or clip item has its media on disk, every twin has its parent, and every date obeys Eq. (22), or Eq. (24) for a twin; a manifest pins every released file by its content hash, and the build log counts each gate's rejections by cause (Table 6).

Table 7: **Released items by domain**: the answer types of Table 5 in order, settling modality, deadline, and items on a cross-domain chain.

		Answer type							Settling modality					Deadline			
Domain	**Items**	Num	MC	Ent	Bin	Pert	Date	Thr	T	I	D	A	V	$b{+}7$	$b{+}14$	$b{+}21$	**Chain**
Politics & Gov.	250	87	49	52	33	18	10	1	197	29	7	7	10	92	100	58	194
Geopolitics	245	95	47	37	39	20	7	0	205	21	5	9	5	96	90	59	177
Law & Crime	132	52	16	29	21	11	3	0	103	14	6	5	4	48	61	23	87
Environment	169	70	45	23	20	11	0	0	149	7	12	0	1	65	63	41	90
Business	248	102	49	43	28	16	8	2	174	45	16	10	3	83	114	51	156
Technology	107	46	17	14	16	9	5	0	85	11	3	5	3	47	39	21	90
Health	131	41	31	19	20	15	5	0	114	6	9	1	1	42	52	37	52
Sports	248	93	60	36	34	19	6	0	188	54	6	0	0	94	107	47	105
All	1,530	586	314	253	211	119	44	3	1,215	187	64	37	27	567	626	337	951

D.3 Evaluation Metric Design

Answer contract. Every system receives the question (options listed beneath it) and one wrapper: "*Today is 2026-07-21; the outcome is not yet known and must be forecast. Your answer must be* {*form*}. *End your response with a final line of the form: Final answer:* ⟨*answer*⟩", where {form} is "Yes or No" for the three yes/no types, "exactly one of the options, copied verbatim", "a single number" (plus the unit when the item has one), "a single date in YYYY-MM-DD format", or "the name only". Neither the grid nor the existence of twins is disclosed.

Extraction and normalisation. The scorer takes $\hat{y} = \epsilon(o)$ from the reply o: the text after the last "Final answer:" (or a `\boxed{}` answer, or else the last line), stripped of quotes, markdown and a leading "the answer is". A yes/no reply keeps its first *yes* or *no*; a multiple-choice reply maps onto an option by a bare letter, an exact option, the unique option it contains, or else highest token F1. Strings are compared under $\nu(\cdot)$: lower-case, strip punctuation and *a/an/the*, collapse whitespace.

Typed matching. With $\pi(\cdot)$ reading a number through thousands separators, dollar and per-cent signs and k/M/B/T suffixes, and $\delta(\cdot)$ a date in any of ten common formats, the match $\mu_\kappa(\hat{y}, y^*) \in \{0, 1\}$ is

$$\mu_\kappa(\hat{y}, y^*) = \begin{cases} \mathbb{1}\big[\mathrm{yn}(\hat{y}) = \mathrm{yn}(y^*)\big] & \kappa \text{ yes/no}, \\ \mathbb{1}\big[\nu(\hat{y}) = \nu(y^*)\big] & \kappa = \text{multiple choice}, \\ \mathbb{1}\big[\,|\pi(\hat{y}) - \pi(y^*)| \le 0.15\,|\pi(y^*)|\,\big] & \kappa = \text{numeric value}, \\ \mathbb{1}\big[\,|\delta(\hat{y}) - \delta(y^*)| \le 3 \text{ days}\,\big] & \kappa = \text{date}, \\ \mathbb{1}\big[\hat{T} \subseteq T^* \;\vee\; T^* \subseteq \hat{T} \;\vee\; \mathrm{F1}(\hat{T}, T^*) \ge 0.6\big] & \kappa = \text{named entity}, \end{cases} \tag{25}$$

where $\hat{T}, T^*$ are the token sets of $\nu(\hat{y}), \nu(y^*)$ and $\mathrm{F1} = 2|\hat{T} \cap T^*|/(|\hat{T}| + |T^*|)$. A numeric gold of 0 requires $\pi(\hat{y}) = 0$; when π or δ cannot read a side, the match falls back to $\nu(\hat{y}) = \nu(y^*)$; an empty reply scores 0. The tolerances are fixed per type in the scorer, committed before any system was run, and equal for every item of a type. No model is called, so a grade replays bit-for-bit.

Accuracy and floors. For a set Q (a domain, or all 1,530 items for **Overall**), $\mathrm{Acc}(Q) = \frac{100}{|Q|}\sum_{q\in Q}\mu_{\kappa_q}(\hat{y}_q, y^*_q)$, a micro-average in which every item weighs the same and a refused or unparseable reply counts as wrong. The *parametric* floor is each block's *No source* row, the same backbone and contract with $\mathcal{K}$ withheld. The *trivial* floor answers *Yes* to every yes/no item and picks a multiple-choice option uniformly:

$$\mathrm{Acc}_{\text{triv}}(Q) = \frac{100}{|Q|}\Big(\,\big|\{q \in Q_{\text{yn}} : y^*_q = \text{Yes}\}\big| + \tfrac{1}{4}\,|Q_{\text{mc}}|\Big), \tag{26}$$

which scores 59.8% on the 333 yes/no items (199 Yes, 134 No), 25% on multiple choice and 0 on the open types: 18.1% Overall, and 17.7, 20.3, 16.7, 18.5, 15.8, 16.1, 20.4 and 19.0% from Politics to Sports. Always-*Yes* scores 0 on all 117 twins, so the yes/no block rewards timing an event.

Significance and strata. Rows are single-seed at temperature 0, with binomial standard error $\sqrt{p(1-p)/N}$ ($\approx \pm 1.2$ Overall). Two rows on one backbone are compared item-paired, by McNemar's $z = (n_{10} - n_{01})/\sqrt{n_{10} + n_{01}}$ and a 95% percentile bootstrap over 4,000 item resamples (seed 0). Strata are fixed before any run: deadline, chain versus single-domain, and the 1,249 items the S4 probe did *not* answer from memory, where a margin is forecasting, not recall.

E BASELINE DETAIL

This appendix expands §5.1 into a per-method specification of every row of Table 1 and Table 2. For each of the twelve baselines we give (i) *Overview*, the method and the store it builds over $\mathcal{K}$; (ii) *Implementation*, what the port keeps verbatim from the released code and what the benchmark fixes; and (iii) *Hyperparameters*, the values used in our runs. Table 8 lists the repository and commit each port starts from. Throughout, Search_k is the shared pool index returning k hits, LLM_β the row's backbone, $\mathrm{C}(q)$ the question with its typed answer contract and final-answer line, and M_q the three stills and one clip the index ranks highest for it.

Shared harness. `google/gemini-2.5-flash-lite`, `qwen/qwen3.5-flash-02-23` (reasoning disabled) and `mistralai/mistral-nemo` are served through OpenRouter with the provider pinned and fallbacks off; every call runs at temperature 0 with seed 20260615 and is disk-cached and metered, and a Nemo prompt stays under 300k characters. Every row searches $\mathcal{K}$ through one index: BM25 (k_1=1.5, b=0.75) over each resource's text view, fused by reciprocal rank (k=60) with 3,072-d `gemini-embedding-2` vectors under a quota per modality. A store is built once with Gemini 2.5 Flash Lite over the released pool and shared by all backbones. Copied code is kept verbatim with every edited line marked; web search becomes Search_k, the original model LLM_β, and the answering call carries $\mathrm{C}(q)$ and M_q (natively, or as OCR/ASR text on Nemo), so every baseline has the same media access; no ported prompt is tuned on test items.

Table 8: **Every row of Table 1 at a glance.** Code: the upstream repository and commit the port starts from. Store and query: what the method builds once over $\mathcal{K}$, and what it does per item.

Method	Code	Store over $\mathcal{K}$ / query-time procedure
No source	—	none / one call on $\mathrm{C}(q)$, no evidence
REACT	`ysymyth/ReAct@6bdb3a1`	none / Thought–Action loop over Search, Lookup, Finish
TKG-ICL	`usc-isi-i2/isi-tkg-icl@951c969`	shared temporal KG / history of the linked entity in context
EXPEL	`LeapLabTHU/ExpeL@e41ec9a`	insights of 150 training items / ReAct with recalled trajectories
HYPERGRAPHRAG	`LHRLAB/HyperGraphRAG@d587cdf`	text hypergraph / entity and hyperedge keys, one-hop expansion
MEMOTIME	`SteveTANTAN/MemoTime@84ed588`	shared temporal KG / Tree-of-Time decomposition, nine toolkits
ATOM	`AuvaLab/itext2kg@840aba2`	temporal KG of atomic facts / top quintuples before the cutoff
FUTUREOMNI	`OpenMOSS/FutureOmni@5b292a0`	none / ten text resources and M_q inline, one call
MIROFLOW	`MiromindAI/MiroFlow@9d180f1`	none / MCP search and read turns with rollback, boxed summary
TIMESEEK	`coys17/timeseek-anonymous@f27050f`	none / research, agent, tools and forecast nodes
mini-prophet	`ai-prophet/mini-prophet@fdf121b`	source board / one tool per step, a probability per outcome
MMHRAG	`ShenAoChen2001/MMHRAG@1c87847`	image–text hypergraph / low- and high-order keys, refined edges
LEAP	`layingfish/LEAP@3677fbb`	snapshot / blind prior, isolated likelihoods, Bayesian update
FUTUREMAP	ours	omnimodal hypergraph, rules / one deadline-blind deduction

E.1 AGENTIC SEARCH FORECASTERS

REACT (Yao et al., 2023). (i) Reasoning interleaved with acting: thoughts and tool actions alternate until an answer is committed; no store. (ii) The HotpotQA instruction, the six `webthink_simple6` few-shot trajectories and the `webthink` loop are kept; Wikipedia becomes Search_8 over articles, documents and transcripts, the best hit opened as the page; after Finish$[\cdot]$ one call reads the trajectory with $\mathrm{C}(q)$ and M_q. (iii) 7 steps, 5 searches, 256 tokens per step, 1,024 for the answer.

EXPEL (Zhao et al., 2024). (i) An experiential learner: a ReAct agent gathers trajectories on training tasks, distils insights from them, and recalls both at test time. (ii) Gathering, Reflexion and insight extraction (AGREE/REMOVE/EDIT/ADD, $\leq$20 rules) are kept, on 150 domain-stratified items of an earlier, disjoint release; search is Search_5 with 250-character extracts. (iii) 7 steps, 5 searches, one Reflexion retry, 3 recalled trajectories.

MIROFLOW (Su et al., 2026). (i) A deep-research agent: one tool call per turn under an MCP contract, rollback of malformed or duplicate turns, a summary ending in `\boxed{}`. (ii) The prompts and the rollback loop are kept; the web tools become `search_corpus` (top-8, 700-character snippets) and `read_resource` (6,000 characters). M_q joins the summary call. (iii) 30 turns, 5 consecutive and 3 duplicate rollbacks, 1,536 tokens per turn, 2,048 for the summary.

TIMESEEK (Mostafa et al., 2026). (i) An agentic forecaster run as a state machine that ends in a structured prediction with a confidence. (ii) Prompts, routing and answer extraction are kept,

the market price off as by default; SerpAPI becomes Search_8 with the best-matching 320-character window; the forecast node always issues the final prompt, so M_q reaches it. (iii) 10 turns, 5 searches, 4,096 tokens per call.

mini-prophet (Murphy, 2026). (i) A forecasting agent that keeps a board of annotated sources and submits a probability per outcome. (ii) Release v0.1.10 is kept: one tool per step, a 10-message window, a three-turn grace period; search is Search_5 with 1,500-character snippets; the answer is the most probable outcome, and one more turn shows M_q before the final submit. (iii) 12 steps, 5 searches.

LEAP (Chen et al., 2026d). (i) Likelihood elicitation: a blind prior, one isolated likelihood per evidence item, and deterministic Bayesian aggregation over a frozen, hashed snapshot. (ii) The snapshot code is kept and the stages follow the paper's Algorithm 1, with dependency clustering and η=1; M_q is one more evidence item. (iii) 6 rounds, 5 searches, 1 fetch, $\leq$6 items.

E.2 Graph and Hypergraph Forecasters

ATOM (Lairgi et al., 2026a). (i) A dynamic temporal KG: resources are split into atomic facts, facts into quintuples $(s, p, o, t_{\text{start}}, t_{\text{end}})$ observed at publication, and snapshots merged in time order. (ii) Both extraction modules and the merge thresholds (entity 0.8, relation 0.7) are kept, giving 157k atomic facts, 39k entities and 194k quintuples; the reader gets the top-60 quintuples dated before the cutoff. TKG-ICL and MemoTime share this graph. (iii) Chunks of $\leq$400 tokens, K=60.

TKG-ICL (Lee et al., 2023). (i) Temporal KG forecasting by in-context learning over the query entity's history. (ii) `construct_history_by_search` is kept unchanged; the entity is linked from the question, and the relation stub of the query is replaced by $\mathrm{C}(q)$. (iii) History 100, entity-level, unidirectional, 2 linked entities.

MemoTime (Tan et al., 2026). (i) Memory-augmented temporal KG reasoning: Tree-of-Time decomposition, operator-aware toolkits, pruning and a sufficiency test per sub-question. (ii) All prompts and the nine toolkits are kept, and experience is written only on MemoTime's own sufficiency verdict. (iii) 4 sub-questions, 2 toolkits, 2 regenerations, memory top-10.

HyperGraphRAG (Luo et al., 2025). (i) Retrieval over n-ary facts, one hyperedge per fact. (ii) Extraction with two gleaning rounds, merge rules and hybrid query are kept (151k entity and 81k hyperedge nodes); the generation prompt of its evaluation pipeline answers with M_q appended. (iii) Chunks of 1,200 tokens, top-60, 4,000-token caps.

MMHRAG (Chen et al., 2026b). (i) Multimodal hypergraph RAG over text and figure units, refined around anchor vertices. (ii) Extraction, refinement and `hyper_query` are kept; the 470 stills are image units (43k vertices, 112k hyperedges). (iii) Top-60, anchors on $\geq$3 units.

FutureOmni (Chen et al., 2026a). (i) Passive omnimodal perception in one call. (ii) The Gemini prompt of its inference script is kept, its nouns widened to the evidence served. (iii) Seven articles, two documents and one transcript of 2,000 characters each, three stills and one clip, answered in one call of 1,024 output tokens.

E.3 FutureMap and Its Ablations

FutureMap. (i) Stage 1 reads each resource once in its own medium into 23,097 event reports fused into 20,220 typed hyperedges; stage 2 mines and backtests hyper-rules, kernels and anchors in code; both are built once and shared by every backbone. (ii) Grounding needs no model call, and one deadline-blind deduction reads the grounded blocks, excerpts and media; the typed answer is read off in code, a yes/no from the horizon curve blended with stage 2's first-passage probability (w=0.25) and a number as the weighted median of the branch values and the persistence branch (w=0.75) (App. C). (iii) 12 options, 5 anchors, $\leq$50 timeline hyperedges, seven 1,200-character excerpts, two stills and one clip, 1,600 output tokens. **Ablations.** Each row of Table 2 sets one switch and leaves the rest of the run unchanged. *Causal hyper-rules* withholds $\mathcal{O}_q$ and the anchors; *magnitude kernels* the comparable quantities; *horizon curve* reads a yes/no from branch dates alone; *first-passage vote* drops stage 2's probability; *persistence* drops the series' last value; *hypergraph* keeps only the source excerpts, in a single retrieve-then-deduce call over the same backbone and answer contract.

F EXPERIMENT DETAIL

This appendix expands §5: for each research question it states the protocol behind the main text and the per-backbone values its figures summarise. Unless named otherwise, every number is recomputed from the cached runs of Table 1.

F.1 SHARED PROTOCOL

Every row answers all 1,530 items once per backbone at temperature 0, seed 20260615, with every call cached (Appendix E); only §F.3 and §F.5 add runs, and the typed matcher of Appendix D.3 grades every answer. A paired interval resamples the items 4,000 times (seed 0) and takes the 2.5th and 97.5th percentiles of the difference of two rows on one backbone; McNemar's $z = (n_{10} - n_{01})/\sqrt{n_{10} + n_{01}}$ counts the items only one row answers. The fixed item sets are the 333 yes/no items (199 *Yes*), 117 twin pairs, 586 numeric items (500 with a positive gold), seeded random draws of 500 (thinning) and 400 (cost) items, and the strata.

F.2 MAIN RESULTS AND ABLATIONS (RQ1, RQ2)

Table 9: **Table 1 per domain, mean of the three backbones.** Best baseline: the strongest of the twelve in each column (LEAP, except ExpeL on Tech and ATOM on Hlth).

	Pol	Geo	Law	Env	Biz	Tech	Hlth	Spt	Overall
FUTUREMAP	**30.7**	**34.0**	**26.8**	**25.9**	**27.0**	**33.3**	**36.7**	**37.4**	**31.5**
Best baseline	24.4	24.5	24.0	21.9	25.0	31.8	33.6	31.9	26.4
No source	21.3	20.5	18.2	20.1	20.3	23.4	26.7	21.9	21.3
Margin over the best	+6.3	+9.5	+2.8	+4.0	+2.0	+1.5	+3.1	+5.5	+5.1

RQ1. Averaged over the backbones (Table 9), FUTUREMAP leads the best baseline of every domain. On single backbones it is first or tied-first in 6 (Gemini), 5 (Qwen) and 7 (Nemo) of the eight domains, and above *No source* in all 24 cells by at least +4.6 points, a floor no baseline clears in every cell. **RQ2.** Each ablation of Table 2 reruns all items on each backbone with one switch set (Appendix E) and every store, cache and hyperparameter shared, so a difference isolates one component; none raises **Overall** on any backbone. Removing stage 1 leaves the same source excerpts without hyperedges, rules or anchors.

F.3 OMNIMODAL TEMPORAL HYPERGRAPH (RQ3)

Stage 1 reads the 3,531 resources of $\mathcal{K}$ into 23,097 event reports and fuses them into 20,220 hyperedges over 23,965 entities (Table 10); a hyperedge is *fused* when its reports span two or more media and *repeated* when two or more resources report it. Coverage is the share of an item's gold precursors, never read by the method, that support its (Gemini) prompt, against the top-k hits of the shared index. The thinned stores keep a nested, hash-seeded part of $\mathcal{K}$, rebuild stage 1 and re-mine stage 2; the best baseline uses the full archive on the same 500 items.

Table 10: **Stage 1 in numbers.** Left: the store; validity grows with media and with reports, as Proposition 1 implies. Right: gold-precursor coverage (all items) and accuracy from part of $\mathcal{K}$ (500 items, mean of the backbones).

Share of hyperedges	**Store**	**Options**	**Anchors**
Fused	1.1%	11.4%	8.5%
Repeated	6.3%	34.7%	27.3%
Single-medium: 15,709 T, 2,238 D, 988 A, 658 I, 414 V			
Fused: 105 I+T, 44 D+T, 41 A+T, 8 T+V, 7 A+I+T, 3 A+D+T, 3 A+T+V, 2 D+I+T			
Mean v_e by media, 1 / 2 / 3		0.43 / 0.85 / 1.00	
Mean v_e by reports, 1 / 2 / 3 / ≥4		0.42 / 0.66 / 0.91 / 1.00	

Coverage	**Chars**	**Covered**
Flat top-3 / 5	7.8k / 12.7k	48.4 / 61.2%
Flat top-10 / 20	25.2k / 50.6k	72.8 / 82.4%
FUTUREMAP	24.8k	**84.9%**
Share of $\mathcal{K}$	**FUTUREMAP**	**Best base.**
One third	30.3	
Two thirds	30.9	28.2
All	32.8	

F.4 Backtested Hyper-rules and Kernels (RQ4)

We rerun the miner of §4.2 on the archive cut at $T_0 = b-14$ and $b-7$. Each event of the week before T_0 forecasts the consequences of the rules on its type, a hit being such an event sharing a participant in $(T_0, T_0+14]$; ρ_ϕ is compared with the co-occurrence confidence $P(\theta^{\mathrm{post}} \mid \theta^{\mathrm{pre}})$ and the base rate (ECE, ten bins), and each realised first lag is mapped through κ_ϕ on the window (randomised PIT).

Table 11: **Stage 2 out of sample.** Left: calibration at two earlier cutoffs. Right: numeric items within a tolerance t, $|\log(\hat{y}/y)| \le \log(1+t)$, 500 items, mean of the backbones; best baseline per column.

Cutoff T_0	$b-14$	$b-7$
Rules / forecasts	838 / 85,894	863 / 96,533
Observed hit rate	10.7%	7.9%
ECE, backtested ρ_ϕ	**0.014**	**0.017**
ECE, co-occurrence	0.023	0.019
ECE, base rate	0.093	0.065
Realised lags	9,192	7,639
KS, time kernel κ_ϕ	**0.074**	**0.109**
KS, uniform lag	0.203	0.181

t (%)	5	10	15	25	50	75	100
FutureMap	**11.3**	**13.9**	**16.3**	**21.0**	**30.5**	**36.9**	**44.3**
− magnitude kernel	9.3	12.5	15.0	19.7	28.9	34.9	42.3
− persistence	9.9	13.0	14.4	19.1	28.5	35.1	42.9
Best baseline	9.5	11.2	12.7	16.2	22.0	25.8	31.3

At both cutoffs ρ_ϕ has the lowest calibration error and the time kernels the lowest KS distance, so the stage-2 quantities that enter Eq. (10) as probabilities hold on weeks the miner never saw. On numeric items, FutureMap leads the best baseline at every tolerance by $+1.8$ to $+13.0$ points, and removing either the magnitude kernel or the persistence branch lowers the curve at every tolerance.

F.5 Deadline-Blind Branching and Read-out (RQ5)

A pair has an *identical deduction* when parent and twin cite the same rule options and read the same horizon curve (the case of Proposition 3); read-outs are compared over the 117, 117 and 112 pairs where both runs return one. Balanced accuracy averages the true-positive and true-negative rates; the deadline-shown run only keeps the deadline in the target.

Table 12: **Stage 3 per backbone**: twin pairs and yes/no items, the deadline shown against withheld, and the numeric read-out (586 items).

	Gemini	**Qwen**	**Nemo**
Parent / twin accuracy	86.3 / 38.5	90.6 / 41.0	94.9 / 30.8
Both right: FutureMap / best baseline	**25.6** / 13.7	**34.2** / 15.4	**25.6** / 12.8
Balanced accuracy: FutureMap / best other row	**59.9** / 53.1	**59.3** / 52.8	**58.5** / 51.6
Pairs with $\hat{p}_q$(parent) $>$ $\hat{p}_q$(twin)	93.2%	95.7%	97.3%
Identical deductions, parent never lower	65 of 65	58 of 58	42 of 42
Deadline shown → withheld: dated branches on it	47.9 → 17.5%	26.8 → 24.2%	30.9 → 12.3%
twin accuracy	30.8 → **38.5**	29.1 → **41.0**	13.7 → **30.8**
both right	21.4 → **25.6**	23.1 → **34.2**	13.7 → **25.6**
yes/no accuracy	61.6 → **64.0**	60.4 → **63.1**	62.2 → **64.0**
Overall accuracy	30.7 → **31.2**	33.0 → **33.6**	29.3 → **29.7**
Numeric: median / mean / weighted / top branch	11.4 / 10.4 / 10.4 / 12.1	15.4 / 14.7 / 14.5 / 15.7	11.4 / 11.4 / 12.1 / 10.9
Numeric: typed read-out of Eq. (11)	**13.7**	**16.9**	**12.6**

On every backbone FutureMap gets both questions of a pair right about twice as often as the best baseline and is the only row with balanced accuracy above 55; withholding the deadline raises every accuracy shown, and the typed read-out beats the four other operators on the same branches.

F.6 Cost and Efficiency (RQ6)

For the 400 sampled items we average, per row and backbone, the provider-billed usage stored with each cached call (tokens and cost, retries included); stages 1 and 2 are built once and shared, so they add no per-item call. FutureMap makes 1.00, 1.00 and 1.07 calls per item (Gemini, Qwen,

Nemo) with prompts of 25.4k, 30.0k and 8.4k tokens, at 1.58, 2.09 and 0.17 thousandths of a dollar, below 9, 9 and 8 of the twelve baselines. The agentic baselines spend 2.6 to 20.0 calls per item and MemoTime up to 28.0, with up to 182.8k prompt tokens (MiroFlow on Gemini); the rows that cost less (TKG-ICL, ATOM, FutureOmni, and ReAct on Nemo) trail FUTUREMAP by 6.1 to 11.9 points.

F.7 ROBUSTNESS AND SIGNIFICANCE OF THE MARGIN (RQ7)

Table 13: **Paired differences**, FUTUREMAP minus each other row of Table 1: accuracy difference (points), paired bootstrap 95% interval and McNemar's z. Bottom: margin over the best baseline of each stratum, chosen per stratum and backbone.

	Gemini			Qwen			Nemo		
Other row	Δ	95% CI	z	Δ	95% CI	z	Δ	95% CI	z
No source	+10.3	[7.7, 13.0]	7.5	+10.8	[8.1, 13.5]	7.7	+9.4	[6.5, 12.2]	6.4
ReAct	+8.7	[6.3, 11.2]	6.8	+7.5	[4.8, 10.1]	5.5	+8.1	[5.4, 10.7]	5.9
TKG-ICL	+11.6	[9.0, 14.1]	8.6	+9.8	[7.3, 12.4]	7.3	+11.9	[9.1, 14.6]	8.3
ExpeL	+10.0	[7.4, 12.6]	7.5	+9.1	[6.5, 11.6]	6.6	+8.0	[5.4, 10.7]	5.6
HyperGraphRAG	+7.1	[4.6, 9.6]	5.6	+9.2	[6.5, 11.7]	6.8	+5.2	[2.4, 8.0]	3.6
MemoTime	+10.3	[7.6, 12.9]	7.6	+12.8	[10.1, 15.5]	9.1	+12.3	[9.7, 15.0]	8.9
ATOM	+7.3	[5.0, 9.7]	5.9	+7.5	[5.0, 10.0]	5.7	+8.1	[5.6, 10.8]	6.0
FutureOmni	+6.1	[3.8, 8.6]	5.1	+6.5	[4.0, 9.0]	5.0	+6.2	[3.4, 9.0]	4.3
MiroFlow	+8.8	[6.3, 11.2]	6.9	+5.3	[2.6, 7.9]	3.9	+12.0	[9.4, 14.5]	8.7
TimeSeek	+4.4	[2.1, 6.9]	3.6	+7.0	[4.4, 9.6]	5.2	+8.6	[6.1, 11.1]	6.5
mini-prophet	+8.0	[5.6, 10.3]	6.3	+9.2	[6.6, 11.7]	7.0	+6.7	[4.4, 8.9]	5.8
MMHRAG	+6.9	[4.4, 9.5]	5.4	+9.5	[6.9, 12.2]	6.9	+6.1	[3.7, 8.6]	4.7
LEAP	+4.1	[1.9, 6.3]	3.6	+7.1	[4.6, 9.3]	5.9	+4.2	[1.8, 6.7]	3.4
Stratum (items)	**Margin, Gemini**			**Margin, Qwen**			**Margin, Nemo**		
deadline b+7 (563)	+5.3			+7.1			+1.6		
deadline b+14 (630)	+1.4			+2.2			+2.2		
deadline b+21 (337)	+3.6			+6.2			+6.2		
cross-domain chain (951)	+3.8			+5.0			+4.2		
single domain (579)	+4.5			+5.5			+4.3		
not from memory (1,249)	+2.6			+2.1			+2.2		

All 39 intervals lie above zero, the lowest bound being +1.8 (Nemo, LEAP), and every z is at least 3.4. The margin over the best baseline is positive in all 18 stratum cells, including the 1,249 items not answerable from memory. FUTUREMAP alone answers 42, 48 and 31 items that no other retrieval row answers, against at most 21, 24 and 19 for a baseline (LEAP, LEAP, ExpeL).

F.8 CASE STUDY (RQ8)

Table 3 shows item `e_3e85438396e2_q1` (Politics, 21-day horizon, not answerable from memory), settled by four articles of 2026-08-11 ("Blanche was confirmed in a 50-49 vote"). Below: verbatim lines of the Qwen prompt and output.

Target. Will Todd Blanche be confirmed with a Senate vote margin of less than 3?
O11 | 2026-07-15 Acting Attorney General Todd Blanche faces a Senate confirmation hearing today. | meet => meet => appoint | 0.04 | 7/18/35
O12 | 2026-07-15 Blanche is set to appear before the Senate Judiciary Committee, which will decide… | vote => appoint | 0.05 | 3/11/33
E21 | 2026-07-15 | meet | podcast+article+TV clip, Up First from NPR | v=1.00 | Acting Attorney General Todd Blanche faces a Senate confirmation hearing today.
Output. p 0.45: "2026-07-23: Committee votes to advance nomination (via O12 lag p50)", "2026-07-28: Full Senate confirms Blanche by 51–49 vote margin"; p 0.30: … "2026-08-04: Full Senate confirms after delayed floor vote (via O11 lag p90)"; p 0.25: … "2026-07-25: Nomination withdrawn or indefinitely stalled".

A rule is kept only when its backtested hit rate exceeds the base rate of its consequence, so a small ρ marks a rare consequence, not an unreliable rule. The first-passage vote of stage 2 alone gives 0.60 on every backbone, and the read-out $\hat{p}_q$ is 0.60, 0.71 and 0.75 on Gemini, Qwen and Nemo: *Yes* on all three backbones, although no resource of the known zone reports the vote or its final margin.

G Related Work

G.1 LLM-based Future Prediction

Converting incomplete evidence into calibrated predictions links acquisition, learning, and evaluation in the Forecasting Agents survey (Xu et al., 2026b). ReAct (Yao et al., 2023), MiroFlow (Su et al., 2026), and human-level forecasting (Halawi et al., 2024) establish evidence-driven workflows; ExpeL (Zhao et al., 2024), ForecastCompass (Chang et al., 2026b), ForeDreamer (Zhong et al., 2026), and Milkyway (Wei et al., 2026) make experience reusable. SCATTER (Chang et al., 2026a), Analogical Deep Research (Chen et al., 2026c), and Agentic Time Machine (Chai et al., 2026) broaden hypothesis exploration, while Outcome-based RL (Turtel et al., 2025), OpenForecaster (Chandak et al., 2026), FutureWorld (Han et al., 2026), and temporal-search distillation (Cai et al., 2026) internalize forecasting skills. To turn evidence into probabilities, mini-prophet (Murphy, 2026), BBC (Dai et al., 2026), LEAP (Chen et al., 2026d), AuditForecast (Cao et al., 2026), and competence-gated pooling (Tiwari et al., 2026) structure belief updating and calibration; Diverse by Reasoning (Chetlapalli et al., 2026) and Artificial Deliberative Crowds (Barrera-Lemarchand et al., 2026) exploit complementary judgments.

These improvements require evaluation that separates predictive ability from information access: MIRAI (Ye et al., 2024), ForecastBench (Karger et al., 2025), and FutureX (Zeng et al., 2026) measure event forecasting, with FutureX-Pro (Liu et al., 2026a) and FutureOmni (Chen et al., 2026a) extending domain and modality coverage. TimeSeek (Mostafa et al., 2026), Hindcast (Ye et al., 2026), and SocietyBench (Wang et al., 2026) examine temporal evaluation, while Simulated Ignorance (Li et al., 2026), Proper Scoring Rules (Turtel et al., 2026), and Corroboration Illusion (Lu & Zhang, 2026) expose knowledge leakage, objective sensitivity, and evidence poisoning, respectively. OmniFP-Bench and FutureMap contribute cross-domain, five-modality evaluation with frozen evidence and temporally separated outcomes, alongside training-free prediction grounded in backtested event rules and empirical time kernels.

G.2 Graph-based Retrieval and Reasoning

Structured retrieval makes evidence relations explicit: GraphRAG (Edge et al., 2025), HyperGraphRAG (Luo et al., 2025), Hyper-RAG (Feng et al., 2026b), and HyperRAG (Lien et al., 2026) organize relational evidence, while PRoH (Zai et al., 2026), Graph-R1 (Luo et al., 2026), HyperProve (Phu et al., 2026), and EXT++/PPR (Khrouf et al., 2026) improve its retrieval. NICO-RAG (Serna-Aguilera et al., 2026a), Hyper-M2RAG (Chen et al., 2026b), GraphLoom (Ali et al., 2026), and PRISM-RAG (Serna-Aguilera et al., 2026b) extend retrieval across modalities; ECHO (Chu et al., 2026) and EvoGraph-R1 (Lin et al., 2026) construct and evolve high-order representations, with HTKGH (Ahrabian et al., 2026) and DyMRL (Zhao et al., 2026) targeting forecasting. Temporal reasoning additionally requires valid, reusable knowledge: ATOM (Lairgi et al., 2026a), OKH-RAG (Wu et al., 2026), and MemStrata (Yadav, 2026) preserve temporal structure; MemoTime (Tan et al., 2026) and SCoP (Guo et al., 2026) constrain temporal evidence. LivingRAG (Cui et al., 2026a), HyperSkill (Xu et al., 2026a), MAGE (Feng et al., 2026a), and EdgeMem (Cui et al., 2026b) make structured experience reusable.

Beyond retrieval, historical regularities support demonstrations in TKG-ICL (Lee et al., 2023), neural-symbolic inference in ONSEP (Yu et al., 2024) and INFER (Li et al., 2025), and rule learning in TILP (Xiong et al., 2023) and CountTRuCoLa (Gastinger et al., 2026); TKG-Thinker (Jiang et al., 2026) extends agentic reasoning to temporal question answering. Event-CausNet (Niu et al., 2025), AECR (Zheng et al., 2026), and Event-Causal RAG (Yan et al., 2026) incorporate causal evidence; C-Unseen (Lairgi et al., 2026b) discovers weak signals, and trace-guided interventions (Yu & guk Ha, 2026) test alternative outcomes. TIEM (Liu et al., 2026b) integrates temporal hypergraph evidence and skill memory for financial forecasting. FutureMap contributes a unified omnimodal temporal hypergraph whose backtested rule confidences and lag distributions constrain future branches, coupling deadline-blind timing with deterministic answer read-out.