Assessing The (In)Ability of LLMs To Reason in Interval Temporal Logic

Pietro Bellodi[1,2] Pietro Casavecchia[1,2] Alberto Paparella[1,2]
Guido Sciavicco[1,2] Ionel Eduard Stan[3,2]

[1]Department of Mathematics and Computer Science, University of Ferrara, Italy
[2]Applied Computational Logic and Artificial Intelligence Lab (ACLAI)
[3]Department of Informatics, Systems, and Communications, University of Milano-Bicocca, Italy

29/08/2025

I

Introduction

Objective

1 Generate a synthetic formal logic benchmark

→

2 Evaluate LLM performance

Why synthetic benchmarks?

Contamination

Often solutions become part of the training corpus of LLMs.

Training data decontamination procedures are often only partially effective.

Scalability

Manually created benchmarks are difficult to scale both in *size* and *complexity*, requiring significant human effort and financial investments.

II.I

Generation (Logic)

Halpern–Shoham's Logic (HS) is a modal logic where *intervals*, rather than points, are the fundamental states.

Key idea: Accessibility between intervals is determined by *Allen's interval relations*.

Given a linear order $\mathbb{D} = \langle D, < \rangle$, a *strict interval* is an ordered pair $[x, y]$ with $x, y \in D$ and $x < y$.

Two intervals $[x, y]$ and $[w, z]$ are compared by their endpoints, and their relation is captured by one of Allen's modalities.

HS introduces an *existential modality* $\langle X \rangle$ for each Allen relation R_X. The six basic relations A, L, B, E, D, O each have an inverse $\overline{X}$, yielding 12 binary relations in total.

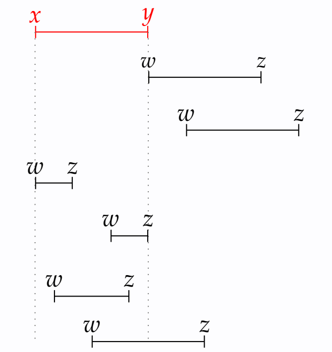
HS modality	Definition w.r.t. the interval structure	Example
		
$\langle A \rangle$ (adjacent)	$[x,y] R_A [w,z] \quad \Leftrightarrow \quad y = w$	
$\langle L \rangle$ (later)	$[x,y] R_L [w,z] \quad \Leftrightarrow \quad y < w$	
$\langle B \rangle$ (begins)	$[x,y] R_B [w,z] \quad \Leftrightarrow \quad x = w \wedge z < y$	
$\langle E \rangle$ (ends)	$[x,y] R_E [w,z] \quad \Leftrightarrow \quad y = z \wedge x < w$	
$\langle D \rangle$ (during)	$[x,y] R_D [w,z] \quad \Leftrightarrow \quad x < w \wedge z < y$	
$\langle O \rangle$ (overlaps)	$[x,y] R_O [w,z] \quad \Leftrightarrow \quad x < w < y < z$	

Table: Allen's interval relations and HS modalities.

Syntax of HS

Alphabet: propositional letters $\mathcal{P}$, classical connectives $\neg, \vee$, and modalities $\langle X \rangle$ for $X \in \mathcal{X}$ wit:

$$\mathcal{X} = \{A, \overline{A}, L, \overline{L}, B, \overline{B}, E, \overline{E}, D, \overline{D}, O, \overline{O}\}.$$

Grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \langle X \rangle \varphi \qquad (p \in \mathcal{P},\ X \in \mathcal{X}).$$

Derived connectives: $\varphi \wedge \psi \equiv \neg(\neg\varphi \vee \neg\psi)$, $\varphi \rightarrow \psi \equiv \neg\varphi \vee \psi$, $\top \equiv p \vee \neg p$.

Universal modality: $[X]\varphi \equiv \neg\langle X \rangle \neg\varphi$.

Interval model: $M = \langle \mathbb{I}(\mathbb{D}), V \rangle$ with $\mathbb{D}$ a linear order, $\mathbb{I}(\mathbb{D})$ the set of strict intervals over $\mathbb{D}$, and $V : \mathcal{P} \to 2^{\mathbb{I}(\mathbb{D})}$ a valuation.

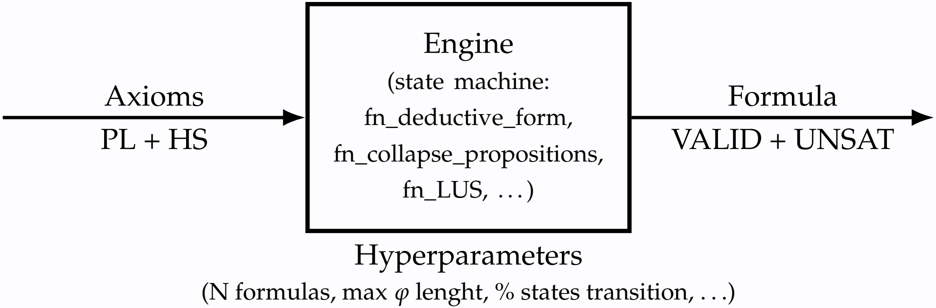
Truth on an interval $[x, y]$:

$$
\begin{array}{lcl}
M, [x,y] \models p & \Leftrightarrow & [x,y] \in V(p) \quad (p \in \mathcal{P}), \\
M, [x,y] \models \neg\psi & \Leftrightarrow & M, [x,y] \not\models \psi, \\
M, [x,y] \models \psi_1 \vee \psi_2 & \Leftrightarrow & M, [x,y] \models \psi_1 \text{ or } M, [x,y] \models \psi_2, \\
M, [x,y] \models \langle X \rangle \psi & \Leftrightarrow & \exists [w,z] \text{ s.t. } [x,y] R_X [w,z] \text{ and } M, [w,z] \models \psi.
\end{array}
$$

II.II

Generation (Scheme)

Scheme


Axioms
PL + HS
Engine
(state machine:
fn_deductive_form,
fn_collapse_propositions,
fn_LUS, ...)
Formula
VALID + UNSAT
Hyperparameters
(N formulas, max φ lenght, % states transition, ...)

II.III

Generation (Input)

● ● ● ● ○ ○ ○ ○ ○

Axiom	Type
$(a \rightarrow b) \leftrightarrow (\neg b \rightarrow \neg a)$	contrapositive
$(a \rightarrow b) \leftrightarrow (\neg a \vee b)$	implication as disjunction
$a \wedge (b \vee c) \leftrightarrow (a \wedge b) \vee (a \wedge c)$	distributivity of $\wedge$ over $\vee$
$\neg(a \wedge b) \leftrightarrow \neg a \vee \neg b$	De Morgan
$\neg(a \vee b) \leftrightarrow \neg a \wedge \neg b$	De Morgan
$a \wedge a \leftrightarrow a$	idempotence
$a \wedge \top \leftrightarrow a$	identity
$\neg\neg a \leftrightarrow a$ a	double negation
$a \wedge (a \rightarrow b) \rightarrow b$	modus ponens
$(a \rightarrow b) \wedge \neg b \rightarrow \neg a$	modus tollens
$\neg a \wedge (a \vee b) \rightarrow b$	disjunctive syllogism

Table: All propositional axioms used.

HS Axioms

Axiom	Type
$\langle \overline{B} \rangle \langle \overline{E} \rangle p \leftrightarrow \langle \overline{E} \rangle \langle \overline{B} \rangle p$	commutativity
$\langle L \rangle p \leftrightarrow \langle A \rangle \langle A \rangle p$	definability
$\neg [B] p \leftrightarrow \langle B \rangle \neg p$	duality
$[A](p \to q) \to ([A]p \to [A]q)$	K axiom
$\langle B \rangle \langle B \rangle p \to \langle B \rangle p$	transitivity
$\langle B \rangle [\overline{B}] p \to p$	temporality
$p \to [A] \langle \overline{A} \rangle p$	inverse of temporality
$\langle A \rangle \langle \overline{A} \rangle p \to [A] \langle \overline{A} \rangle p$	stability

Table: Selected HS axioms, one for each type.

II.IV

Generation (Output)

Output

<table>
<tr><th>id_enriched</th><th>premise_size</th><th>conclusion_size</th><th>total_size</th><th>compositional_hops</th><th>id_base</th><th>id_plus</th></tr>
<tr><td>Te1</td><td>6</td><td>7</td><td>14</td><td>1</td><td>Ae9</td><td>Ae41</td></tr>
<tr><td colspan="7">base: $\neg(t \vee q) \models (\neg t \wedge \neg q)$</td></tr>
<tr><td colspan="7">plus: $\neg[\text{later}](s) \models \langle\text{later}\rangle\neg s$</td></tr>
<tr><td colspan="7">enriched: $\neg(t \vee \neg[\text{later}](s)) \models (\neg t \wedge \neg\langle\text{later}\rangle\neg s)$</td></tr>
<tr><td colspan="7">enriched_valid: $\neg(\neg(t \vee \neg[\text{later}](s))) \wedge (t \vee \langle\text{later}\rangle\neg s)$</td></tr>
<tr><td colspan="7">enriched_unsat: $\neg(t \vee \neg[\text{later}](s)) \wedge (\neg(\neg t) \vee \langle\text{later}\rangle\neg s)$</td></tr>
<tr><td>Ti8</td><td>7</td><td>7</td><td>15</td><td>2</td><td>Ti3</td><td>Ae54</td></tr>
<tr><td colspan="7">base: $\langle\text{meets}\rangle\langle\text{met\_by}\rangle\neg[\text{begins}](q) \models [\text{meets}]\langle\text{met\_by}\rangle\langle\text{begins}\rangle\neg q$</td></tr>
<tr><td colspan="7">plus: $\langle\text{during}\rangle\neg p \models \neg[\text{during}](p)$</td></tr>
<tr><td colspan="7">enriched: $\langle\text{meets}\rangle\langle\text{met\_by}\rangle\neg[\text{begins}](\langle\text{during}\rangle\neg p) \models [\text{meets}]\langle\text{met\_by}\rangle\langle\text{begins}\rangle(\neg\neg[\text{during}](p))$</td></tr>
<tr><td colspan="7">enriched_valid: $[\text{meets}][\text{met\_by}][\text{begins}](\langle\text{during}\rangle\neg p) \vee [\text{meets}]\langle\text{met\_by}\rangle\langle\text{begins}\rangle[\text{during}](p)$</td></tr>
<tr><td colspan="7">enriched_unsat: $(\langle\text{meets}\rangle\langle\text{met\_by}\rangle\neg[\text{begins}](\langle\text{during}\rangle\neg p)) \wedge \langle\text{meets}\rangle[\text{met\_by}]\langle\text{begins}\rangle\neg[\text{during}](p)$</td></tr>
</table>

Table: Example of raw output of two sets of valid and unsatisfiable formulas.

II.V

Generation (Engine)

Formula Structures

Equivalence form: $\varphi \leftrightarrow \psi$

Implicative form: $\varphi \rightarrow \psi$

Formula Roles

Base types: equivalence or implicative

Plus types: only equivalence

Substitution Principle

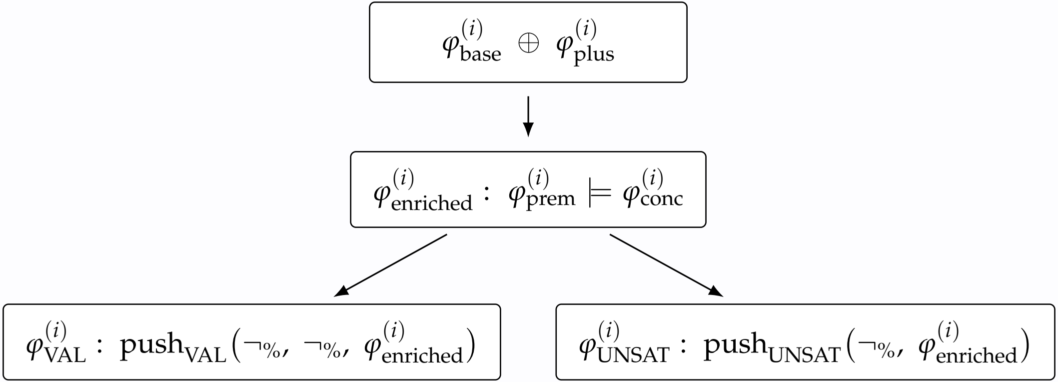
Inspired by uniform substitution:

$$\varphi^{base}_{prem}[p/\psi] \models \varphi^{base}_{conc}[p/\psi]$$

Instead of a WFF ψ, we use a plus formula:

$$\varphi_{enriched} : \varphi^{base}_{prem}[p/\varphi^{plus}_{prem}]_{partial} \models \varphi^{base}_{conc}[p/\varphi^{plus}_{conc}]$$

Valid and unsatisfiable formulas



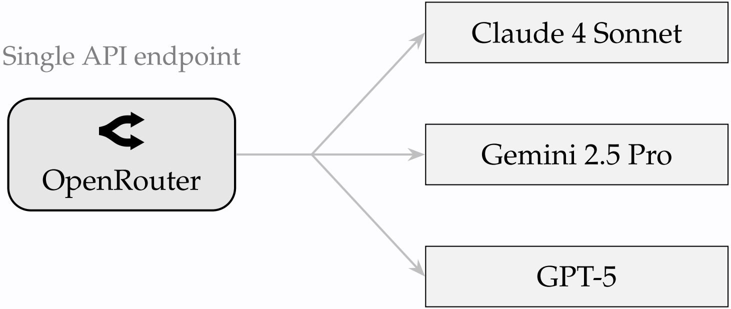
Tree view of Formula Construction

```
Ti77: ([finished](!(<begun_by>(!((p & !([before](t))))))) | //
|       // [met_by](<meets>(<overlaps>((!(p) | !(<before>(!(t))))))))
|-- Ti32b: ([finished]([begun_by]((p & t))) | //
|       // [met_by](<meets>(<overlaps>((!(p) | !(t))))))
|   |-- Ti18b: ([finished]([begun_by](!(p))) | //
|   |   // [met_by](<meets>(<overlaps>(p))))
|   |   |-- Ti15b: (!(q) | [met_by](<meets>(q)))
|   |   |   |-- Ai38b: p |= [met_by](<meets>(p))
|   |   |   |-- Ae16p: q |= !(!(q))
|   |   |-- Ae30p: <finished>(<begun_by>(p)) |= <overlaps>(p)
|   |-- Ae7p: !((p & t)) |= (!(p) | !(t))
|-- Ae43p: !([before](t)) |= <before>(!(t))
```

III

Evaluation

OpenRouter


Single API endpoint
OpenRouter
Claude 4 Sonnet
Gemini 2.5 Pro
GPT-5

Adopted prompting strategies

Few-Shot

Provides **complete examples** of problems with their solutions to facilitate learning by analogy without explicit instructions.

CoT

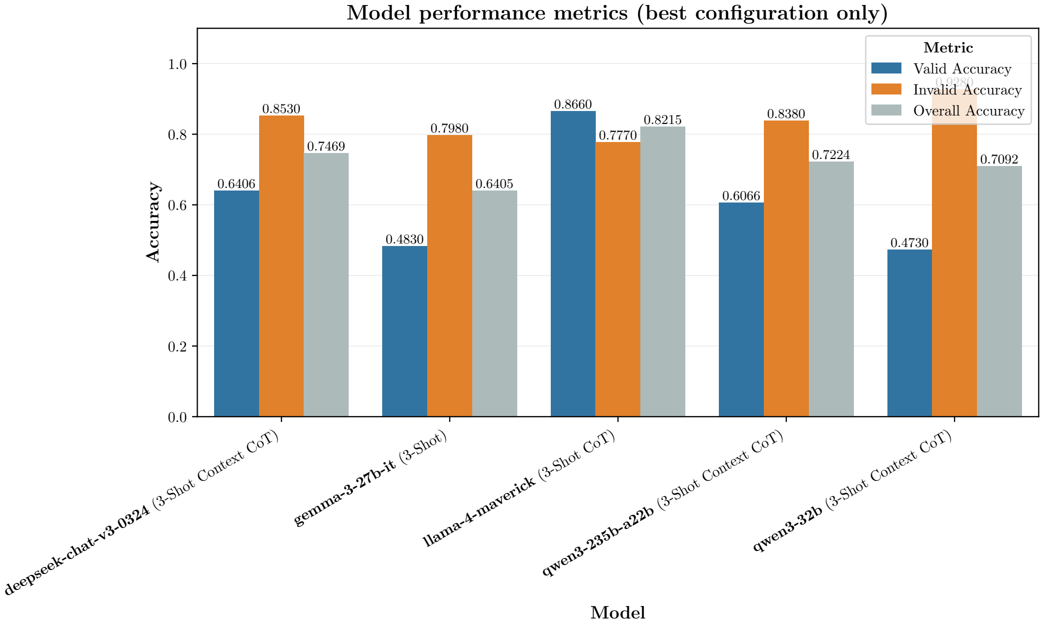
Guides the model to **decompose** the problem into multiple components to facilitate its resolution.

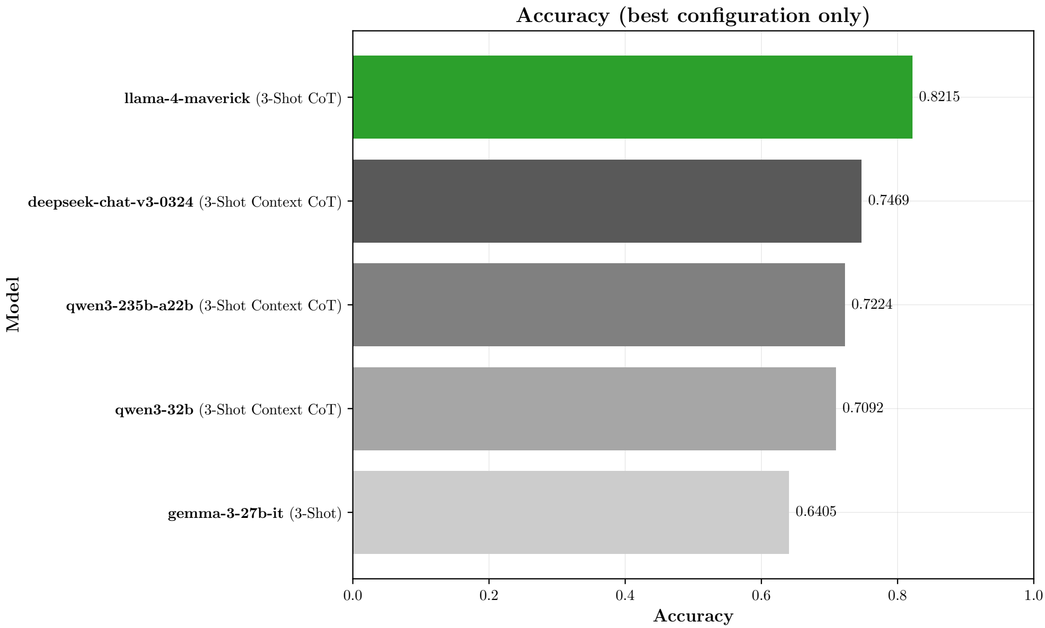
Context

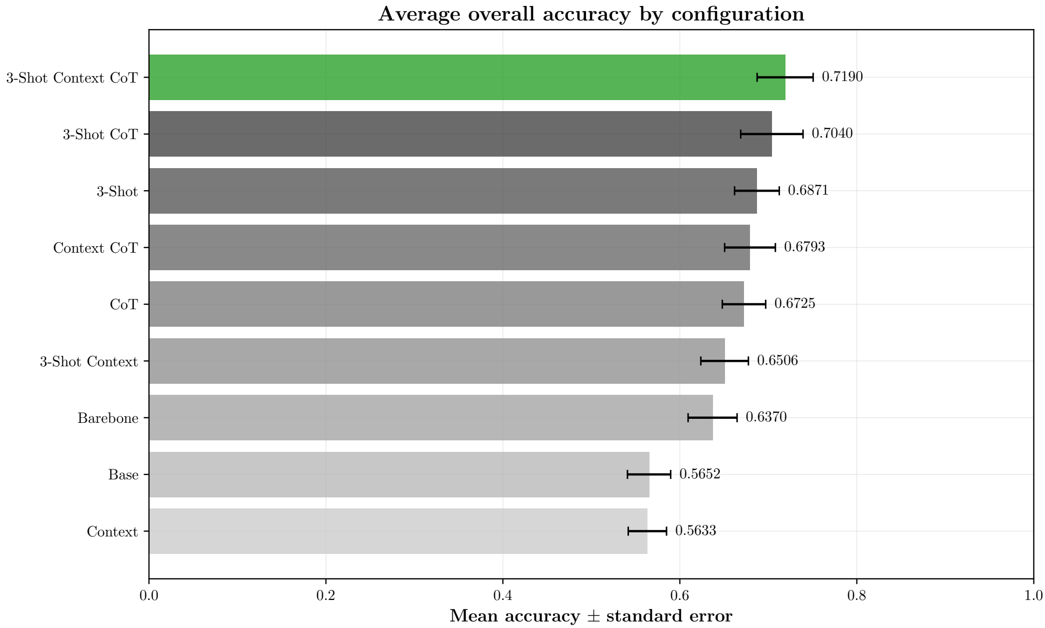
A natural language **introduction** to LTL is provided followed by its syntax and semantics. A natural language **introduction** to HS is provided followed by its syntax and semantics.

IV

Results

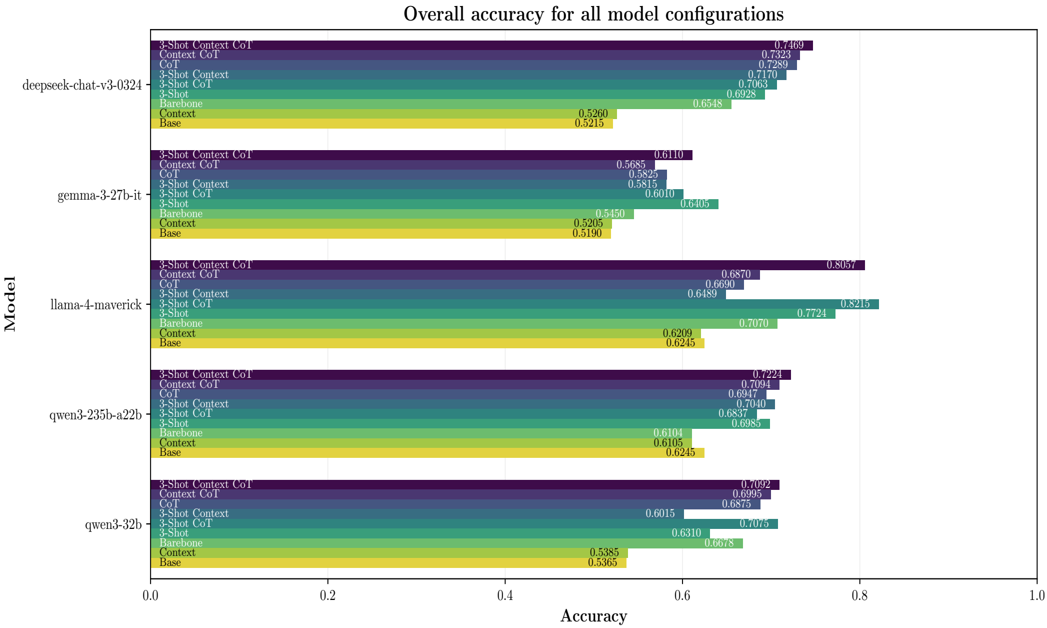

Model performance metrics (best configuration only)
Metric
Valid Accuracy
Invalid Accuracy
Overall Accuracy
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.6406
0.8530
0.7469
0.4830
0.7980
0.6405
0.8660
0.7770
0.8215
0.6066
0.8380
0.7224
0.4730
0.9280
0.7092
deepseek-chat-v3-0324 (3-Shot Context CoT)
gemma-3-27b-it (3-Shot)
llama-4-maverick (3-Shot CoT)
qwen3-235b-a22b (3-Shot Context CoT)
qwen3-32b (3-Shot Context CoT)
Model


Accuracy (best configuration only)
llama-4-maverick (3-Shot CoT)
deepseek-chat-v3-0324 (3-Shot Context CoT)
qwen3-235b-a22b (3-Shot Context CoT)
qwen3-32b (3-Shot Context CoT)
gemma-3-27b-it (3-Shot)
Model
0.8215
0.7469
0.7224
0.7092
0.6405
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy


Average overall accuracy by configuration
3-Shot Context CoT
0.7190
3-Shot CoT
0.7040
3-Shot
0.6871
Context CoT
0.6793
CoT
0.6725
3-Shot Context
0.6506
Barebone
0.6370
Base
0.5652
Context
0.5633
0.0
0.2
0.4
0.6
0.8
1.0
Mean accuracy ± standard error

V

Appendix



Overall accuracy for all model configurations
Model
deepseek-chat-v3-0324
gemma-3-27b-it
llama-4-maverick
qwen3-235b-a22b
qwen3-32b
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
3-Shot Context CoT 0.7469
Context CoT 0.7323
CoT 0.7289
3-Shot Context 0.7170
3-Shot CoT 0.7063
3-Shot 0.6928
Barebone 0.6548
Context 0.5260
Base 0.5215
3-Shot Context CoT 0.6110
Context CoT 0.5685
CoT 0.5825
3-Shot Context 0.5815
3-Shot CoT 0.6010
3-Shot 0.6405
Barebone 0.5450
Context 0.5205
Base 0.5190
3-Shot Context CoT 0.8057
Context CoT 0.6870
CoT 0.6690
3-Shot Context 0.6489
3-Shot CoT 0.8215
3-Shot 0.7724
Barebone 0.7070
Context 0.6209
Base 0.6245
3-Shot Context CoT 0.7224
Context CoT 0.7094
CoT 0.6947
3-Shot Context 0.7040
3-Shot CoT 0.6837
3-Shot 0.6985
Barebone 0.6104
Context 0.6105
Base 0.6245
3-Shot Context CoT 0.7092
Context CoT 0.6995
CoT 0.6875
3-Shot Context 0.6015
3-Shot CoT 0.7075
3-Shot 0.6310
Barebone 0.6678
Context 0.5385
Base 0.5365