Summary of "Language Models with Rationality" (REFLEX system)

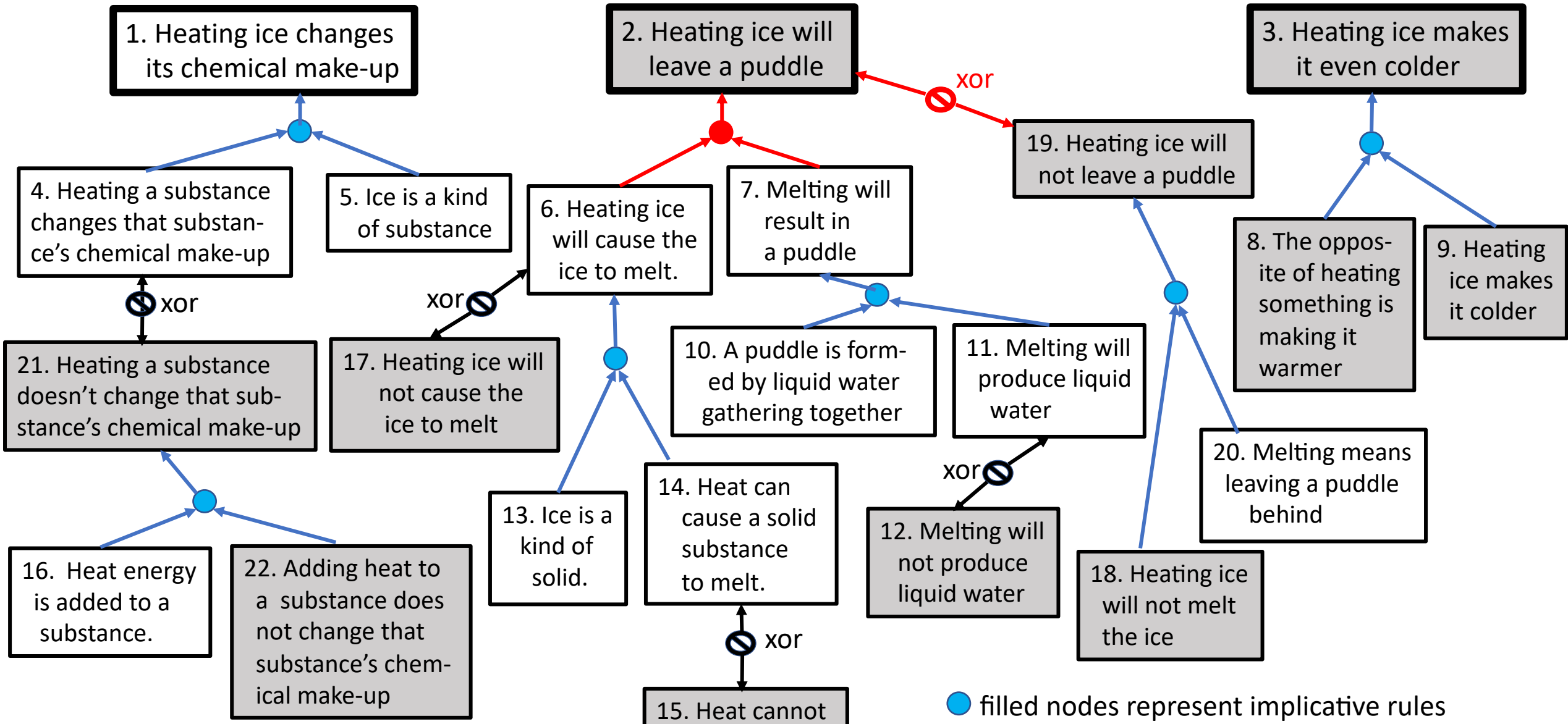
(Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schütze, & Peter Clark, Allen Institute for AI, Seattle, WA), arXiv2305.14250v2 [cs.CL] 29 Oct 2023)

- Arguably the most interesting LLM-based reasoning paper to date (as of Nov. 2023);
- Target domains: Multiple-choice questions from AP science; general "textual entailment";
  Simple example: "Which animal gives live birth? (A) giraffe (B) spider"
- ***Only the LLM's internal knowledge is used – and is made explicit;***
- **Problem-solving control is not by the LLM, but by a recursive algorithm layered on top of it**;
- While QA gains are small (75% vs 74%), consistency gains are large (96% vs 88%)

The algorithm goes roughly like this (omitting LLM-based assignment of certainty factors):

- Query the LLM about each choice (in sentence form), hence mark them T/F (tentatively);
- Introduce the negated claims, connected via XOR "rules"; use the LLM to mark them T/F;
- Regardless of the T/F values, for each claim C (and negations) ask the LLM for a rule of form, e.g., (C1 & C2) => C; add these "rules" to the proof graph; mark C1, C2 T/F (use LLM);
- Continue chaining back in this way, till a depth limit is reached;
- Find inconsistencies; "flip" a minimal set of T/F values to get consistency; return likeliest answer.

Multiple choice task: Determine which of 1, 2, 3 is true



White boxes: Claims that the LLM considers true

Grey boxes: Claims that the LLM considers false;
NB: Most negated claims have been omitted.

filled nodes represent implicative rules

XOR nodes connect each claim to its denial

RED connections indicate conflicts

"Flipping" 2 to "true" resolves the conflicts

Probabilities

Such graphs are thought of as *factor graphs* (generalizing *Bayes & Markov nets*), allowing representation of *a joint distribution as a product of factors*):

- The statements ***s1, …, sk*** in the boxes are probabilistic Boolean ***variables***;
- The round nodes ***r1, …, rl*** are ***rules***, denoting 0/1 functions of the var's they connect;
- An ***unnormalized prob. dist.***^{n} Φ over any 0/1 truth values ***s1*** = ***x1, …, sk*** = ***xk*** of the var's is obtained as a product of exponentials over all k statements and l rules

$$\Phi = e^{-(y1+y2+ \dots + yk) - (z1+z2+ \dots + zl)},$$

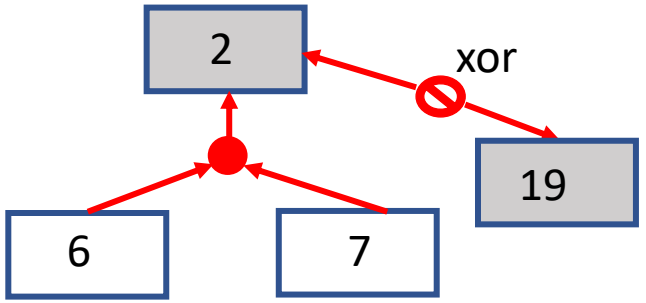
The yi are determined by the xi, & the zj are determined by the var. values at the nodes connected by rj

- where yi = 0 if xi agrees with the LLM's truth judgement about ***si***, and
  = ci otherwise, where ci is the confidence (in [0,1]) expressed by the LLM about its truth value judgement; (So, xi determines yi)
  zj = 0 if the value of rule rj is 1 according to the LLM's truth value judgements,
  = cj' otherwise, where cj' is the confidence (in [0,1]) expressed by the LLM about the correctness of rule rj that it proposed.
- NB: $\Phi = e^{-y1} \cdot e^{-y2} \dots e^{-zl}$; thus, a factored distribution.
- To normalize Φ at ***s*** = *x*, divide it by the sum over all such variable assignments.

*As far as I can tell, Kassner et al. don't actually **use** such distributions (e.g., for updating confidence levels).*

Consistency algorithm

Recall the example from the figure, where 6 & 7 are judged true & entail 2, yet the LLM judges 2 to be false;



...and 2, 19 are XOR-related, yet the LLM judges both false

"Flipping" the truth value of 2 repairs both conflicts. But how do we find the best "flips", disrupting the LLM's T/F judgements as little as possible?

Simply minimizing the number of flips may not be good, because the LLM assigns different degrees of confidence to both the variables (sentences claimed to be T or F) and to the rules it proposes.

The authors convert the set of beliefs (variables with presumed truth values and degrees of confidence) into a weighted MaxSAT problem. Both flipping truth values, and violating rules, have associated costs, based on (recalibrated versions of) the LLM's level of confidence in the truth values and rules. The minimal-cost MaxSAT solution then provides the flips that maximize consistency.

Then they delete rules that are responsible for any remaining inconsistencies.

The resulting graph enables giving consistent justifications of answers.

Assessment: A major gain in logical consistency, but …

How close does this get us to reliable general reasoning using LLMs?

- The test accuracy is still only 75% -- short of that of many students;
- The method relies on the fact that exactly one of the choices is correct;
- There is no "open-ended" reasoning (which is more natural; hard to evaluate);
- The statements comprising LLM-proposed "inference rules" are too imprecise to support reliable reasoning; e.g.,

  "Giraffes give live birth" should really be,
  "When a giraffe gives birth, it gives live birth"

  "Dogs are mammals", & *"Dogs are barking"* should really be
  "All dogs are mammals", & *"Some dogs are barking"* respectively;

  "Heating ice will leave a puddle" should really be
  "Heating water ice continuously on a level, solid surface will eventually leave a puddle". (cf. ice cubes in your beverage, dry ice)

- Perhaps most importantly, *there is no abstract model building and manipulation* of the type people employ in solving more combinatorially intricate problems … e.g., *Missionaries & Cannibals, Towers of Hanoi, planning your college courses, creating a start-up company*, …

See examples of bad rules resulting from this, p.8 of Kassner et al.; e.g., *"Some people don't mind not moving for an hour" + "Breathing is a kind of moving"* ➔ *"Some people don't mind not breathing for an hour."*