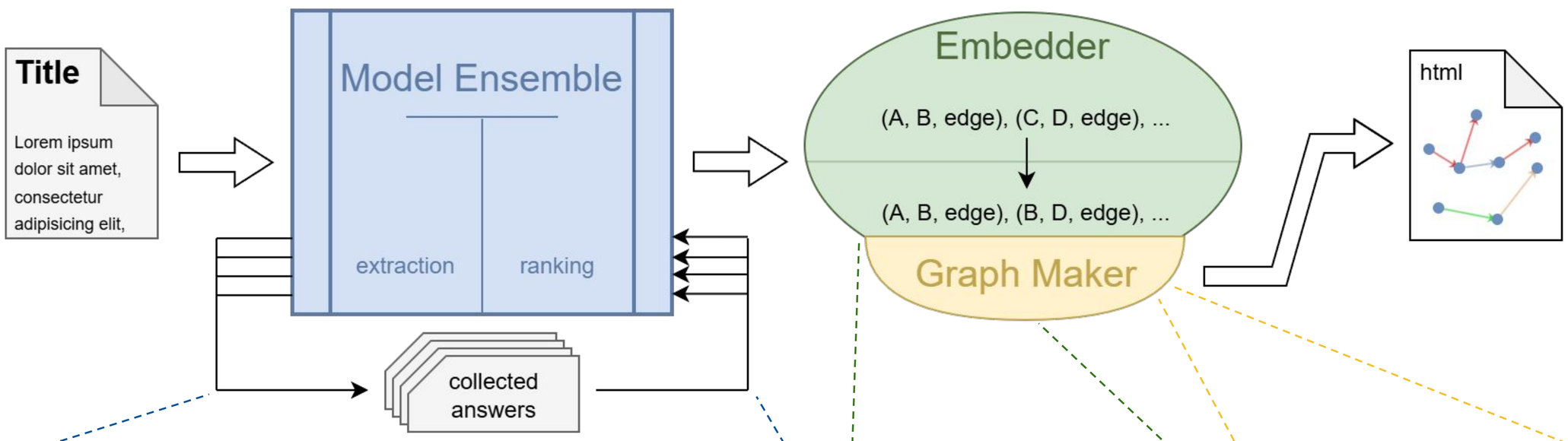
ReX-GG: a LLM Ensemble Pipeline for Relation Extraction and Graph Generation

Giacomo Magnifico, Eduard Barbu

This work proposes a customisable ensemble of coordinated LLMs that leverages JSON-structured outputs and anonymous peer-review ranking to mitigate hallucinations and single-model failure points. We demonstrate the robustness of the implementation on a relation extraction task applied to popular science articles in English.



"Your lack of sleep is making you clumsy. That's why you spilled the milk."

[{**bond**: strong causation}, {**span_1**: lack of sleep}, {**span_2**: clumsy}],

[{**bond:** strong causation}, {**span_1:** clumsy}, {**span_2:** spilled milk}]

"Diseases with uncertain causes were about 50% more likely to attract religious or magical treatments."

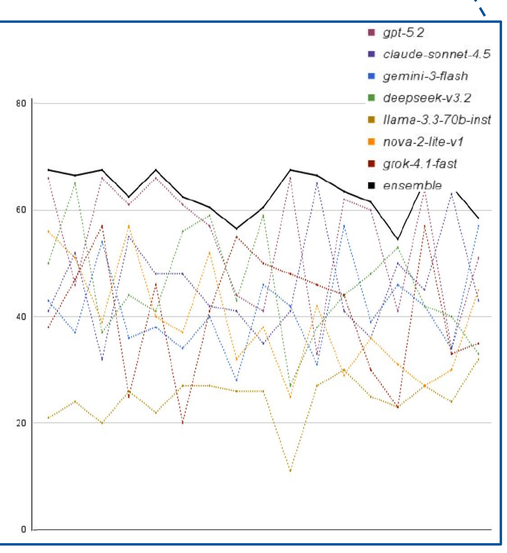
[{**bond:** weak causation}, {**span_1:** diseases with uncertain causes}, {**span_2:** esoteric treatment}]

"In the treatment group, 70% recovered; in the placebo group, only 30% recovered."

[{**bond:** contrastive}, {**span_1:** treatment group, 70% recovered}, {**span_2:** placebo group, 30% recovered}]

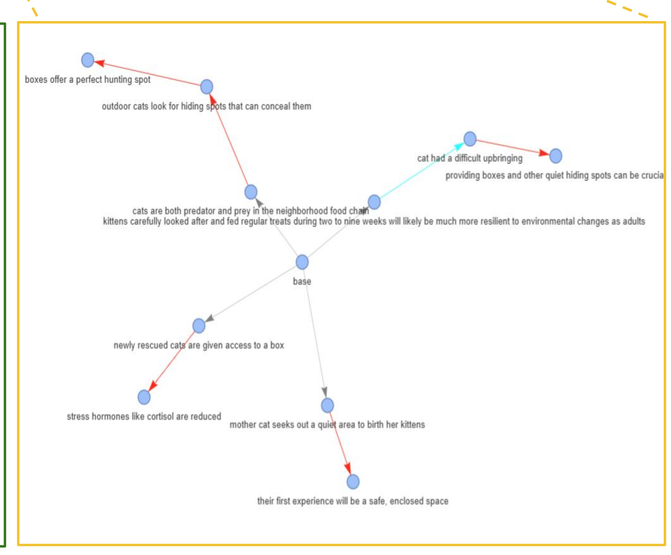
"Protons and neutrons are combinations of even tinier particles, called quarks."

[{**bond:** compositional}, {**span_1:** protons and neutrons}, {**span_2:** quarks}]

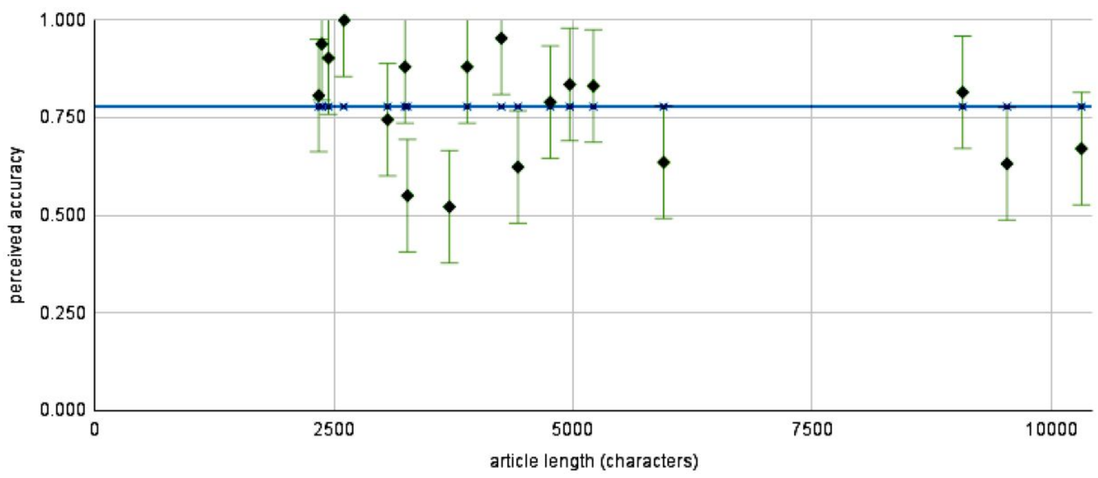


Algorithm 1 Semantic Triples Node Assimilation

$inputs \leftarrow$ list of dictionaries
$relations = [\,]$
for i in $inputs$ **do**
  if $i[relation] \neq$ none **then**
    $pair \leftarrow (i[A], i[B], i[relation])$
    $relations \leftarrow relations + pair$
  end if
end for
for $i = 1$ to $i = \#relations$ **do**
  $x \leftarrow relations[i-1][1]$
  $y \leftarrow relations[i][0]$
  $X, Y = embedder.encode(x, y)$
  if $embedder.similar(X, Y) \geq value$ **then**
    $y \leftarrow x$
  end if
end for



HUMAN EVALUATION



Evaluation of graph correctness via anonymous survey on the Prolific web app. Each article was scored by 6 testers as follows:

- 4 points, usefulness of the graph
- 5 points, accuracy of the information in the graph
- 4 points, score for the tester's confidence
  (used for weighted average across same-article testers)

Scores for author-annotated graphs were used as a gold standard baseline for the evaluation.

	Accuracy	Length	$\times\sigma$
A1	0.807	2342	0.2
A13	0.939	2374	1.1
A14	0.903	2444	0.9
A7	1.000	2603	1.6
A10	0.745	3062	0.2
A12	0.880	3245	0.7
A5	0.550	3270	1.6
A18	0.521	3709	1.8
A3	0.880	3895	0.7
A15	0.953	4253	1.2
A6	0.623	4426	1.1
A4	0.790	4764	0.1
A8	0.835	4967	0.4
A16	0.831	5215	0.4
A11	0.636	5950	1.0
A17	0.815	9078	0.3
A2	0.632	9542	1.0
A9	0.670	10321	0.8
Avg.	*0.778*	*4748*	*1*

PERFORMANCE ANALYSIS

Ensemble vs Single-model Rank. Single LLM performance has wide margin of potential error for malformed outputs. Ensemble rankings mitigate catastrophic failure states.

Document Length vs Accuracy. Result analysis shows no relation between document length and perceived accuracy, both highest and lowest accuracy are within initial range (2k - 5k).

Human-rated Accuracy. Pipeline produces high-quality outputs with near-human peaks and 0.778 average. Normal-adjacent distribution.

Robustness. Architecture design forces JSON structure outputs, with multiple error checks in place. Peer-review ranking is anonymous and independent, avoids error propagation. Design allows for single-LLM multi-agent setups for lighter-weight scenarios. In-house and API usage equally available.

www.tartunlp.ai

{name.surname}@ut.ee

TARTU ÜLIKOOL
arvutiteaduse instituut
1632