VerifAI: Verified Generative AI (Vision)

Nan Tang  
HKUST (GZ)  
China

Chenyu Yang  
HKUST (GZ)  
China

Ju Fan  
Renmin Univ.  
China

Lei Cao  
Univ. of Arizona/MIT  
USA

Yuyu Luo  
HKUST (GZ)  
China

Alon Halevy  
Amazon  
USA
Outline

• What
• How
• Results
• Opportunities
2018: GPT-1 and GPT-2
2020: GPT-3 (Wow)
2022: ChatGPT (Amazing)
2023: GPT-4 (Unbelievable)

100,000+ citations
Do you trust the results?

Hallucination

- Inaccurate decision-making
- Spread misinformation
- Privacy violation
- Legal liabilities
- Damage to trust
- ... ...

Downstream tasks?
Errors can be propagated ...
• **LLMs: Generative AI**
  • Learn a distribution from finite data
  • Generalize to infinite data points

• **Goals of Generative AI**
  • Generalizable and natural
  • Not factual

• **LLMs vs. data compression models?**
  • Is not
  • Can not
Outline

• **What**: Verify the outputs of LLMs (True/False)
• **How**
• **Results**
• **Opportunities**
Generative AI

(a) ChatGPT for tuple generation/completion/augmentation

<table>
<thead>
<tr>
<th>district</th>
<th>incumbent</th>
<th>candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ohio's 1st congressional district</td>
<td>Steve Chabot (R)</td>
<td>Steve Chabot (R), John Cranley (D), James Berns (Libertarian)</td>
</tr>
<tr>
<td>Ohio's 2nd congressional district</td>
<td>Rob Portman (R)</td>
<td>Rob Portman (R), Charles W. Sanders (D), Edward S. Focke (Green)</td>
</tr>
<tr>
<td>Ohio's 3rd congressional district</td>
<td>Tony P. Hall (D)</td>
<td>Tony P. Hall (D), Richard E. Stobbs (R), John Fockler (Libertarian)</td>
</tr>
</tbody>
</table>

(b) ChatGPT for text generation

Does Meagan Good play a role in Stomp the Yard?

No, Meagan Good does not appear in the 2007 film “Stomp the Yard.” “Stomp the Yard” is a dance drama film directed by Sylvain White, and Meagan Good is not listed as part of the main cast or credited with a role in the film. The main cast of “Stomp the Yard” includes actors such as Columbus Short, Chris Brown, Ne-Yo, and Brian White, among others. Meagan Good is an American actress known for her roles in films such as “Think Like a Man,” “Eve’s Bayou,” and “Waist Deep,” but she did not appear in “Stomp the Yard.”
(a) ChatGPT for tuple generation/completion/augmentation

You

Please help me fill in the missing values denoted by 'NULL' in this table and provide me with the completed one:

<table>
<thead>
<tr>
<th>Party</th>
<th>Candidate</th>
<th>Votes</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Republican</td>
<td>Mike Turner</td>
<td>107,286</td>
<td>59.90</td>
</tr>
<tr>
<td>Democratic</td>
<td>Jane Mitakides</td>
<td>119,448</td>
<td>37.71</td>
</tr>
<tr>
<td>Republican</td>
<td>hold</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) ChatGPT for text generation

**Stomp the Yard**

From Wikipedia, the free encyclopedia

The film stars **Columbus Short**, Meagan Good, Brian White, Laz Alonso, and Valerie Pettiford, with Harry Lennix, and, in their film debuts, R&B singers Ne-Yo & Chris Brown. **Stomp the Yard** was filmed in Atlanta, Georgia, on the campuses of Morris Brown College, Georgia Institute of Technology, Morehouse College, and Clark Atlanta University, and in the MAK Historic District of Decatur, Georgia.

Elsewhere Short, Alonso and Brown had also starred together in the film **This Christmas**.

Meagan Good’s Filmography

<table>
<thead>
<tr>
<th>Year</th>
<th>Title</th>
<th>Role</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Venom</td>
<td>Cece</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Miles from Home</td>
<td>Natasha Freeman</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Stomp the Yard</td>
<td>April Palmer</td>
<td></td>
</tr>
</tbody>
</table>
Outline

• What: Verify the outputs of LLMs (True/False)
• How
• Results
• Opportunities
Retrieve relevant dataset(s) to verify factual (not subjective) answers.
Outline

• What: Verify the outputs of LLMs (True/False)
• How: Retrieval
• Results
• Opportunities
A Deep Look into Neural Ranking Models for Information Retrieval

Jiafeng Guo\textsuperscript{a,b}, Yining Fan\textsuperscript{a,b}, Liang Peng\textsuperscript{a,c}, Liu Yang\textsuperscript{a}, Qingyou Ai\textsuperscript{a}, Hamed Zamani\textsuperscript{a}, Chen Wu\textsuperscript{a,c}, W. Bruce Croft\textsuperscript{a}, Xueqi Cheng\textsuperscript{a,b}

\textsuperscript{a}University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{b}CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
\textsuperscript{c}Center for Intelligent Information Retrieval, University of Massachusetts Amherst, Amherst, MA, USA

Abstract

Ranking models lie at the heart of research on information retrieval (IR). During the past decades, different techniques have been proposed for constructing ranking models, from traditional heuristic methods, probabilistic methods, to modern machine learning methods. Recently, with the advance of deep learning technology, we have witnessed a growing body of work in applying shallow or deep neural networks to the ranking problem in IR, referred to as neural ranking models in this paper. The power of neural ranking models lies in the ability to learn from the raw text inputs for the ranking problem to avoid many limitations of hand-crafted features. Neural networks have sufficient capacity to model complicated tasks, which is needed to handle the complexity of relevance estimation in ranking. Since there have been a large variety of neural ranking models proposed, we believe it is the right time to summarize the current status, learn from existing methodologies, and gain some insights for future development. In contrast to existing reviews, in this survey, we will take a deep look into the neural ranking models from different dimensions to analyze their underlying assumptions, major design principles, and learning strategies. We compare these models through benchmark tasks to obtain a comprehensive empirical understanding of the existing techniques. We will also discuss what is missing in the current literature and what are the promising and desired future directions.
VerifAI

The singer of the song "Mama's Song" did not win any specific contest related to the song. The song was released as a single by country music artist Carrie Underwood and was well-received by fans and critics alike.

You are wrong!!!

1. LLM

The singer of the song "Mama's Song" won the Junior Eurovision Song Contest in 2010. This is evident from the passage: "The contest was won by Vladimir Arzumanyan from Armenia"

2. RAG

Passage 1: The Eurovision Song Contest 2001 was ...
Passage 2: The Eurovision Song Contest 1996 was ...
Passage 3: The Junior Eurovision Song Contest 2010 was ...

3. VerifAI

VerifAI

American Idol

Question:
Which contest did the singer of the song Mama's Song win?

Ground Truth: American Idol
• Given a query, find
• textual files
• table (tuples)
• (sub)graphs
• images
• HTML files
• or a combination thereof

How to align?
“match” -> similar vectors
query Q → Encoder

data D → Encoder

(D, Q) match closer

(D, Q) do not match farther
Dense Passage Retriever (DPR) for Text-Text

\[ \text{sim}(q, p) = \text{enc}_1(p) \cdot \text{enc}_2(q) \]
Dense Passage Retriever (DPR) for Tuple-Tuple

\[ \text{sim}(t, s) = \text{enc}_1(t) \cdot \text{enc}_2(s) \]
Dense Passage Retriever (DPR) for Text-Table

\[ \text{sim}(q, T) = \text{enc}1(T) \cdot \text{enc}2(q) \]

Encoder 1

Encoder 2

table T

text q

dot-product
Retrieval over multi-modal data lakes

- We know (a little bit about) how to
  - align two encoders of the same/different data modalities
  - the main effort is to find positive/negative pairs

- We do not know how to
  - align multiple encoders
  - encode/index large tables
  - encode/index large graphs
Outline

• What: Verify the outputs of LLMs (True/False)
• How: Retrieval  Rerank
• Results
• Opportunities
Retrieval (Coarse-grained) -> Rerank (Fine-grained)
Text-Text: Cross-Encoder (monoBert)

Rerank

_tuple s_
tuple t_

Tuple-Tuple

text x

Classifier

score

Encoder
text y

Classifier

score

Encoder
tuple t
tuple s
Text-Table: Cross-Encoder-Decoder (monoT5)

Figure 1: Illustration of the T5 ranking model structure from prior work and the two variants of our proposed RankT5 model. The differences in training mechanism is not reflected in this illustration.

Note: Figure from the Google paper “RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses”
Outline

• **What:** Verify the outputs of LLMs (True/False)

• **How:** Retrieval, Rerank, Verifier

(Q, A) → Vector Database → top-K → top-k (k << K) → True/False
Outline

• **What**: Verify the outputs of LLMs (True/False)
• **How**: retrieval, rerank, verifier
• **Results**
• **Opportunities**
Verify Tuple(s) via retrieving Tuple and Text

Retrieved Evidence and Verification
**E1:** Brendan Daly (2 February 1940 - 6 July 2023) was an Irish Fianna Fáil politician. He was a long-serving Teachta Dála(TD) for the Clare constituency, a government minister, and Senator.

**Verification result:** True. **Explanation:** The attribute ‘party’ value for Brendan Daly is Fianna Fáil, which matches with the claim.

Retrieved Evidence and Verification
**E2:** members of the 20th dáil

<table>
<thead>
<tr>
<th>constituency</th>
<th>name</th>
<th>party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tipperary South</td>
<td>Brendan Griffin</td>
<td>Fine Gael</td>
</tr>
</tbody>
</table>

**Verification result:** Not Related. **Explanation:** The evidence talks about the members of the 20th Dáil and the TDs elected by constituencies Tipperary South and Clare. It mentions Brendan Griffin as a member of the 20th Dáil from Tipperary South constituency, but it does not provide any information about Brendan Daly’s party affiliation.
Verify Tuple(s) via retrieving Tuple and Text

Generative AI task: **tuple completion**

- **100** tuples with missing values
- ChatGPT’s accuracy in completion: **0.52**

Data lake: **13,796** textual files and **19,498** tables containing **269,622** tuples

<table>
<thead>
<tr>
<th>Retrieval</th>
<th>Recall@3</th>
<th>both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuple</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Verify Text using via retrieving Tables

**Claim:** In 1954 u.s. open (golf), the cash prize for tommy bolt, fred haas, and ben hogan was 960 in total.

**Ground Truth:** a false claim that should be **Refuted**

### Retrieved Evidence and Verification

#### Table E1: 1954 u.s. open (golf)

<table>
<thead>
<tr>
<th>Place</th>
<th>Player</th>
<th>Country</th>
<th>Score</th>
<th>Par</th>
<th>Money</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ed Furgol</td>
<td>United States</td>
<td>71 + 70 + 71 + 72 = 284</td>
<td>4</td>
<td>16000</td>
</tr>
<tr>
<td>2</td>
<td>Gene Littler</td>
<td>United States</td>
<td>70 + 69 + 76 + 70 = 285</td>
<td>6</td>
<td>13600</td>
</tr>
<tr>
<td>3</td>
<td>Floyd Mangrum</td>
<td>United States</td>
<td>72 + 71 + 72 + 71 = 286</td>
<td>6</td>
<td>1500</td>
</tr>
<tr>
<td>3</td>
<td>Dick Mayer</td>
<td>United States</td>
<td>72 + 71 + 70 + 73 = 286</td>
<td>6</td>
<td>1500</td>
</tr>
<tr>
<td>5</td>
<td>Bobby Locke</td>
<td>South Africa</td>
<td>74 + 70 + 74 + 70 = 288</td>
<td>8</td>
<td>960</td>
</tr>
<tr>
<td>6</td>
<td>Tommy Bolt</td>
<td>United States</td>
<td>72 + 72 + 73 + 72 = 289</td>
<td>9</td>
<td>570</td>
</tr>
<tr>
<td>6</td>
<td>Fred Haas</td>
<td>United States</td>
<td>73 + 73 + 71 + 72 = 289</td>
<td>9</td>
<td>570</td>
</tr>
<tr>
<td>6</td>
<td>Ben Hogan</td>
<td>United States</td>
<td>71 + 70 + 76 + 72 = 289</td>
<td>9</td>
<td>570</td>
</tr>
<tr>
<td>6</td>
<td>Shelley Mayfield</td>
<td>United States</td>
<td>73 + 75 + 72 + 69 = 289</td>
<td>9</td>
<td>570</td>
</tr>
<tr>
<td>6</td>
<td>Billy Joe Patton</td>
<td>United States</td>
<td>69 + 76 + 71 + 73 = 289</td>
<td>9</td>
<td>101</td>
</tr>
</tbody>
</table>

**Verification result:** **Refuted.** **Explanation:** The cash prize for Tommy Bolt, Fred Haas, and Ben Hogan was $570 each, totaling $1710.

#### Table E2: 1959 u.s. open (golf)

<table>
<thead>
<tr>
<th>Player</th>
<th>Country</th>
<th>Year(s)</th>
<th>Total</th>
<th>Par</th>
<th>Finish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ben Hogan</td>
<td>United States</td>
<td>1948, 1950, 1951, 1953</td>
<td>287</td>
<td>7</td>
<td>T8</td>
</tr>
<tr>
<td>Cary Middlecoff</td>
<td>United States</td>
<td>1949, 1956</td>
<td>294</td>
<td>14</td>
<td>T19</td>
</tr>
<tr>
<td>Jack Fleck</td>
<td>United States</td>
<td>1955</td>
<td>294</td>
<td>14</td>
<td>T19</td>
</tr>
<tr>
<td>Julius Boros</td>
<td>United States</td>
<td>1952, 1957</td>
<td>17</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>Tommy Bolt</td>
<td>United States</td>
<td>1958</td>
<td>301</td>
<td>21</td>
<td>T38</td>
</tr>
</tbody>
</table>

**Verification result:** **Not related.**
Verify Text using via retrieving Tables

A controlled study: 1,300 textual claims from TabFact benchmark
Retrieved from 19K Wikipedia Tables

<table>
<thead>
<tr>
<th>Retrieval</th>
<th>Recall@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table</td>
<td>0.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verifier</th>
<th>Reference</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChatGPT</td>
<td>Reference</td>
<td>0.54</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>Tables</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Outline

• What: Verify the outputs of LLMs (True/False)
• How: retrieval, rerank, verifier
• Results
• Opportunities
Opportunities (DB)

1. Table learning using LLMs
2. Table reasoning using LLMs
3. Table retrieval (small tables, large tables, database tables)
4. Cross-modal data discovery (aligning multiple encoders)
5. Cross-modal matching
6. Cross-modal reranking
7. Cross-modal reasoning