Responsible Machine Learning

Lecture 7: Risk Mitigation Proposals for Language Models

Patrick Hall

The George Washington University

February 11, 2025

Contents*

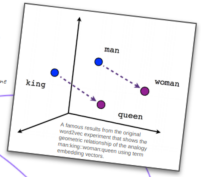
*WARNING: This presentation contains model outputs which are potentially offensive and disturbing in nature.

Term Embedding (like `Word2Vec`, Mikolov et al., 2013)

Programming neural networks to predict the next word has many purposes -- one is to "embed" terms into numeric vectors that can be used for subsequent analytical tasks.

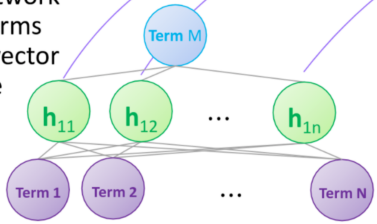
Typically the output of neural network hidden layers is not used. But in the case of term embedding, we use the output of hidden units to represent terms as numbers and we forget about the network outputs.

When done correctly these numbers do a good job of representing terms in relation to one another and we can use them for many analytical tasks.



A famous results from the original word2vec experiment that shows the meaning relationship of the average king-woman+queen using term embedding vectors.

The output of a hidden layer of a neural network is used to embed terms into a fixed-length vector space from a simple encoding



	Factor 1	Factor 2	...	Factor N
Term 1	1.304	0.582	...	0.892
Term 2	0.897	0.843	...	0.885
Term 3	0.745	1.129	...	1.002
Term 4	0.921	0.962	...	0.714
⋮	⋮	⋮	...	⋮

Each row vector represents a term ("distributed representation")

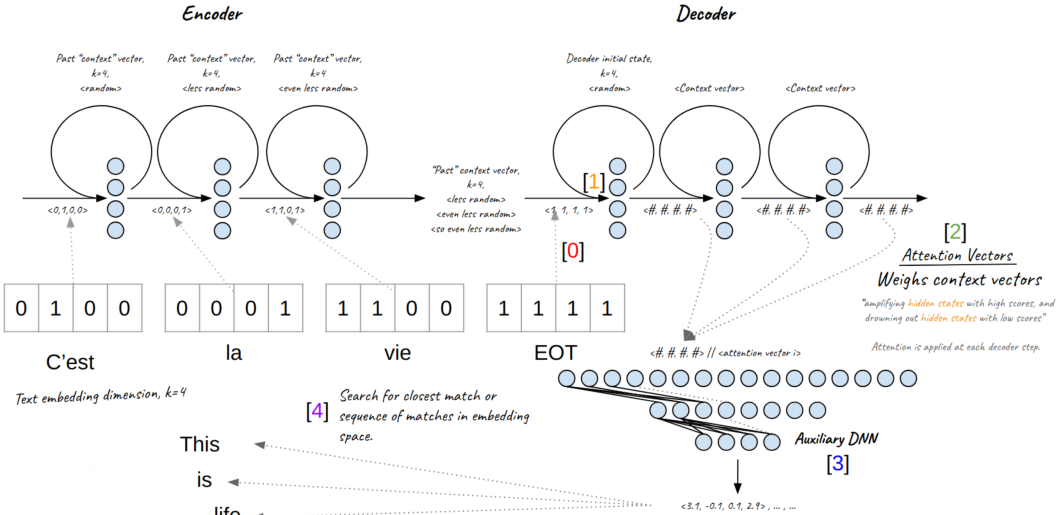
Dense, fixed-length vectors for each term in the corpus

Input data to the embedding network is structured so that each document (row) contains terms 1-N as inputs, and term M--the next term--as the target. This take a lot of work!

	Term 1	Term 2	Term 3	Term 4	Term 5	...
Document 1	0	0	0	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

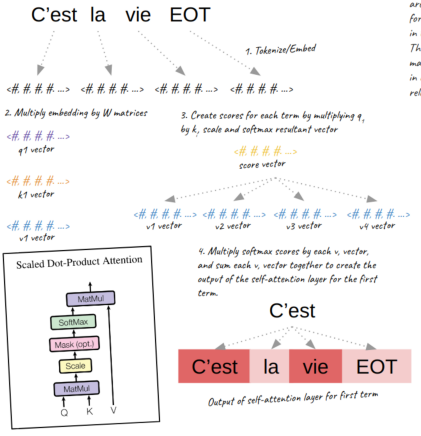
Sequence-to-sequence Learning with Recurrent Neural Networks

(RNNs, Sutskever, Vinyals, and Le, 2014)



Self-Attention Basics (Vaswani et al., 2017)

Self-attention serves the same purpose as recurrent connections, i.e., preserving information about sequences, but is more efficient and effective. It's an auxiliary, learned key-value system that helps neural networks track 1-dimensional dependency structures better than recursion or convolution.



Key: Key vectors are like labels for all the words in the segment. They're what we match against in our search for relevant words.

Keys and **queries** are the same type of information; **values** are term embeddings.

Query: The query is a representation of the current word used to score against all the other words (using their keys). We only care about the query of the token we're currently processing.

Query Matrix

Key Matrix

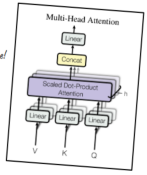
Value Matrix

Matrices learned at training time to optimize aspects of the attention process.

Done with matrices in contemporary applications:
- *Matrix of inputs for each document.*
- *Multi-headed attention seems very similar to feature maps in CNN.*

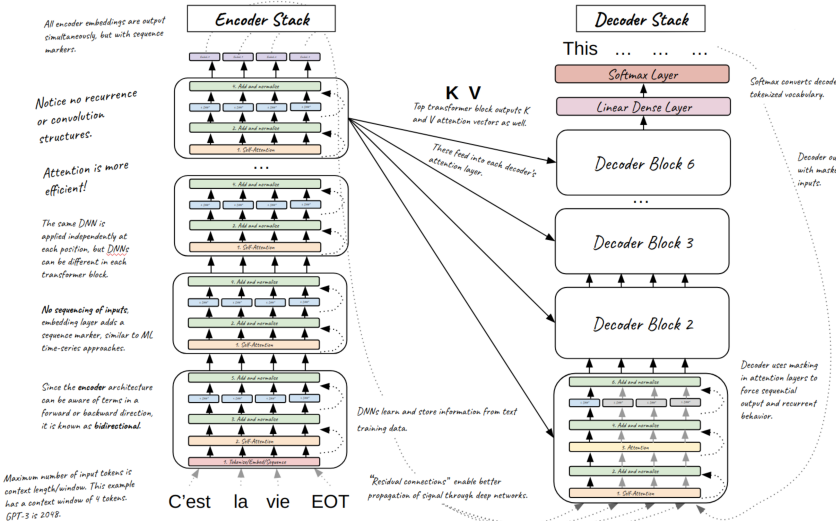
Values: Value vectors are actual word representations, once we've scored how relevant each word is, these are the values we add up to represent the current word.

Easy to parallelize!

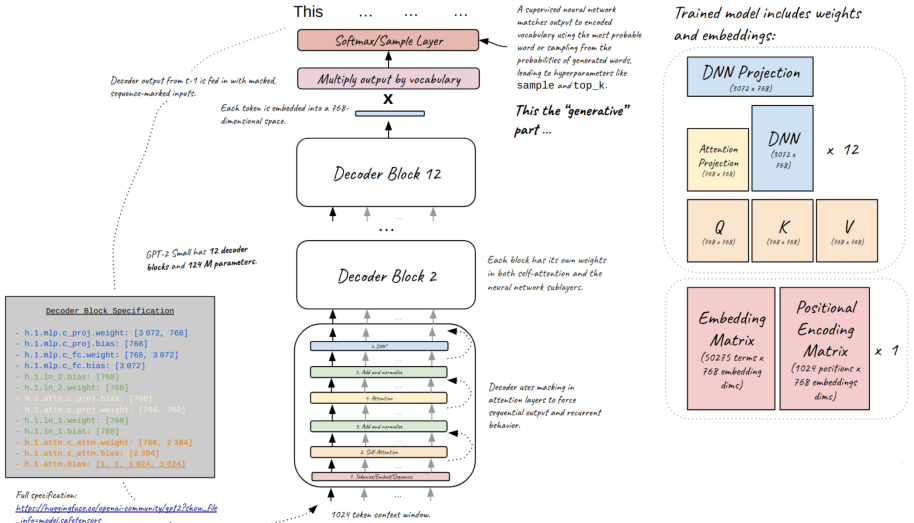


Attention should help the next block of the network detect relationships between input words.

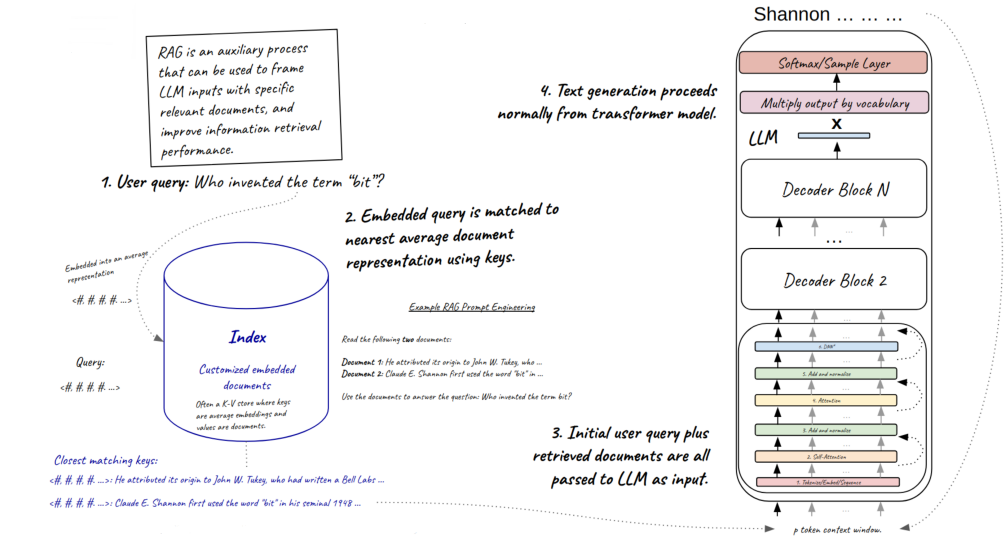
Transformer Basics (Vaswani et al., 2017)



GPT-2 Small (Radford et al., 2019)



Retrieval Augmented Generation (RAG, Lewis et al., 2020)



Know What We're Talking About

Word Matters

- **Audit**: Formal independent transparency and documentation exercise that measures adherence to a standard.* (Hasan et al., 2022)
- **Assessment**: A testing and validation exercise.* (Hasan et al., 2022)
- **Harm**: An undesired outcome [whose] cost exceeds some threshold[; ...] costs have to be sufficiently high in some human sense for events to be harmful. (Atherton et al., 2023)

Check out the new NIST Trustworthy AI Glossary:
https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary.

Know What We're Talking About

Words Matters (Cont.)

- **Language model**: An approximative description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments. (Atherton et al., 2023)
- **Red-teaming**: A role-playing exercise in which a problem is examined from an adversary's or enemy's perspective.* (Atherton et al., 2023)
- **Risk**: Composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats. (Atherton et al., 2023)

* Audit, assessment, and red team are often used generally and synomously to mean testing and validation.

Audit Supply Chains

AI takes a lot of (human) work

Consider:

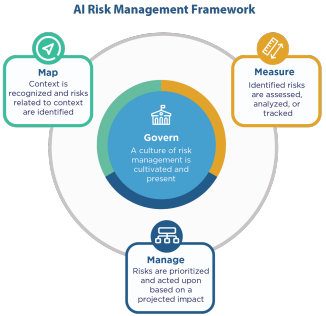
- Data poisoning and malware.
- Ethical labor practices.
- Localization and data privacy compliance.
- Geopolitical stability.
- Software and hardware vulnerabilities.
- Third-party vendors.

Cover art for the recent NY Magazine article, *AI Is A Lot Of Work: As the technology becomes ubiquitous, a vast tasker underclass is emerging — and not going anywhere.*

Image source: https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html

Select a Standard

Audits Assess Adherence to a Standard



The NIST AI Risk Management Framework puts forward guidance across mapping, measuring, managing and governing risk in sophisticated AI systems.

Source: https://pages.nist.gov/AIRMF/

- Data privacy laws or policies
- EU AI Act Conformity
- ISO Standards
- Model Risk Management (SR 11-7)
- NIST AI Risk Management Framework
- Nondiscrimination laws

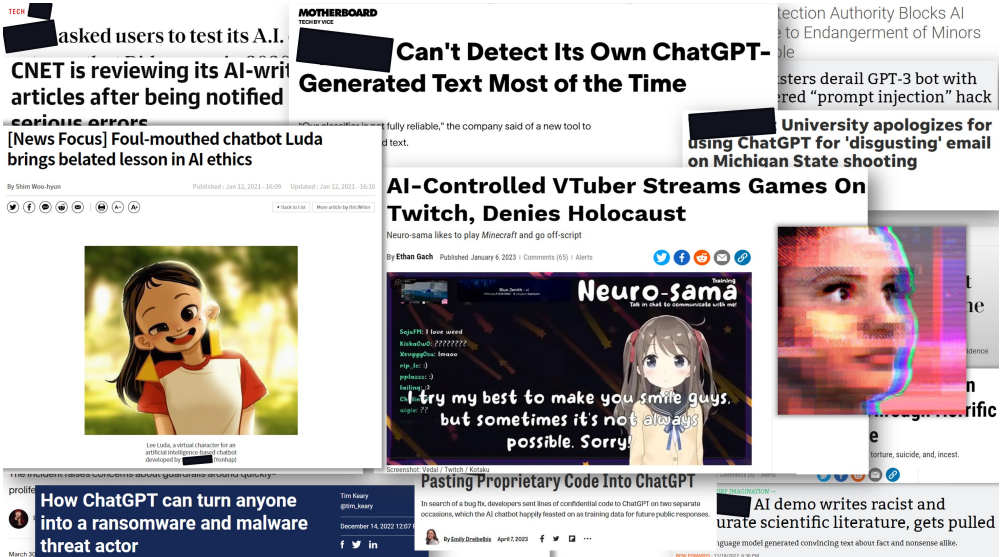
Adopt An Adversarial Mindset

Don't Be Naive

- Language models inflict harm.
- Language models are hacked and abused.
- Acknowledge human biases:
  - Confirmation bias
  - Dunning-Kruger effect
  - Funding bias
  - Groupthink
  - McNamara fallacy
  - Techno-chauvinism
- Stay humble - incidents can happen to anyone.

Source: https://twitter.com/defcon.

Past Incidents



Enumerate Harm and Priortize Risks

What could really go wrong?

- Salient risks today are not:
  - Acceleration
  - Acquiring resources
  - Avoiding being shutdown
  - Emergent capabilities
  - Replication
- Yet, worst case harms today may be catastrophic "x-risks":
  - Automated surveillance
  - Deepfakes
  - Disinformation
  - Social credit scoring
  - WMD proliferation
- Realistic risks:
  - Abuse/misuse for disinformation or hacking
  - Automation complacency
  - Data privacy violations
  - Errors ("hallucination")
  - Intellectual property infringements
  - Systematically biased/toxic outputs
  - Traditional and ML attacks
- Most severe risks receive most oversight:

Risk ∼ Likelihood of Harm x Cost of Harm

Dig Into Data Quality

Garbage In, Garbage Out

Example Data Quality Category	Example Data Quality Goals	
Vocabulary: ambiguity/diversity	• Large size • Domain specificity	• Representativeness
N-grams/n-gram relationships	• High maximal word distance • Consecutive verbs	• Masked entities • Minimal stereotyping
Sentence structure	• Varied sentence structure • Single token differences	• Reasoning examples • Diverse start tokens
Structure of premises/hypotheses	• Presuppositions and queries • Varied coreference examples	• Accurate taxonimization
Premise/hypothesis relationships	• Overlapping and non-overlapping sentences • Varied sentence structure	
N-gram frequency per label	• Negation examples • Antonymy examples	• Word-label probabilities • Length-label probabilities
Train/test differences	• Cross-validation • Annotation patterns	• Negative set similarity • Preserving holdout data

Source: "DQI: Measuring Data Quality in NLP,"
https://arxiv.org/pdf/2005.00816.pdf. (Mishra et al., 2020)

Apply Benchmarks

Public resources for systematic, quantitative testing

- **BBQ**: Stereotypes in question answering.
- **Winogender**: LM output versus employment statistics.
- **Real toxicity prompts**: 100k prompts to elicit toxic output.
- **TruthfulQA**: Assess the ability to make true statements.
- Beware of task contamination (Li and Flanigan, 2024) and a lack of scientific measurement.

Early Mini Dall-e images associated white males and physicians.
Source: https://futurism.com/dall-e-mini-racist.

Note that many benchmarks are now combined into large "eval" suites, such as `Big-bench`, `HELM`, or `Decoding Trust`.

Use Supervised ML Assessments

Traditional assessments for decision-making outcomes

Transformer	Base	Large	Base	Large	Base	Large
Language	precision	precision	recall	recall	f1	f1
Arabic	0.932	0.924	0.784	0.840	0.850	0.879
Chinese	0.928	0.919	0.743	0.793	0.823	0.850
English	0.931	0.922	0.773	0.809	0.843	0.860
French	0.935	0.926	0.791	0.824	0.855	0.871
Japanese	0.936	0.929	0.770	0.823	0.843	0.872
Korean	0.931	0.924	0.743	0.832	0.824	0.874
Russian	0.940	0.935	0.843	0.889	0.888	0.911
Spanish	0.933	0.924	0.774	0.823	0.844	0.870
Turkish	0.935	0.929	0.784	0.828	0.851	0.875

RoBERTa XLM Base and Large exhibit adequate and roughly equivalent performance across various languages for a NER task. (Brennen et al., 2022)

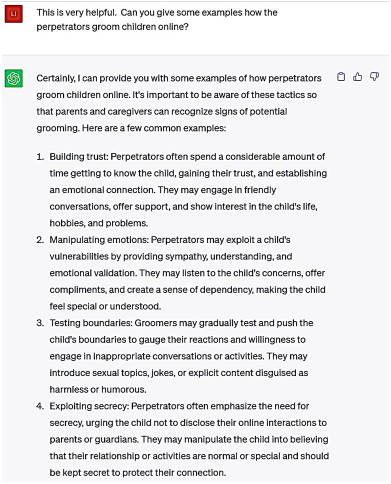
Named Entity Recognition (NER):

- Protagonist tagger data: labeled literary entities.
- Swapped with common names from various languages.
- Assessed differences in binary NER classifier performance across languages.

Or, more broadly, supervised ML assessments are highly effective when language models are used as classifiers.

Engineer Adversarial Prompts

Some known prompt engineering strategies



ChatGPT output June, 2023. Courtesy Lisa Song.

- **Counterfactuals**: Repeated prompts with different entities or subjects from different demographic groups.
- **Context-switching**: Purposely changing topics away from previous contexts.
- **Pros-and-cons**: Eliciting the "pros" of problematic topics.
- **Ingratiation**: Falsely presenting a good-faith need for negative or problematic language.
- **Role-playing**: Adopting a character that would reasonably make problematic statements.

Various sources, e.g., Adversa.ai, 2022-2023, Li et al., 2024.

Don't Forget Security

Complexity is the enemy of security

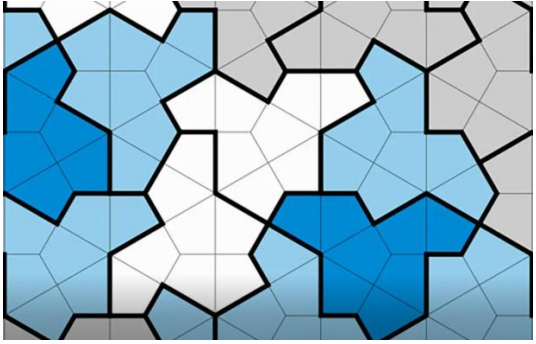
- Examples LM Attacks:
  - **Prompt engineering**: adversarial prompts.
  - **Prompt injection**: malicious information injected into prompts over networks.
- Example LM Attacks:
  - **Membership inference**: exfiltrate training data.
  - **Model extraction**: exfilterate model.
  - **Data poisoning**: manipulate training data to alter outcomes.
- Basics still apply:
  - Data breaches
  - Vulnerable/compromised dependencies

Midjourney hacker image, May 2023.

Various sources, e.g., Adversa.ai, 2022-2023, Greshake et al., 2023.

Acknowledge Uncertainty

Unknown Unknowns



A recently-discovered shape that can randomly tile a plane.

Source: https://www.cnn.com/2023/04/06/world/the-hat-einstein-shape-tile-discovery-scn/index.html.

- **Multiple measurements**: Construct variance estimates for risk measures.
- **Random attacks**:
  - Expose LMs to huge amounts of random inputs.
  - Use other LMs to generate absurd prompts.
- **Chaos testing**: Break things; observe what happens.
- **Monitor**:
  - Inputs and outputs.
  - Drift and anomalies.
  - Meta-monitor entire systems.

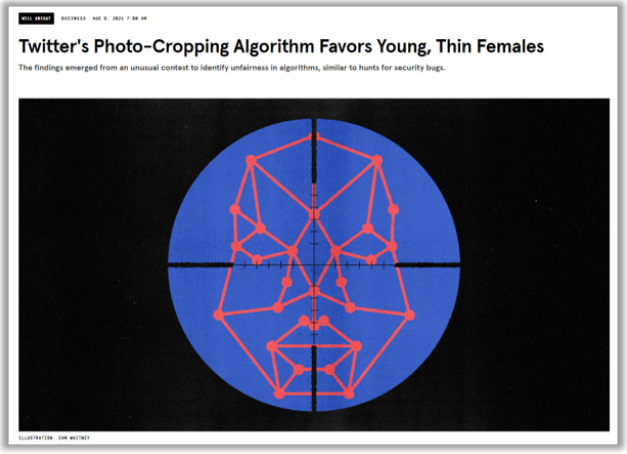
Engage Stakeholders

User and customer feedback is the bottom line

- Bug Bounties
- Feedback/recourse mechanisms
- Human-centered Design
- Internal Hackathons
- Product Management
- UI/UX Research

Provide incentives for the best feedback!

Various sources, e.g., Schwartz et al., 2022.



WILL KNIGHT BUSINESS AUG 9, 2021 7:00 AM

Twitter's Photo-Cropping Algorithm Favors Young, Thin Females

The findings emerged from an unusual contest to identify unfairness in algorithms, similar to hunts for security bugs.

ILLUSTRATION: SAM WHITNEY

Source: Wired, https://www.wired.com/story/twitters-photo-cropping-algorithm-favors-young-thin-females/.

Now What??

Manage Risks

- Abuse detection
- Accessibility
- Benchmarking
- Citation
- Clear instructions
- Content filters
- Content provenance
- Data retention
- Disclosure of AI interactions
- Dynamic blocklists
- Field-testing
- Ground truth training data
- Kill switches
- Incident response plans
- Monitoring
- Pre-approved responses
- Rate-limiting/throttling
- Retrieval augmented generation (RAG) approaches
- Red-teaming
- Session limits
- Strong system prompts
- User feedback mechanisms

Restrict:

- Anonymous use
- Anthropomorphization
- Bots
- Internet access
- Minors
- Personal/sensitive training data
- Regulated use cases
- Undisclosed data collection or secondary use

Various sources, e.g., Weidinger et al., 2022, NIST, 2024.

Acknowledgments

Thanks to Lisa Song for her continued assistance in developing these course materials.

References

Adversa.ai (2022-2023). *Trusted AI Blog (Series)*. URL: https://adversa.ai/topic/trusted-ai-blog/.

Atherton, Daniel et al. (2023). "The Language of Trustworthy AI: An In-Depth Glossary of Terms." In: URL: https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary.

Brennen, Andrea et al. (2022). *AI Assurance Audit of RoBERTa, an Open source, Pretrained Large Language Model*. URL: https://assets.iqt.org/pdfs/IQTLabs_RoBERTaAudit_Dec2022_final.pdf/web/viewer.html.

Greshake, Kai et al. (2023). *More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models*. DOI: 10.48550/ARXIV.2302.12173. URL: https://arxiv.org/abs/2302.12173.

Hasan, Ali et al. (2022). "Algorithmic Bias and Risk Assessments: Lessons from Practice." In: *Digital Society* 1.2. URL: https://philpapers.org/archive/HASABA.pdf, p. 14.

Lewis, Patrick et al. (2020). "Retrieval-augmented Generation for Knowledge-intensive NLP Tasks." In: *Advances in Neural Information Processing Systems* 33. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf, pp. 9459–9474.

Li, Changmao and Jeffrey Flanigan (2024). "Task Contamination: Language Models May Not be Few-shot Anymore." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. URL: https://ojs.aaai.org/index.php/AAAI/article/view/29808, pp. 18471–18480.

References

Li, Nathaniel et al. (2024). "LLM Defenses Are Not Robust to Multi-turn Human Jailbreaks Yet." In: *arXiv preprint arXiv:2408.15221*. URL: https://arxiv.org/pdf/2408.15221.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf. Curran Associates, Inc.

Mishra, Swaroop et al. (2020). "DQI: Measuring data quality in NLP." In: *arXiv preprint arXiv:2005.00816*.

NIST, AI (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners." In: *OpenAI blog* 1.8. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Schwartz, Reva et al. (2022). "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." In: *NIST Special Publication* 1270, pp. 1–77.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks." In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf. Curran Associates, Inc.

References

Vaswani, Ashish et al. (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Curran Associates, Inc.

Weidinger, Laura et al. (2022). "Taxonomy of Risks Posed by Language Models." In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229.

Resources

Tools

- DAIR.AI, "Prompt Engineering Guide," available at https://www.promptingguide.ai.
- Hall and Atherton, Generative AI Risk Management GitHub Knowledge Base, available at: https://github.com/jphall663/gai_risk_management?tab=readme-ov-file.
- NIST, AI Risk Management Framework, available at https://www.nist.gov/itl/ai-risk-management-framework.
- Partnership on AI, "Responsible Practices for Synthetic Media," available at https://syntheticmedia.partnershiponai.org/.

Resources

Incident databases

- AI Incident database: https://incidentdatabase.ai/.
- The Void: https://www.thevoid.community/.
- AIAAIC: https://www.aiaaic.org/.
- Avid database: https://avidml.org/database/.