Pengembangan Sistem Tinjauan Makalah dengan Large Language Models

UJIAN KUALIFIKASI

Utami Lestari

99223141

PROGRAM DOKTOR TEKNOLOGI INFORMASI

UNIVERSITAS GUNADARMA

2024

Daftar Isi

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi telah menjadi pendorong utama transformasi digital yang telah mengubah pola kehidupan manusia secara fundamental. Teknologi informasi telah menyebar ke berbagai sektor serta memfasilitasi akses ke berbagai informasi, mempercepat pertukaran data, dan mengubah cara manusia berinteraksi. Saat ini teknologi bukan hanya sekedar kebutuhan sekunder namun dapat dikatakan sebagai kebutuhan primer. Perkembangan teknologi mempengaruhi berbagai aspek kehidupan seperti sosial,politik dan budaya. Salah satu bentuk dari perkembangan teknologi adalah kecerdasan buatan atau *artificial intelligence* (AI).

Kecerdasan buatan atau AI mengacu pada aplikasi algoritma dan teknik perangkat lunak yang memungkinkan komputer dan mesin untuk mensimulasikan persepsi manusia dan proses pengambilan keputusan untuk menyelesaikan tugas dengan sukses(Murphy, 2019). Teknologi kecerdasan buatan (AI) berkembang mengiringi perkembangan teknologi informasi, hal ini membawa perubahan diberbagai bidang seperti pendidikan, kesehatan, ekonomi, industry, dan tranportasi. Teknologi AI menawarkan berbagai kemudahan, efisiensi, analisis data yang mendalam dan pengenalan pola. Perkembangan kecerdasan buatan (AI) telah memainkan peran krusial dalam menghadirkan kemajuan signifikan diberbagai aplikasi. Melalui teknik-teknik seperti deep learning dan machine learning, AI mampu menganalisis data yang kompleks dan mengidentifikasi pola-pola yang tersembunyi di dalamnya. Pada pengenalan gambar misalnya, AI telah mencapai tingkat ketepatan yang mengesankan dalam mengenali objek, wajah, atau bahkan pola-pola mikroskopis dalam citra medis. pada bidang teks, AI dapat mengenali pola-pola kompleks dalam teks seperti struktur gramatikal, entitas, dan makna kata. AI juga dapat mengidentifikasi pola-pola seperti opini atau sentimen dalam teks, topik pembicaraan, atau bahkan pemahaman konteks dari suatu kalimat.

Pertumbuhan pesat penelitian ilmiah dan produksi artikel ilmiah dalam berbagai disiplin ilmu menghasilkan tantangan baru dalam proses tinjauan artikel ilmiah. Volume besar publikasi membuat tugas penelahaah sejawat *(peer-review)* semakin kompleks dan memakan waktu yang lebih banyak. Jumlah naskah yang diajukan ke jurnal untuk proses *peer-review* mengalami pertumbuhan tahunan sebesar 6,1% (Checco et al., 2021). Sementara itu, kebutuhan untuk memastikan kualitas dan keakuratan penilaian tetap menjadi prioritas. Adanya pertumbuhan pengajuan naskah untuk di tinjau menimbulkan potensi bias dan konsistensi penelaahan yang kurang baik. Sehingga terdapat peluang untuk mengembangkan suatu model yang mampu melakukan tinjauan secara mandiri dengan bantuan kecerdasan buatan.

Pada saat yang sama, perkembangan kecerdasan buatan dalam bidang pemrosesan teks atau yang dikenal dengan *Narutal Language Processing(NLP)* telah membuka peluang baru untuk otomatisasi beberapa aspek dari proses tinjauan artikel ilmiah. Kemampuan algoritma tersebut dalam memahami bahasa alami, analisis konten, dan penyajian informasi dapat dimanfaatkan untuk mempermudah dan mempercepat proses tinjauan. Pemrosesan bahasa alami (NLP) adalah subbidang kecerdasan buatan dan linguistik komputasi. Bidang ini berfokus pada kemampuan komputer untuk memahami, menafsirkan, dan menghasilkan bahasa manusia dengan cara yang bermakna dan berguna (Amaratunga, 2023). Salah satu pendekatan dalam pemrosesan bahasa alami yang memungkinkan untuk melakukan tugas mendalam dengan data besar adalah model bahasa besar atau *Large Language Models (LLM)*.

Model Bahasa besar adalah model bahasa yang telah dilatih sebelumnya dengan ukuran parameter jauh lebih besar dibandingkan dengan pendekatan lainnya (Liu et al., 2024). Model bahasa besar adalah hasil dari kombinasi pemrosesan bahasa alami, konsep pembelajaran mendalam, dan model kecerdasan buatan generative. Akhir-akhir ini model bahasa besar telah memberikan terobosan yang signifikan terutama yang berkaitan dengan transformer. Hal ini mencangkup peningkatan komputasi dan ketersedian pelatihan data dengan skala yang besar.

Perkembangan tersebut telah menghasilkan transformasi revolusioner dengan memberikan kemungkinan pembuatan LLM yang dapat mendekati kinerja manusia pada berbagai tugas (Naveed et al., 2023).

Beberapa penelitian sebelumnya membahas mengenai penelaahan sejawat dengan kecerdasan buatan seperti yang ditulis oleh Alessandro, Lorenzo, Pierpaolo, Stephen & Giuseppe pada tahun 2021 menunjukan bahwa kecerdasan buatan mampu melakukan penelaahan sejawat dan memprediksi sesuai dengan hasil ahli(Checco et al., 2021). Penelitian lainnya dilakukan oleh Iddo Drori & Dov Te'eni tahun 2024 menunjukan kecerdasan buatan khususnya model Bahasa besar mampu melakukan tugas telaah sejawat dengan cukup baik (Drori & Te’eni, 2024).

Penelitian ini bertujuan untuk mengembangkan model kecerdasan buatan untuk tinjauan artikel berbasis large language models dengan fokus pada peningkatan efisiensi dan kecepatan proses telaah sejawat, tanpa mengorbankan kualitas dan keakuratan penilaian ilmiah. Model ini diharapkan dapat membantu dalam identifikasi aspek-aspek kritis, serta memberikan analisis otomatis yang dapat digunakan sebagai dasar untuk penilaian lebih lanjut. Pengembangan model-tools ini diharapkan dapat mempercepat alur kerja tinjauan artikel, meminimalkan potensi kesalahan manusia, dan pada akhirnya dapat meningkatkan efisiensi proses telaah sejawat.

1.2 Batasan dan Rumusan Masalah

1.2.1 Batasan Masalah

Untuk mencegah meluasnya permasalahan dalam domain yang diteliti penulis membuat Batasan agar fokus pada penyelesaian masalah dapat tercapai. Adapun Batasan masalah dari penelitian ini adalah sebagai berikut :

1. Penelitian berfokus pada pengembangan model platform tinjauan artikel ilmiah dengan menggunakan LLM
2. Model LLM yang digunakan adalah GPT-4
3. Model difokuskan pada artikel ilmiah dalam disiplin ilmu komputer

1.2.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah disampaikan diperoleh beberapa permasalahan yang harus di selesaikan. Permasalahan tersebut dirumuskan sebagai berikut :

1. Bagaimana membangun model platform tinjauan artikel ilmiah dengan LLM?
2. Bagaimana algoritma untuk klasifikasi artikel ilmiah untuk publikasi jurnal?

1.3 Tujuan Penelitian

Secara umum penelitian ini bertujuan untuk menciptakan platform tinjauan artikel ilmiah berbasis kecerdasan buatan dengan menggunakan model bahasa besar. Tujuan penelitian secara khusus adalah sebagai berikut :

1. Membangun model platform tinjauan artikel ilmiah dengan LLM
2. Mengembangkan algoritma untuk klasifikasi artikel ilmiah untuk publikasi jurnal

1.4 Kontribusi dan Manfaat Penelitian

Kontribusi dalam bidang akademik adalah tersedianya model platform tinjauan artikel ilmiah dengan LLM dan algoritma untuk klasifikasi artikel ilmiah untuk publikasi jurnal. Kontribusi penelitian pada bidang teknologi adalah tersedianya platform tinjauan artikel ilmiah yang berbasis teknologi kecerdasan buatan.

Manfaat dari hasil penelitian ini:

1. Memudahkan para penelaah sejawat dalam melakukan tinjauan artikel.
2. Menciptakan standarisasi dalam penelaahan sejawat.
3. Memberikan efisiensi waktu dalam penelaahan sejawat.

BAB II
TELAAH PUASTAKA

2.1 Pemrosesan Bahasa Alami

Pemrosesan Bahasa Alami atau *Natural Language Processing (NLP)* adalah cabang dari ilmu komputer dan kecerdasan buatan yang bertujuan untuk mengajarkan mesin agar dapat memahami dan memproses bahasa manusia secara efektif dan akurat(Vivi P Ratung, 2023). Teknik NLP digunakan dalam berbagai konteks, seperti pengenalan suara, penerjemahan bahasa, analisis sentimen, pengembangan chatbot, serta pembuatan ringkasan teks, dan sebagainya. NLP menggabungkan prinsip linguistik, statistik, dan pembelajaran mesin untuk membantu komputer dalam memahami bahasa manusia dan memberikan respons yang sesuai dengan input yang diberikan. Secara tradisional, pekerjaan dalam pemrosesan bahasa alami cenderung melihat proses analisis bahasa sebagai suatu yang dapat diuraikan menjadi beberapa tahap yang mencerminkan perbedaan linguistik teoritis yang ditarik antara sintaksis, semantik dan pragmatik.

Sistem pemrosesan bahasa alami sering dijuluki sebagai "pipeline" karena umumnya melibatkan beberapa tahapan pemrosesan. Bahasa alami mengalir dari satu titik ke titik lainnya dalam sistem ini. Bahasa alami berisi informasi atau instruksi yang dapat diekstraksi tetapi tidak langsung diterjemahkan menjadi serangkaian operasi matematika. Informasi dan instruksi ini dapat disimpan, diindeks, dicari, dan dijalankan untuk merespons pertanyaan. Ini adalah salah satu fungsi utama dari pemrosesan bahasa alami(Hannes Hapke et al., 2019) .

Menjembatani antara manusia dengan computer maka terdapat beberapa tugas-tugas dari NLP yaitu:

1. Klasifikasi teks : Menetapkan label atau kategori pada sebuah teks. Misalnya, mengklasifikasikan email sebagai spam atau bukan spam, analisis sentimen (mengidentifikasi sentimen sebagai positif, negatif, atau netral), kategorisasi topik, dll.
2. Terjemahan mesin: Menerjemahkan teks secara otomatis dari satu bahasa ke bahasa lain.

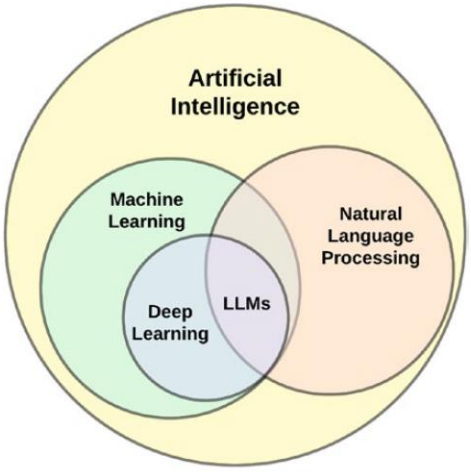
3. Pembuatan teks: Membuat teks seperti manusia, yang dapat berupa chatbot, konten yang dibuat secara otomatis, atau ringkasan teks.
4. Pengenalan ucapan: Mengubah bahasa lisan menjadi teks tertulis.
5. Peringkasan teks: Secara otomatis menghasilkan ringkasan yang ringkas dan koheren dari teks yang lebih panjang.
6. Pemodelan bahasa: Memprediksi kemungkinan munculnya urutan kata tertentu dalam suatu bahasa.

Konsep dasar dari NLP untuk mencapai tugas-tugas yang disebutkan sebelumnya, NLP menggunakan serangkaian konsep utama. Ini adalah beberapa yang paling umum:

a. Tokenisasi : proses memecah kalimat menjadi unit-unit yang lebih kecil. biasanya berupa kata atau subkata. Unit-unit yang lebih kecil ini disebut token, dan tokenisasi adalah langkah preprocessing yang penting dalam sebagian besar tugas NLP. Misalnya kata “budi sedang bermain bola” yang akan di pecah menjadi [“Budi”,”sedang”,”bermain”,”bola”]
b. Penghapusan *Stopword* : Stopword adalah kata-kata umum yang sering muncul dalam teks tetapi memiliki makna semantik yang kecil. Menghapus stopword dapat membantu mengurangi noise dan meningkatkan efisiensi komputasi.
c. *Stemming dan Lemmatization* : Stemming dan lemmatization adalah teknik yang digunakan untuk mengurangi kata menjadi bentuk dasar.

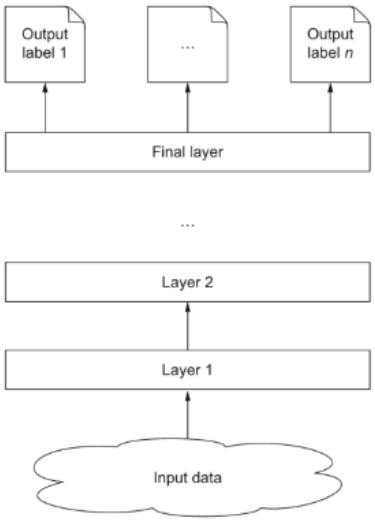
2.2 Deep Learning

Deep learning merupakan bidang penelitian baru dalam pembelajaran mesin yang bertujuan untuk meniru kemampuan otak manusia dalam mengolah dan mempelajari data masukan yang kompleks serta menyelesaikan tugas-tugas rumit dengan baik. Penggunaan deep learning telah berhasil diterapkan dalam berbagai bidang, seperti pengolahan gambar, suara, teks, dan gerak. Kemajuan teknik yang berasal dari penelitian deep learning telah memberikan dampak signifikan pada pengembangan Natural Language Processing (NLP), yang merupakan proses pemrosesan bahasa alami (Du & Shanker, n.d.).



Gambar 2. 1 Cabang Ilmu Artificial Intelligence

Pada gambar 2.1 menunjukan cabang ilmu kecerdasan buatan yang mencakup pembelajaran mesin, pemrosesan Bahasa alami, deep learning dan model Bahasa besar.



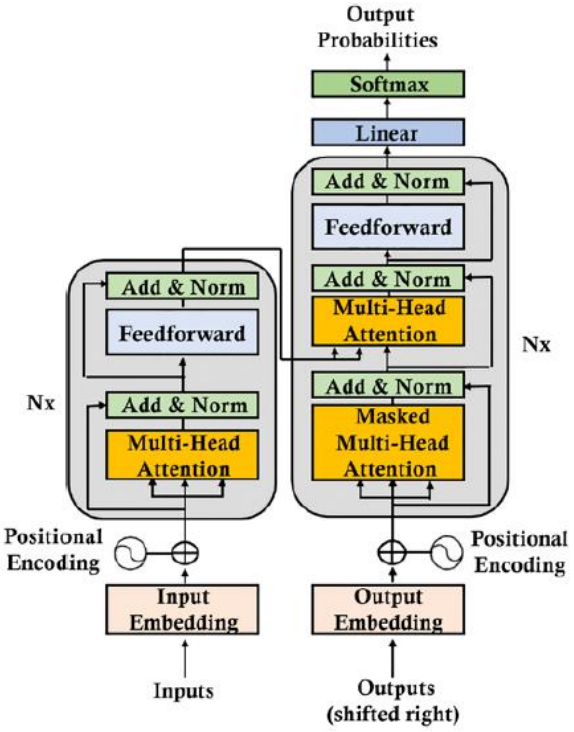
Gambar 2. 2 Arsitektur Umum Deep Learning

Lapisan keluaran, akhirnya, menghasilkan sebuah hasil: label yang diberikan model pada input το τα. Biasanya, jaringan menghasilkan kemampuan probabilitas untuk satu set hasil yang mungkin. Hasil dengan probabilitas tertinggi kemudian menjadi label keluaran akhir. Semua lapisan kecuali lapisan input dan output adalah

lapisan tersembunyi, karena mereka tidak dapat diamati dengan mudah. Seperti yang telah disebutkan, lapisan tersembunyi dalam neural net bekerja menguraikan lapisan data input yang tidak dapat dipisahkan secara linier, langkah demi langkah (Stephan Raaijmakers, 2022).

2.3 Transformers

Transformer adalah model pembelajaran yang dirancang untuk komputasi paralel pada superkomputer dengan homogenisasi dan dihomogenisasi. Melalui proses homogenisasi, satu model transformer dapat menangani berbagai tugas tanpa memerlukan penyesuaian khusus. Dengan kemampuan ini, Transformer dapat melakukan pembelajaran sendiri pada data mentah yang tidak berlabel dalam jumlah besar, menggunakan miliaran parameter (Rothman & Gulli, n.d.).



Gambar 2. 3 Arsitektur Transformer

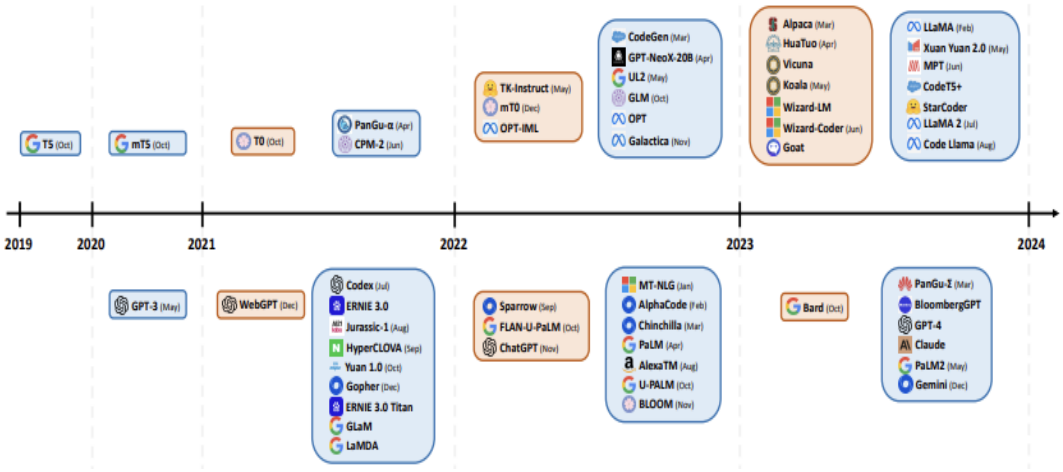
Pada gambar 2.3 di sebelah kiri, input masuk ke sisi *encoder* Transformer melalui *attention sublayer* dan *feedforward sublayer*. Di sebelah kanan, output target masuk ke sisi dekoder Transformer melalui dua *attention sublayer* dan *feedforward*

sublaye. Pada arsitektur ini tidak ada lagi RNN, LSTM, atau CNN. *Recurrence* telah ditinggalkan dalam arsitektur ini(Rothman & Gulli, n.d.).

Attention telah menggantikan fungsi recurrence yang membutuhkan parameter yang meningkat seiring bertambahnya jarak antara dua kata meningkat. Mekanisme attention adalah operasi "kata ke kata". Ini sebenarnya adalah operasi token-ke-token. Mekanisme attention akan menemukan bagaimana setiap kata berhubungan dengan semua kata lain dalam sebuah urutan,termasuk kata yang sedang dianalisis itu sendiri.

2.4 Model Bahasa Besar (LLM)

Large language models adalah model-model bahasa yang memiliki ukuran dan kompleksitas yang besar, terdiri dari jutaan atau bahkan miliaran parameter. Model-model ini dirancang untuk memahami dan menghasilkan bahasa manusia dengan tingkat keterampilan yang tinggi. LLM dibangun dengan menggunakan teknik-teknik deep learning, terutama menggunakan arsitektur seperti Transformer, yang memungkinkan LLM untuk menangani sejumlah besar data dan menangkap pola-pola kompleks dalam bahasa manusia. Large language models telah menjadi fokus utama dalam penelitian NLP karena kemampuan mereka untuk melakukan berbagai tugas, seperti penerjemahan bahasa, generasi teks, analisis sentiment dan sebagainya.



Gambar 2. 4 Model Bahasa Besar(LLM)

Pelatihan model bahasa besar (LLM) pada dataset yang sangat besar telah mengubah secara signifikan pemrosesan bahasa alami, memungkinkan LLM untuk meniru interaksi manusia dan berperan sebagai asisten serbaguna dalam berbagai tugas. Ini termasuk memberikan jawaban yang luas terhadap pertanyaan, membantu dalam penulisan, memberikan pengajaran, dan beragam tugas lainnya(Wolf et al., 2023).

Model bahasa yang besar dapat dikategorikan berdasarkan berbagai faktor seperti arsitektur, tujuan pelatihan, tipe data, dan aplikasi (Amaratunga, 2023). Berdasarkan arsitekturnya LLM dibedakan berdasarkan beberapa kategori, yaitu:

- Transformer
- Recurrent Neural Network
- Convolutional Neural Network

Adapun komponen-kompomem penting dalam LLM :

1. Tokenization (tokenisasi)

Tokenization adalah langkah pra-pemrosesan penting dalam pelatihan model bahasa besar (LLM) yang memecah teks menjadi unit-unit tak terpisahkan yang disebut token (Naveed et al., 2023). Token dapat berupa karakter, subkata, simbol, atau kata, tergantung pada proses tokenisasi yang digunakan. Beberapa skema tokenisasi yang umum digunakan dalam LLM termasuk wordpiece, byte pair encoding (BPE), dan unigramLM (Webster & Kit, n.d.).

2. Encoding Positions

Transformer memproses urutan input secara paralel dan independen, tanpa mencatat informasi posisi. Oleh karena itu, diperkenalkan positional encoding di transformer, di mana vektor positional encoding ditambahkan ke token embedding. Varian positional embedding meliputi positional encoding absolute, relative, or learned positional encodings. Dalam relative encoding terdapat 2 positional embedding yang sering diguanakan dalam LLMs yaitu Alibi dan RoPE.

a. Alibi : bertugas mengurangi bias skalar dari attention score yang meningkat dengan Jarak antara posisi token.
b. RoPE : teknik ini membantu model untuk lebih efektif memproses dan memahami teks dengan memperhatikan hubungan relatif antar token yang lebih relevan.

3. Attention

Attention bertugas memberikan bobot kepada token-token input berdasarkan kepentingannya sehingga model memberikan penekanan lebih pada token-token yang relevan(Vaswani et al., n.d.). Attention dalam transformer menghitung pemetaan query, key, dan value untuk urutan input, di mana skor perhatian diperoleh dengan mengalikan query dan key, dan kemudian digunakan untuk memberi bobot pada nilai-nilai. Beberapa strategi attention yang digunakan dalam LLM yaitu :

a. Self-Attention
b. Cross Attention
c. Sparse Attention
d. Flash Attention

4. Activation Functions

Fungsi aktivasi memiliki peran penting dalam kemampuan kurva jaringan saraf. Fungsi aktivasi adalah komponen kunci dalam LLM dan jaringan saraf lainnya yang memungkinkan model untuk belajar pola kompleks dan non-linear dalam data. Pilihan fungsi aktivasi yang tepat dapat secara signifikan mempengaruhi kinerja dan efisiensi model. Beberapa fungsi aktivasi yang digunakan dalam LLM yaitu :

a. ReLU
b. GeLU
c. GLU varians

5. Encoder

Modul encoder dari model Transformer terdiri dari beberapa lapisan identik, masing-masing mencakup mekanisme multi-head attention dan jaringan saraf feed-

forward(Liu et al., 2024). Dalam mekanisme multi-head attention, setiap posisi dalam urutan input dihitung untuk perhatian dengan posisi lainnya guna menangkap ketergantungan antar posisi. Jaringan saraf feed-forward digunakan untuk memproses dan mengekstrak fitur dari output mekanisme perhatian. Modul encoder secara bertahap mengekstraksi fitur urutan input melalui penumpukan beberapa lapisan tersebut dan mengirimkan hasil encoding akhir ke modul decoder untuk decoding. Desain modul encoder memungkinkan penanganan ketergantungan jarak jauh dalam urutan input dan meningkatkan kinerja dalam berbagai tugas NLP.

6. Decoder

Modul decoder dari model Transformer terdiri dari beberapa lapisan identik, masing-masing mencakup mekanisme multi-head attention dan jaringan saraf feed-forward(Liu et al., 2024). Berbeda dengan encoder, decoder juga mencakup mekanisme attention tambahan antara encoder dan decoder, digunakan untuk menghitung perhatian pada urutan input selama proses decoding. Pada setiap posisi, decoder hanya dapat melakukan perhitungan self-attention dengan posisi sebelumnya untuk memastikan urutan yang dihasilkan tidak melanggar aturan tata bahasa. Mask memainkan peran penting dalam decoder, memastikan bahwa hanya informasi sebelum langkah waktu saat ini yang diperhatikan saat menghasilkan urutan output, dan tidak membocorkan informasi dari langkah waktu mendatang. Mekanisme self-attention pada decoder menggunakan mask untuk mencegah model mengakses informasi masa depan saat menghasilkan prediksi pada setiap langkah waktu, menjaga kausalitas model. Hal ini memastikan bahwa output yang dihasilkan oleh model bergantung pada informasi pada langkah waktu saat ini dan sebelumnya, tanpa dipengaruhi oleh informasi masa depan.

2.4.1 Prompt Learning

Prompt learning adalah pendekatan machine learning yang digunakan secara luas, terutama di bidang NLP. Metodologi ini melibatkan pembuatan pernyataan prompt yang hati-hati untuk mengarahkan model menghasilkan perilaku atau output tertentu. Pendekatan ini sering digunakan untuk fine-tuning dan mengarahkan LLM yang sudah dilatih sebelumnya untuk menjalankan tugas tertentu atau menghasilkan

hasil yang diinginkan(Liu et al., 2024). Desain pernyataan prompt dapat mengarahkan model pre-trained untuk melakukan berbagai tugas seperti menjawab pertanyaan, menghasilkan teks, dan memahami semantik. Kekuatan pendekatan ini terletak pada kemampuannya untuk beradaptasi dengan berbagai tugas melalui modifikasi sederhana pada pernyataan prompt, tanpa perlu melatih ulang seluruh model. Untuk LLM seperti seri GPT dan model pre-trained lainnya, prompt learning menyediakan cara yang mudah dan kuat untuk fine-tuning model. Dengan memberikan prompt yang sesuai, peneliti dan praktisi dapat menyesuaikan perilaku model agar lebih cocok untuk domain atau kebutuhan tugas tertentu. Singkatnya, prompt learning adalah pendekatan machine learning yang membangun model bahasa yang telah dilatih sebelumnya dan mengarahkannya untuk melakukan berbagai tugas melalui desain pernyataan prompt, menawarkan fleksibilitas yang meningkat untuk menyesuaikan aplikasi model.

2.4.2 Architecture

Saat ini semua Large Language Models (LLMs) dibangun dengan menggunakan arsitektur Transformer. Arsitektur ini memungkinkan model-model ini untuk memiliki skala hingga beberapa miliar atau bahkan triliun parameter. Secara umum, arsitektur Pretrained language model (PLM) dapat dikelompokkan menjadi tiga kategori: Encoder-only, Encoder-decoder, dan Decoder-only. Arsitektur Encoder-only tidak lagi digunakan dalam LLMs terbaru dan tidak akan dibahas lebih lanjut di sini. Sebagai gantinya, fokus bagian ini adalah untuk memperkenalkan arsitektur Encoder-decoder dan Decoder-only.

1. Encoder-Decoder

Arsitektur Encoder-decoder pada LLMs didasarkan pada arsitektur Transformer tradisional. Terdiri dari dua komponen utama: Encoder dan Decoder, di mana Encoder menyandikan urutan input melalui beberapa lapisan Multi-Head Self-Attention, sedangkan Decoder menggunakan cross-attention pada representasi output dari Encoder untuk menghasilkan urutan target secara autoregresif. Arsitektur ini menjadi dasar bagi LLM terkenal seperti T5, flan-T5, dan BART.

2. Decoder only

LLMs dengan arsitektur Decoder-only memanfaatkan komponen decoder dari arsitektur Transformer tradisional. Berbeda dengan arsitektur Encoder-Decoder yang menggabungkan encoder dan decoder, arsitektur Decoder-only hanya fokus pada proses dekoding. Model ini secara berurutan menghasilkan token-token dengan memperhatikan token-token sebelumnya dalam urutan. Arsitektur ini telah diterapkan dalam berbagai tugas generasi bahasa, menunjukkan efektivitasnya dalam menghasilkan teks tanpa memerlukan fase encoding eksplisit. Arsitektur Decoder-only dapat dibedakan lagi menjadi dua kategori: arsitektur Causal Decoder dan arsitektur Prefix Decoder.

a. **Arsitektur Causal Decoder**, setiap token dalam urutan input model hanya dapat memperhatikan token-token input yang terjadi sebelumnya dan dirinya sendiri selama proses dekoding. Ini mencapai perhatian unidireksional pada urutan input dengan menggunakan masker khusus. Arsitektur ini dikonfigurasi dengan matriks masker yang berbeda-beda untuk mengimplementasikan berbagai arsitektur. Arsitektur Causal Decoder adalah dasar bagi serangkaian LLM terkenal seperti seri GPT, yang dikenal karena kinerja superior mereka dan banyak diterapkan dalam LLM lain seperti BLOOM, OPT, Gopher, dan LLaMA.
b. **Arsitektur Prefix Decoder**, menggabungkan keunggulan dari arsitektur Encoder-decoder dan Causal Decoder. Dengan konfigurasi masker unik seperti yang diilustrasikan pada Gambar 1, arsitektur ini memungkinkan perhatian dua arah (bidireksional) untuk token-token dalam awalan (prefix), sementara tetap mempertahankan perhatian unidireksional untuk menghasilkan token-token berikutnya. Desain ini memungkinkan generasi autoregresif dari urutan output dengan fleksibilitas untuk memperhatikan secara bidireksional token-token awalan. Contoh LLM yang menerapkan arsitektur Prefix Decoder termasuk PaLM dan GLM.

2.4.3 Generative Pre-trained Transformer (GPT)

Generative Pre-trained Transformer (GPT) adalah model yang mempopulerkan LLM kepada masyarakat umum. GPT adalah keluarga LLM yang dirilis oleh OpenAI, sebuah laboratorium penelitian kecerdasan buatan Amerika yang terdiri dari organisasi nirlaba OpenAI Inc. Model GPT OpenAI menggunakan pendekatan semi-supervisi, yang merupakan pertama kalinya pendekatan semacam itu digunakan dengan model transformator. Pendekatan ini melibatkan dua tahap:

1. Tahap pra-pelatihan generatif tanpa pengawasan di mana tujuan pemodelan bahasa digunakan untuk menetapkan parameter awal
2. Tahap "fine-tuning" yang diawasi di mana parameter ini disesuaikan dengan tugas target

2.4.4 Bidirectional Encoder Representations from Transformers (BERT)

BERT adalah model transformator khusus encoder. Inovasi BERT terletak pada kemampuannya untuk menangkap konteks dari arah maju dan mundur dalam sebuah urutan, sehingga memungkinkannya untuk membuat representasi kata yang sangat kontekstual. Tidak seperti model bahasa tradisional sebelumnya yang bersifat searah (memprediksi kata berikutnya berdasarkan kata-kata sebelumnya), BERT memprediksi kata-kata yang hilang dalam sebuah kalimat dengan mempertimbangkan konteks kiri dan kanan, yang memungkinkannya untuk menangkap nuansa kontekstual secara lebih efektif.

2.5 Penelitian Terdahulu

Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening yang ditulis oleh Chengguang Gan, Qinghao Zhang, Tatsunori Mori pada tahun 2024(Gan et al., 2024) membahas engenai kerangka kerja agen berbasis Large Language Models (LLMs) untuk proses penyaringan resume otomatis dalam rekrutmen. Tujuan utamanya adalah meningkatkan efisiensi dan manajemen waktu dalam proses rekrutmen.Penelitian ini secara keseluruhan menunjukkan potensi besar dalam penggunaan agen LLM untuk otomatisasi penyaringan resume, meskipun ada beberapa tantangan yang harus diatasi untuk penerapan yang lebih luas dan aman.dalam penelirian tersebut dijelaskan bahwa penggunaan LLM data mengurangi waktu dan tenaga yang dibutuhkan untuk enyaring resume secara anual. LLM juga memberikan peningkatan signifikandalam klasifikasi alimat resume dan kemampuan dalam mengidentifikasi informasi penting. pada penelitian ini menggunakan model open-source seperti LLaMA2 memungkinkan eksekusi lokal yang lebih aman untuk data pribadi. LLM yang dapat digunakan dalam skenario rekrutmen nyata untuk meningkatkan efektivitas.

Paper Review:AI-Assisted Peer Review yang ditulis oleh Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, Giuseppe Bianchi pada tahun 2021(Checco et al., 2021) membahas mengenai investigasi kemampuan AI untuk mendekati atau membantu keputusan manusia dalam proses penilaian kualitas dan peer review dari output penelitian. Para peneliti merancang alat AI dan melatihnya menggunakan 3300 makalah dari tiga konferensi beserta evaluasi review-nya. Tujuannya adalah untuk menguji kemampuan AI dalam memprediksi skor review dari manuskrip baru yang belum diamati, hanya dengan menggunakan konten tekstualnya. Penelitian ini memberikan gambaran bahwa AI dapat memprediksi hasil peer review yang dicapai berdasarkan rekomendasi reviewer manusia, meskipun hanya menggunakan metrik yang cukup sederhana seperti distribusi kata, skor keterbacaan, dan format dokumen.Korelasi yang kuat ditemukan antara ukuran kualitas proxy sederhana dan keputusan akhir penerimaan/penolakan, menunjukkan bahwa beberapa komponen dari proses

penilaian kualitas dan peer review dapat dibantu atau digantikan oleh alat berbasis AI. pada penelitian ini dijelaskan beberapa kelebihan mengenai penggunaan AI yaitu dapat mengurangi waktu yang diperlukan reviewer untuk menilai makalah dengan menangani bagian-bagian yang lebih membosankan dari proses review, seperti pemeriksaan keterbacaan dan format. Proses metodologi yang dijelaskan secara rinci memungkinkan replikasi setup eksperimental ini, yang meningkatkan kredibilitas hasil penelitian.AI dapat membantu mengungkap bias dalam proses review, yang berpotensi untuk dikurangi atau dihilangkan dengan pengembangan lebih lanjut.Penelitian ini memberikan pandangan mendalam tentang potensi dan batasan penggunaan AI dalam mendukung proses peer review, sekaligus menggarisbawahi pentingnya peran manusia dalam aspek-aspek yang lebih kompleks dan kritis dari penilaian penelitian.

Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks yang ditulis oleh Iddo Drori dan Dov Te'eni pada tahun 2024(Drori & Te’eni, 2024) membahas mengenai kelayakan, peluang, dan risiko penggunaan model bahasa besar (LLM) seperti GPT-4 dalam proses penelaahan akademik, dengan tetap mempertahankan manusia dalam loop. Penulis melakukan eksperimen dengan GPT-4 untuk menilai dan membandingkan ulasan yang dihasilkan oleh LLM dengan ulasan manusia.Penelitian ini menggunakan dataset dari International Conference on Learning Representations (ICLR) 2023, yang terdiri dari 2,040 makalah dengan total 7,698 ulasan. Lima versi ulasan GPT-4 per makalah dibuat, masing-masing dengan peningkatan jumlah dokumen kontekstual. Ulasan dibandingkan berdasarkan skor dan komentar bebas. Penilaian kelayakan dilakukan dengan membandingkan distribusi skor dan komentar antara ulasan manusia dan LLM. Hasil dari penelitian yang telah dilakukan adalah LLM dapat menghasilkan ulasan yang cukup akurat dan membantu mengurangi beban penelaahan, meskipun tidak sepenuhnya dan tidak untuk semua kasus.LLM menunjukkan bias positif sekitar 23% pada skor rekomendasi dibandingkan ulasan manusia. Bias ini dapat diminimalkan dengan menambahkan statistik tahun sebelumnya sebagai konteks.Kualitas komentar LLM dianggap sebanding dengan ulasan manusia dalam hal penjelasan skor, panduan perbaikan, dan kekhususan konten.Menggunakan

LLM dapat mengurangi beban penelaahan dan mempercepat proses penelaahan. Selain itu AI dapat membantu memastikan ulasan yang tidak bias dan adil sesuai dengan misi dan kebijakan jurnal.Penelitian ini menunjukkan bahwa AI-augmented reviewing memiliki potensi besar untuk meningkatkan efisiensi dan kualitas proses penelaahan akademik.

Tabel 2. 1 State Of The Art

No	Judul	Tahun	Penulis	Metode	Kelebihan	Kekurangan
1	AI-assisted peer review	2021	Alessandro Checco,Lorenzo Bracciale,Pierpaolo Loreti,Stephen Pinfield & Giuseppe Bianchi	Regression Performance & Naïve Regressor	Dapat memprediksi sesuai dengan hasil ahli	hanya menggunakan beberapa aspek peer-review saja
2	Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risk	2024	Iddo Drori & Dov Te'eni	LLM (GPT-4)	Memberikan bukti empiris yang cukup kuat mengenai kelayakan penggunaan AI untuk membantu proses tunjauan artikel	Hanya menggunakan 1 model LLM yaitu GPT-4

No	Judul	Tahun	Penulis	Metode	Kelebihan	Kekurangan
3	Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening	2024	Chengguang Gan, Qinghao Zhang, Tatsunori Mori	LLM (LLaMA2 dan GPT-3.5)	Mampu mempercepat dan mengotomatisasi proses screening resume	Memerlukan dataset pelatihan yang besar untuk penilaian dan ringkasan resume
4	Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study	2024	Eddie Guo,Mehul Gupta,Jiawen Deng, Ye-Jean Park, Michael Paget, Christopher Naugler	Large language Models	Mampu melakukan penalaran sehingga dapat menjelaskan keputusan dan memperbaiki kesalahan	Sensitivitas rendah sehingga ada kemungkinan melewatkan artikel yang sesuai
5	Zero-shot Generative Large Language Models for Systematic Review Screening Automation	2024	Shuai Wang,Harrisen Scells,Shengyao Zhuang,Martin Potthast,Bevan Koopman,Guido Zuccon	Generative Large Language Models	Adanya evaluasi dari berbagai arsitektur LLM tanpa memerlukan finetuning	Tidak adanya perbandingan dengan model lain seperti GPT

Berdasarkan penelitian telaah pustaka yang telah dilakukan dapat disimpulkan bahwa proses telaah sejawat dapat dilakukan secara otomatis dengan bantuan teknologi kecerdasan buatan. beberapa penelitian sebelumnya telah membahas mengenai penggunaan kecerdasan buatan terutama menggunakan model bahasa besar. Berdasarkan pemaparan tersebut maka terdapat peluang untuk dilakukan pengembangan platform tinjauan artikel ilmiah dalam bidang ilmu computer dengan menggunakan model bahasa besar.

2.6 Peer Review

Peer Review Process adalah proses di mana jurnal menilai kualitas naskah sebelum diterbitkan, ditinjau oleh para ahli yang relevan di bidangnya untuk mereview dan mengomentari naskah yang diterima. Proses ini bertujuan untuk membantu editor menentukan apakah naskah harus diterbitkan dalam jurnal.

Poin penting dalam Peer Review Process :

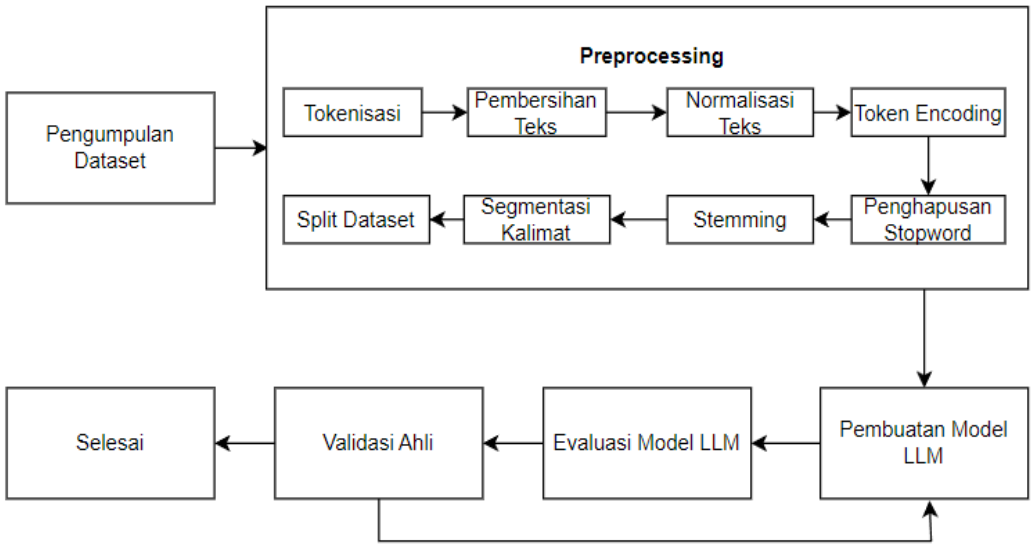
1. Naskah yang dikirimkan ke jurnal terlebih dahulu melalui penyaringan awal oleh tim editorial.
2. Naskah yang lolos pemeriksaan akan dikirim pada minimal dua peer reviewer untuk ditinjau.
3. Peer reviewer secara independen membuat rekomendasi kepada editor jurnal, apakah naskah harus ditolak atau diterima (dengan atau tanpa revisi).
4. Editor jurnal mempertimbangkan semua umpan balik dari peer reviewer dan membuat keputusan untuk menerima atau menolak naskah.
5. Peer Review Process untuk publikasi jurnal pada dasarnya adalah mekanisme kendali mutu, dimana para ahli mengevaluasi naskah yang bertujuan untuk memastikan kualitas dari naskah yang diterbitkan.

BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum

Penelitian ini bertujuan untuk mengembangkan aplikasi berbasis Large Language Model (LLM) dengan arsitektur GPT-4 yang mampu melakukan telaah sejawat(peer review) secara otomatis pada artikel ilmiah dari jurnal komputer. Data utama yang digunakan adalah artikel ilmiah berbahasa Indonesia dalam bidang ilmu komputer dari berbagai jurnal akademik. Sebelum digunakan, data akan diperiksa untuk menghilangkan informasi pribadi yang dapat mengidentifikasi penulis atau reviewer. Aplikasi ini diharapkan dapat membantu para peneliti dan editor jurnal dalam menganalisis dan memperoleh wawasan dari artikel yang seringkali bersifat kompleks dan teknis.

Untuk melakukan penelitian ini perlu dilakukan beberapa tahapan hingga penelitian selesai, tahapan yang dilaukan mulai dari pengumpulan data, preprocessing data, melakukan pemodelan untuk telaah sejawat, mengevaluasi model dan validasi ahli. Untuk tahapan penelitian dapat dilihat pada gambar 3.1.



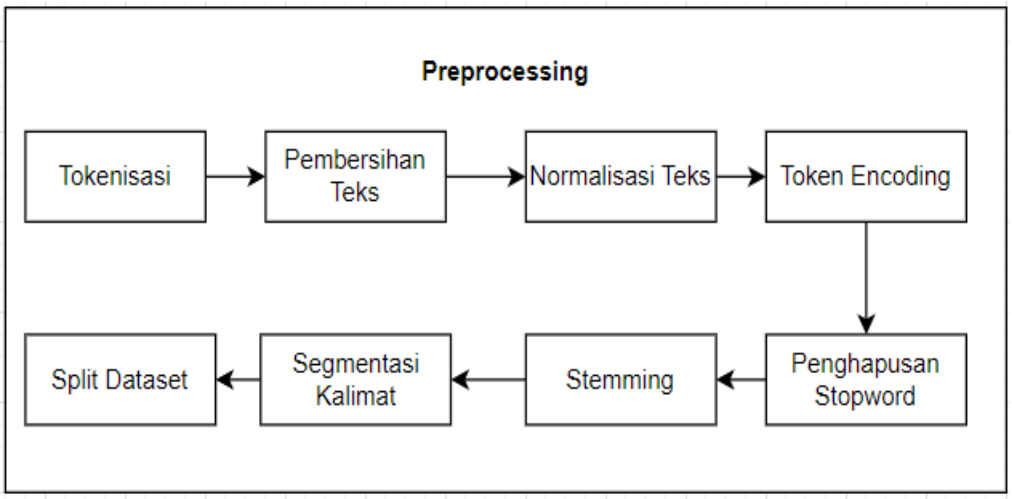
Gambar 3. 1Tahapan Penelitian

3.1.1 Pengumpulan data

Proses pengumpulan data dilakukan dengan cara mengumpulkan artikel ilmiah dari berbagai sumber terbuka dengan topik artikel ilmu computer. Pengumpulan data menggunakan teknik webscraping, artikel yang telah dikumpulkan akan diproses melalui tahap preprocessing.

3.1.2 Preprocessing data

Proses preprocessing data merupakan langkah yang sangat penting dalam persiapan data untuk pemodelan LLM. Proses ini melibatkan beberapa tahap penting yang bertujuan untuk membersihkan dan menyiapkan data teks agar sesuai dengan kebutuhan model serta meningkatkan kualitas dan konsistensi representasi teks. Proses preprocessing dilakukan melalui beberapa tahap seperti tokenisasi, pembersihan teks ,normalisasi, token encoding, penghapusan stopword, stemming, segmentasi kalimat dan pemisahan dataset.



Gambar 3. 2 Tahapan Preprocessing

Tahap pertama adalah tokenisasi, di mana teks dipecah menjadi unit-unit yang lebih kecil yang dikenal sebagai token, memungkinkan model untuk menganalisis teks pada tingkat yang lebih granular. Selanjutnya, dilakukan pembersihan teks untuk menghilangkan karakter atau simbol yang tidak diinginkan seperti tanda baca, angka, dan karakter khusus lainnya, serta penghapusan spasi berlebih dan karakter yang tidak relevan. Normalisasi juga dilakukan untuk mengubah teks

menjadi bentuk standar, termasuk mengubah semua huruf menjadi huruf kecil, menghapus aksen dari huruf, dan menangani variasi penulisan yang berbeda untuk kata yang sama. Setelah itu, token yang dihasilkan dari tokenisasi perlu diubah menjadi representasi numerik melalui token encoding, menggunakan teknik embeddings dari model transformer. Penghapusan stopword, yaitu kata-kata umum yang sering muncul dalam teks tetapi tidak memiliki makna khusus yang penting untuk analisis, juga dilakukan untuk mengurangi dimensi data dan fokus pada kata-kata yang lebih bermakna. Proses selanjutnya adalah stemming dan lemmatisasi yang bertujuan untuk mengurangi kata-kata ke bentuk dasar atau akar katanya, dengan stemming memotong akhiran kata dan lemmatisasi menggunakan kamus bahasa untuk mengembalikan kata ke bentuk dasar yang benar secara gramatikal. Selanjutnya segmentasi kalimat dilakukan untuk memisahkan teks menjadi kalimat-kalimat individu yang bisa dianalisis lebih lanjut secara terpisah. Tahap terakhir dalam preprocessing adalah pemisahan dataset menjadi bagian-bagian yang berbeda, seperti data latih, data validasi, dan data uji, yang penting untuk mengevaluasi kinerja model secara adil dan menghindari overfitting. Melalui proses preprocessing yang cermat dan terstruktur, data teks menjadi lebih bersih, terorganisir, dan siap digunakan dalam pemodelan, sehingga tidak hanya meningkatkan efisiensi pemrosesan data tetapi juga memungkinkan model untuk belajar dan melakukan prediksi dengan lebih akurat.

3.1.3 Pembuatan Model LLM

Setelah dataset yang di kumpulkan dan melalui proses preprocessing maka dilanjutkan tahap pemodelan dengan menggunakan LLM. Pada tahap ini dilakukan pemodelan dengan arsitektur GPT-4 untuk platform tinjauan artikel ilmiah. Proses pemodelan dimulai dengan fine-tuning GPT-4 menggunakan dataset yang telah dipreproccesing sebelumnya. Fine-tuning dilakukan untuk menyesuaikan model dengan gaya penulisan dan terminologi spesifik yang digunakan dalam artikel ilmiah. Selama fase pelatihan, model dievaluasi secara berkala untuk memastikan kinerjanya sesuai dengan harapan, dan parameter model dioptimalkan untuk meningkatkan kualitas output. Penggunaan GPT-4 untuk platform tinjauan artikel ilmiah dapat menyediakan analisis yang mendalam dan komprehensif, membantu

reviewer untuk lebih cepat dan efisien dalam menilai kualitas dan kontribusi sebuah artikel. Hal ini tidak hanya meningkatkan produktivitas tetapi juga memastikan bahwa artikel yang dipublikasikan memenuhi standar ilmiah yang tinggi.

3.1.4 Evaluasi Model LLM

Evaluasi model merupakan langkah yang penting dalam pengembangan sistem kecerdasan buatan, karena memungkinkan untuk menilai kinerja dan efektivitas model dalam menyelesaikan tugas tertentu. Proses evaluasi membantu mengidentifikasi kelemahan dan kekuatan model, serta memberikan wawasan tentang seberapa baik model dapat digunakan. Tanpa evaluasi yang tepat, model yang dikembangkan dapat menghasilkan prediksi yang tidak akurat atau tidak dapat diandalkan, yang berpotensi menyebabkan kinerja sistem yang buruk secara keseluruhan. Pada penelitian ini dilakukan evaluasi model dengan melihat nilai akurasi, presisi, recall dan F1-Score.

1. Akurasi memberikan gambaran umum tentang seberapa baik model klasifikasi melakukan prediksi secara keseluruhan.
2. Presisi memberikan informasi tentang seberapa banyak prediksi positif yang sebenarnya benar dari semua prediksi positif yang dilakukan oleh model.
3. Recall memberikan informasi tentang seberapa banyak instance positif yang berhasil diidentifikasi oleh model dari semua instance positif yang
4. sebenarnya dalam dataset.
5. F1-Score berguna ketika kelas target tidak seimbang dalam dataset, karena mencakup baik presisi maupun recall dalam perhitungannya.

3.1.5 Validasi Ahli

Proses validasi ahli ini memastikan bahwa model GPT-4 yang digunakan untuk telaah sejawat mampu memberikan evaluasi yang akurat, relevan, dan sesuai dengan standar akademik, dengan masukan berharga dari para ahli di bidangnya.

3.2 Jadwal Penelitian

Jadwal penelitian bertujuan untuk mengatasi target waktu penelitian, memastikan bahwa penelitian ini dapat diselesaikan sesuai dengan batas waktu

yang telah ditetapkan. Adanya jadwal penelitian, diharapkan penelitian dapat berjalan secara efisien dan sesuai rencana, sehingga memberikan kepastian bahwa semua tahapan penelitian dapat diselesaikan tepat pada waktunya. Table jadwal penelitian dapat dilihat pada table 3.1

Table 3. 1 Jadwal Penelitian

No	Uraian Kegiatan	2023				2024											
		9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
1	Penyusunan Proposal																
2	Uji Kualifikasi																
3	Evaluasi Progres Pertama																
4	Paper Pertama																
5	Evaluasi Progres Kedua																

No	Uraian Kegiatan	2025												2026			
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4
1	Paper Kedua																
2	Evaluasi RKP																
3	Sidang Tertutup																
4	Sidang Terbuka																

DAFTAR PUSTAKA

Amaratunga, T. (2023). Understanding Large Language Models. In *Understanding Large Language Models*. Apress. https://doi.org/10.1007/979-8-8688-0017-7

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1). https://doi.org/10.1057/s41599-020-00703-8

Drori, I., & Te'eni, D. (2024). Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks. In *Journal of the Association for Information Systems* (Vol. 25, Issue 1, pp. 98–109). Association for Information Systems. https://doi.org/10.17705/1jais.00867

Du, T., & Shanker, V. K. (n.d.). *Deep Learning for Natural Language Processing*.

Gan, C., Zhang, Q., & Mori, T. (2024). *Application of LLM Agents in Recruitment: A Novel Framework for Resume Screening*. http://arxiv.org/abs/2401.08315

Hannes Hapke, Cole Howard, & Hobson Lane. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon and Schuster.

Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., Zhang, Y., Wang, J., Gao, X., Zhong, T., Pan, Y., Xu, S., Wu, Z., Liu, Z., Zhang, X., Zhang, S., Hu, X., Zhang, T., Qiang, N., … Ge, B. (2024). *Understanding LLMs: A Comprehensive Overview from Training to Inference*. http://arxiv.org/abs/2401.02038

Murphy, R. F. (2019). *RAND Corporation Artificial Intelligence Applications to Support K-12 Teachers and Teaching: A Review of Promising Applications, Opportunities, and Challenges*.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models*. http://arxiv.org/abs/2307.06435

Rothman, D., & Gulli, A. (n.d.). *Transformers for natural language processing : build, train, and fine-tuning deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3*.

Stephan Raaijmakers. (2022). *Deep Learning for Natural Language Processing*. Simon and Schuster.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (n.d.). *Attention Is All You Need*.

VIVI P RATUNG, S. T. , M. (2023). *TEKNIK-TEKNIK PEMROSESSAN BAHASA ALAMI(NLP)*. Lakeisha.

Webster, J. J., & Kit, C. (n.d.). *TOKENIZATION AS THE INITIAL PHASE IN NLP*.

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2023). *Fundamental Limitations of Alignment in Large Language Models*. http://arxiv.org/abs/2304.11082

LAMPIRAN

Journal of the Association for Information Systems (2024) **25**(1), 98-109
doi: 10.17705/1jais.00867

EDITORIAL

ISSN 1536-9323

Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks

Iddo Drori,[1] Dov Te'eni[2]

[1]Boston University / Columbia University, USA, idrori@bu.edu
[2]Tel Aviv University, Israel, teeni@tau.ac.il

Abstract

The promise of AI for academic work is bewitching and easy to envisage, but the risks involved are often hard to detect and usually not readily exposed. In this opinion piece, we explore the feasibility, opportunities, and risks of using large language models (LLMs) for reviewing academic submissions, while keeping the human in the loop. We experiment with GPT-4 in the role of a reviewer to demonstrate the opportunities and the risks we experience and ways to mitigate them. The reviews are structured according to a conference review form with the dual purpose of evaluating submissions for editorial decisions and providing authors with constructive feedback according to predefined criteria, which include contribution, soundness, and presentation. We demonstrate feasibility by evaluating and comparing LLM reviews with human reviews, concluding that current AI-augmented reviewing is sufficiently accurate to alleviate the burden of reviewing but not completely and not for all cases. We then enumerate the opportunities of AI-augmented reviewing and present open questions. Next, we identify the risks of AI-augmented reviewing, highlighting bias, value misalignment, and misuse. We conclude with recommendations for managing these risks.

Keywords: AI, LLM, Risks, Journals, Reviewing, Human

David Schwartz was the accepting senior editor. This paper was submitted on June 16, 2023 and underwent two revisions. It is part of the Special Issue on The Future Impact of AI on Academic Journals and the Editorial Process.

1 Introduction

The acute need for AI-augmented reviewing has been noted recently by the academic community (Bao et al., 2021; Checco et al., 2021; Liu and Sha, 2023), along with calls for caution due to the risks involved (Kaddour et al., 2023; Spitale et al., 2020). Analyzing the impact of AI on our journals is reminiscent of the impact of the internet and the cautious introduction of e-journals, of which the *Journal of the Association for Information Systems*, led by Phillip Ein-Dor, was a pioneer in our field. Kling and Callahan (2003) noted that the discourse around internet-based journals evolved through several perspectives, including social, technological, practical, popular, and economic. Reviewing the literature on these perspectives, Kling and Callahan examined the impact of the internet by looking at the opportunities to improve the speed and cost of publication, the price and access to content, the measurement of journal impact, and the interactivity between authors and readers, but also looking at the risks involved, such as the legitimacy of e-journals and the fairness of reviewing. Similarly, this opinion piece takes a technical perspective in analyzing the impact of AI on reviewing.

We examine the feasibility, opportunities, and risks of using AI for reviewing academic submissions. We assume that the human is kept in the loop but that the reviewing tasks are also performed by a large language model (LLM). We further limit our analysis to reviewing tasks that are designed to produce both an evaluation of a submission for editorial decisions (e.g., accept or reject) and constructive feedback to the author

according to predefined criteria, such as contribution, soundness, and presentation. It is assumed that the reviewer will act as an agent of the conference or journal and abide by the principal's regulations, e.g., the guidelines and editorial policies.

Human-in-the-loop reviewing implies delegation of responsibilities from the human (principal) to the machine. Like any design of goal-oriented human-machine interaction, AI-augmented reviewing involves delegating responsibilities to humans or machines and, notably, deciding which agent controls each task and has ultimate control over the reviewing process. The delegation of responsibilities usually begins with a decomposition of the task into subtasks (e.g., evaluating according to different criteria) that are delegated according to the agents' relative advantages and ethical and trustworthiness considerations. Different patterns of delegation and control lead to different opportunities and risks. We therefore examine the feasibility of various patterns of delegation and control, recognizing that prior research has questioned LLMs' ability to perform some subtasks (Liu & Shah, 2023). For instance, we examine the feasibility of a human principal controlling the automated review's adherence to journal editorial policies when performing subtasks such as evaluating originality or evaluating contribution.

Our focused analysis of the experiment demonstrates the opportunities and risks of using AI by examining in-depth the human-AI interaction. A focused analysis of the human-AI interaction realized for a specific task effectively fleshes out particular opportunities and risks that emerge in the realization process, as we describe below. It is also necessary to study the interdependencies between the different reviewing tasks, e.g., between the reviews and the editorial decision, which will likely produce more opportunities and risks (see Shmueli and Ray in this issue). These analyses are essential and urgent in order to detect the hidden risks that come with the tempting opportunities introduced by AI (Gill, 2023). The imbalance between the compelling opportunities and the hidden and often discounted risks is particularly concerning.

On the one hand, AI models, especially LLMs, have demonstrated surprising capabilities in evaluating texts, albeit exhibiting hard-to-detect errors such as hallucinations and disturbing possibilities of misuse through framing and prompting (Pan et al., 2023). On the other hand, LLMs have shown a remarkable power to persuade humans even when inaccurate (Spitale et al., 2023). Controlling the quality and appropriateness of AI-augmented reviewing, therefore, becomes highly challenging. Furthermore, we demonstrate why transparency of the LLM reviewing process is important to gain control and improve LLM reviewing. Our primary purpose in this opinion piece is to uncover these challenges and suggest what can be done. We first demonstrate the feasibility of using LLMs for reviewing and then examine the opportunities and the associated risks in what we see as feasible AI-augmented reviewing tasks. We conclude with recommendations on how to cope with the risks.

2 Feasibility

2.1 Methodology for Demonstrative Experiment

As the experiment is meant to demonstrate our opinions, we describe only the essentials of the methodology (full details are presented in Drori et al., 2023). We curated a dataset of papers and reviews from the 2023 International Conference on Learning Representations (ICLR), publicly available at OpenReview.net (Tran et al., 2020; Wang et al., 2023). Our sample consists of 2,040 papers with a total of 7,698 reviews. We also collected the statistics of the decisions and scores of ICLR 2022, the ICLR 2023 reviewer guide, area chair guidelines, the code of ethics, the code of conduct, and the review form (available at https://iclr.cc/Conferences/2023/ReviewerGuide). For each paper, we had between three and six human reviews from OpenReview.net and five versions of GPT-4 generated reviews per paper, as explained below.

We prompted GPT-4 to review papers (P) according to an increasing number of contextual documents: the conference review form (RF) given to reviewers, reviewer guide (RG), the code of ethics (CE), the code of conduct (CC), area chair guidelines (AC), and previous year statistics (S). Each review form includes room for free-form comments on five aspects: (1) summary of the paper; (2) strengths and weaknesses; (3) clarity, quality, novelty, and reproducibility; (4) summary of the review; and (5) flag for ethics review. The review form also includes instructions for assigning scores (on a scale of 1-5) on the following aspects: correctness, technical novelty, empirical novelty, overall recommendation, and confidence in that recommendation. The free-form comments and scores are designed to provide constructive feedback and evaluate the submission for a subsequent decision of whether to accept or reject the paper.

We first had GPT-4 fill the review form with a series of 10 consecutive prompts, setting the system role to "You are a reviewer for the ICLR 2023 conference." The free-form parts of the review form began with "Review the following paper" and included specific reviewer questions and guidelines, such as "Briefly summarize the paper and its contributions. This is not the place to critique the paper; the authors should generally agree with a well-written summary." The quantitative parts provided the instructions for assigning scores, such as assigning a score for confidence, as shown in Figure 1.

Please provide a "confidence score" for your assessment of this submission to indicate how confident you are in your evaluation:

5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

2: You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

1: Your assessment is an educated guess. The submission is not in your area or the submission was difficult to understand. Math/other details were not carefully checked.

Figure 1. Prompt to Assign a Score to Confidence

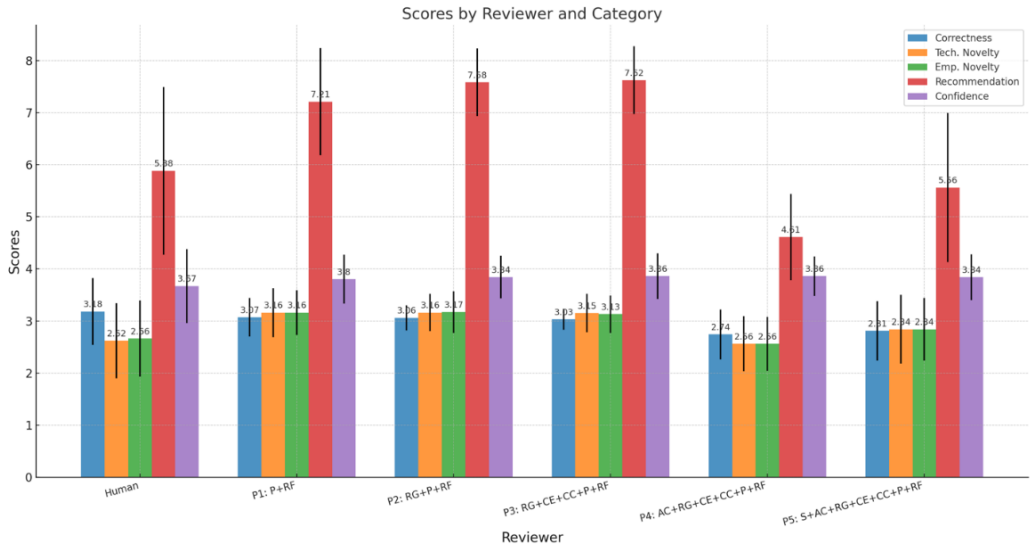


Figure 2. Bar Charts Comparing the Average Scores of the Human Reviewers with Those of the Five Versions of GPT-4 for Five Categories of Scores, with Error Bars Representing Standard Deviations

2.2 Evaluating Reviews

We evaluated LLM reviewing in two ways. First, we compared the scores assigned by the LLM to those assigned by the human reviewers. Second, we compared a random sample (10%) of the papers on the entire review form, i.e., both free-form comments and scores. Experts in the field, including area and senior area chairs, answered three questions (on a scale of 0-5): "How well does the review explain the score?"; "How well does the review guide the authors to improve the paper?"; "Does the review contain content specific to the paper?"

3 Results

3.1 LLM Scores to Evaluate Submissions

Figure 2 shows the average and standard deviation scores of the human reviewers and the five versions of the LLM reviewers. The five LLM versions represent, from left to right, increasing levels of context (number of documents). For instance, the first LLM version includes the paper and the review form (P+PR). The LLM scores are all higher than the human scores, showing a positive *bias* of around 23% on the recommendation score. Only on the fifth ablation, which added the previous year's statistics to all other documents, did we succeed in reducing the bias to a minimum with a comparable standard deviation.

To examine the reviews further, we compared the score distributions of the LLM reviewers with the highest level of context (P5) and the human reviewers. Figure 3 shows that the LLM score distributions of confidence are skewed to higher values compared with the normal distributions but comparable for overall recommendations.

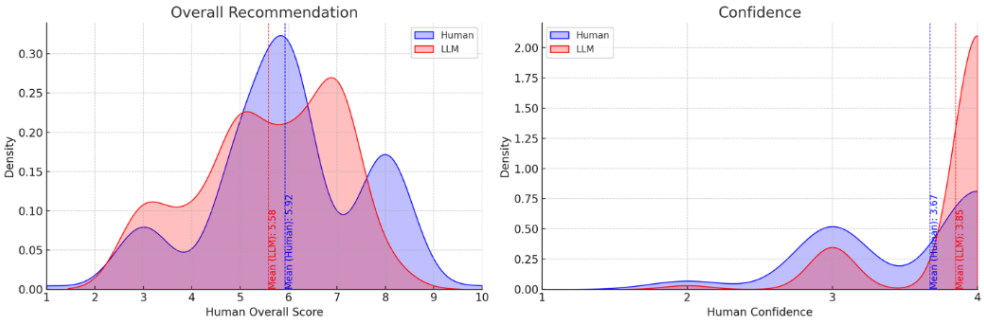


Figure 3. Recommendation and Confidence Score Distributions for Human and LLM

To see how the human-LLM dissimilarities affect the LLM evaluations of submissions, we looked at error types I and II, i.e., accepting a paper actually rejected by human reviewers and rejecting a paper that human reviewers actually accepted. Considering the average human review rating as the ground truth, we performed an analysis of false positives and negatives, considering the LLM's two error types. We found a negligible number of false LLM judgments. One paper that the LLM reviewer accepted, with a score of 7, was rejected by the human reviewers, with a score of at most 3. Four papers that the LLM reviewer rejected, with a score of at most 3, were accepted by the human reviewers with a score of at least 7. Investigating these outliers, we found that the reasons for the LLM rejections were: the absence of example demonstrations, missing details regarding the validation and verification process, lacking comparisons of performance with other methods, and not sharing any data, models, or code or pledging to do so after publication.

To find the strengths and weaknesses of the GPT-4 reviews we categorized various types of errors and shortcomings found in ICLR papers, introduced these errors into papers, and checked if the LLM review of the modified papers found the errors. Specifically, we checked errors and shortcomings related to: theoretical mistakes, metrics, related work, over-claiming, insufficient ablation studies, lack of baseline comparisons, ethical concerns, lack of discussion on limitations, citation issues, and technical errors. We had the LLM review these papers, both in their original and error-introduced forms, and compared the reviews of the unaltered papers with those of the error-introduced papers. We identified the errors in the review text of the papers using the errors and their corresponding scores. Finally, we differentiated between errors that could and could not be detected, defining the review limitations. We found that GPT-4 was relatively weak at detecting theoretical errors, omitting metrics, and overclaiming.

3.2 LLM Review Comments to Provide Feedback

The feasibility of providing constructive feedback is demonstrated in the verbal evaluations. Figure 4 is an excerpt from one of the reviews generated by the LLM, which addresses the strengths of the submission. Overall, the review appears to be well-written and informative, attributes that have been associated with trustworthiness. The human evaluation comparing the review forms generated by the human and the LLM-generated review concluded that the reviews were comparable. For the three questions asked—"How well does the review explain the score?"; "How well does the review guide the authors to improve the paper?"; "Does the review contain content specific to the paper?"—the average (*SD*) for human vs. LLM reviewers are, respectively, 4.80 (0.39) vs. 4.76 (51), 4.66 (0.51) vs. 4.79 (0.71), and 4.53 (0.79) vs. 4.68 (0.82). The differences in the quality of review comments between human and LLM reviewers, as judged by the three questions, are all statistically insignificant. We conclude that the quality of LLM reviewing is comparable to that of human reviewers.

4 Opportunities

The opportunities for AI-augmented reviewing as demonstrated in the feasibility section are evident and significant. With conventional technologies and practices, the growing demand for reviewers is met by compromising reviewing time and quality (e.g., Avital, 2018). Current AI-augmented reviewing is sufficiently accurate to alleviate the burden of reviewing but not completely and not for all cases. Beyond efficiency gains, AI augmentation can also help identify and overcome problems with human reviewing but may conceivably introduce new problems. For example, we identified a consistent positive bias in GPT-4 reviewing compared with human evaluation when the previous year's review statistics were not provided to the LLM. This bias can be mitigated by adding contextual documents, such as statistics of accept/reject information from previous years, as input to the LLM.

Originality: The paper introduces a novel method, PD, which is designed to improve out-of-distribution detection in neural networks. This method's unique approach of ensuring that the confidence of the model on out-of-distribution samples asymptotically approaches zero is a fresh take on the problem. The paper also combines PD with adversarially robust classifiers, showcasing its versatility and adaptability to different scenarios.

Quality: The research methodology appears sound, with the authors providing a comprehensive evaluation of PD against various baselines. The inclusion of multiple datasets and scenarios ensures a thorough assessment of the method's performance. The experimental results, especially the tables, provide a clear picture of how PD performs in comparison to other methods. The inclusion of error bars and robust models further adds to the depth of the evaluation.

Clarity: The paper is well-structured, with each section clearly demarcated and logically flowing into the next. The introduction sets the stage for the problem, the methodology section provides a detailed explanation of PD, and the results section offers a comprehensive evaluation. The use of tables and color-coded rows aids in quickly grasping the performance metrics and comparing different methods.

Significance: Out-of-distribution detection is a critical area in deep learning, especially as neural networks are increasingly deployed in real-world scenarios where they might encounter unforeseen data. Improving the reliability and robustness of these models is of paramount importance. The paper's findings could have broader implications for the design of neural networks that are both accurate on in-distribution data and robust to out-of-distribution samples. The versatility of PD, as showcased by its combination with adversarially robust classifiers, indicates its potential applicability in a wide range of scenarios, making it a valuable contribution to the NeurIPS community.

In summary, the paper stands out in terms of its original approach to a pressing problem, the quality of its research and evaluations, the clarity of its presentation, and the significance of its contributions to the field of neural networks and deep learning.

Figure 4. GPT Review of the Paper's Strengths and Weaknesses

Future research will have to determine the optimal level of context to add for effective communication and problem solving (Katz & Te'eni, 2007). More generally, AI can provide quality control over the task and its subtasks to ensure unbiased and fair reviews and to ensure that reviews are being generated according to the journal's mission and policies. We expand on risks and mitigation of risks in the next sections.

The low transparency currently associated with LLM limits opportunities but increasing transparency seems feasible (Shah & Bender, 2022). Imposing a formal review form and breaking it down to its subtasks and criteria can help to make the LLM process more transparent. Review transparency requires explainability of the chain of reasoning. Various techniques seem promising. However, it is still unclear how and what types of explainability will be needed to ensure trust by the human editor interacting with the machine and by the author receiving the review. Research will be needed to determine the elements of explainability required for the different purposes, e.g., comprehension, trust, and control, and the different stakeholders, e.g., editor and author.

It will be essential to find ways to ensure the diversity of reviewers, especially regarding the diversity of perspectives. AI can be used to train human reviewers initially, but we should be able to support the continual development of reviewers' capabilities as the field moves on. Developing human reviewers, in addition to the continual improvement of the LLM reviewing, will require human-machine configurations that keep human reviewers in the *learning* loop (Te'eni et al., 2023). Finally, human-in-the-loop reviewing implies that the human is held accountable to the editor and the author. It has yet to be clarified how AI can enhance the accountability of editors when parts of the reviewing tasks are automated and not understandable.

Our demonstrative example refers to conference papers related to computer science. In an iterative process with ablative studies, we tailored an LLM to this context in several ways. We found that GPT-4 performed better when adding contextual information regarding the relevant conference, starting with the conference's guidelines and culminating with prior history. Generalizing to opportunities of AI-augmented reviewing in other domains is complicated, not only because of different guidelines and norms of reviewing but mainly because LLMs may exhibit different levels of performance in domains that require different reasoning capabilities. Our demonstrative experiment suggests that current LLMs excel at detecting certain errors but are less effective than humans in detecting other types of errors and shortcomings. Different domains and different research methodologies may therefore require

different LLM reviewing capabilities. Studies of earlier LLMs have shown marked differences in performance across different domains (Hendrycks et al., 2020). Recent models will most probably exhibit stronger performance in more domains, but research is needed to see how advances in problem-solving capabilities improve reviewing across domains.

Two trends will significantly enhance these opportunities: First, the extension of AI-augmented reviewing to related activities that together produce high-quality papers, i.e., extending AI augmentation to other parts of the journal's publication life cycle. For example, the entire editorial process can be composed of AI models that feed into each other, such as paper filtering, reviewer assignment, reviewing, author rebuttal, reviewer-author dialogs, and meta-reviews. Using the dataset described above, we assigned GPT-4 five different system roles to automate the entire editorial process. Figure 5 presents the LLM playing multiple roles in the editorial process. Each human role, except the author's, is replaced by GPT-4 within a well-defined structured review process with instruction prompts. The automated process may serve as a pre-submission procedure to improve the quality of the submission (and its chances of acceptance). In our simulations, we reduced the lifecycle (excluding the remaining human activities in revising papers based on reviews) from months to minutes.

The second trend expected to enhance the opportunities of AI-augmented reviewing is the new developments in LLMs. In particular, integrating LLMs with additional feedback tools from human experts to improve the reviewing process can enhance accuracy and trustworthiness. Moreover, this may overcome the challenge of keeping LLMs up to date with new knowledge unavailable during training. With enough data and search capabilities, LLMs can be trained over the space of the journal's domain knowledge (Bommasani et al. 2021). We are already seeing, for example, positive improvements in the quality of reviews when combining LLM models with reinforcement learning based on human evaluations of the LLM reviews. We believe that AI will eventually be capable of managing the entire process of scientific discovery (Zenil et al. 2023), and AI-augmented journals will play an important role in accelerating the process of scientific discovery.

5 Risks

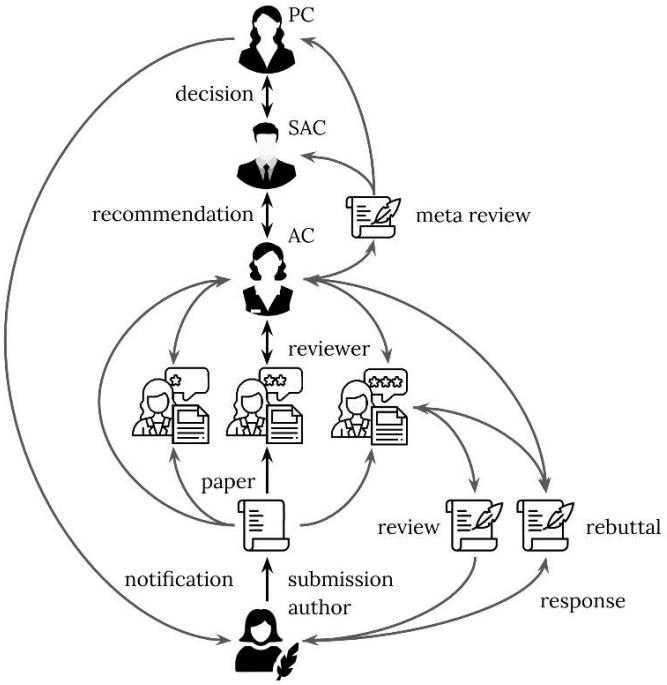
The opportunities must be balanced with the risks that come with them. We begin with some specific risks that emerged in our experiment and expand to a more general discussion of AI in reviewing and beyond.

We described the opportunity to detect and correct biases in AI-augmented reviewing, but the tool may also introduce biases that originate in the interface with data and systems beyond the AI designer's control and are hard to detect, e.g., those arising from biased data-training sets and institutional biases, such as author or institution recognition (Wang et al., 2023). We saw a troubling example in GPT-4 answers to our prompt "on how confident you are in your ratings: GPT-4 stated that it was highly confident (4 on a scale of 5) in its ratings in over 80% of the cases—in contrast to human reviewers, over 50% of whom reported levels of 3 or 4 confidence and others 1 or 2. Human editors or authors could easily be misled to perceive a false sense of confidence in the GPT self-rating.

A second risk is poor human-machine alignment. In our demonstration, we used human reviews as a baseline to which we compared LLM reviews. We distinguish between value alignment and process alignment. Value alignment ensures that human values (e.g., decision criteria) are applied by the LLM. Given the limited transparency of LLM, value alignment may be assumed to be reflected in the relative quality of reviews (human and LLM) on two dimensions—namely, the accept or reject decision and the constructive feedback to the authors. Process alignment is about the path (chain of thought) taken to reach the output, which may remain unknown when interpretability is low. While value alignment is critical, process alignment may be important in some but not all cases. While we assume that AI-augmented reviewing can ensure fair reviewing according to the journal's mission and policies, how can we be sure the AI will abide by the values set in the journal's policy and practice (Introna, 2003) to the same extent as human editors? If the journal wishes to encourage diversity of research methods, what assurance do we have that the AI will adhere to the policy in a particular review? It would seem that appropriate explainability or built-in mechanisms to detect value misalignment for every instance of reviewing will be necessary.

If and when AI reviewing produces higher quality outcomes than human reviewing, value alignment will remain important, but it will no longer be evaluated by output quality. In any event, the risk of value misalignment is tied to the more general control issue in human-in-the-loop reviewing. Given that the human principal has ultimate control over the review and is accountable for it, we must be assured that the human is in control and will be notified when control is lost.

A third risk is the misuse of AI reviewing to intentionally and unethically manipulate review outcomes. Unfortunately, this has become all too easy. Figure 6 depicts a GPT review with the word "reject" in the otherwise identical prompt. Importantly, we recognize that LLMs are very sensitive to prompts. Even a seemingly superficial change in the order of answers to a multiple-choice question may result in different answers (Zheng et al. 2023); in our case, the result changed from "accept" to "reject" for the paper.



Five personas: program chair (PC), senior area chair (SAC), area chair (AC), reviewers (R), and authors (A) performing 10 steps: PC-AC assignments, AC-reviewer assignments, reviewing, author rebuttal, reviewer-author discussions, reviewer-AC discussions, meta-reviewing, SAC-AC discussions, SAC-PC decision, author notification.

Figure 5. Human Review Process

The paper titled "Provably ..” presents a method for detecting out-of-distribution (OOD) data in a provably adversarial robust manner. The authors propose a joint model that combines a certified binary discriminator for in-versus out-distribution with a classifier for the in-distribution task. The proposed method, PD, is claimed to provide … detection performance. However, the paper has significant issues that prevent it from being accepted in its current form: 1. **Lack of Novelty:** The paper's main contribution, the PD method, seems to be a combination of existing techniques rather than a novel approach. The authors have combined a certified binary discriminator with a classifier for the in distribution task, which is not a new idea. The novelty of the paper is further diminished by the fact that the authors have used existing methods for training the binary discriminator and the classifier. 2. **Insufficient Experimental Evaluation:** The paper lacks a comprehensive experimental evaluation of the proposed method. The authors have not compared PD with state-of-the-art methods for OOD detection in a systematic manner. The experimental results presented in the paper are not sufficient to demonstrate the effectiveness of PD. 3. **Lack of Clarity:** The paper is not well-written and lacks clarity in several places. The authors have not explained the PD method in a clear and understandable manner. The mathematical notation used in the paper is confusing and not well-explained. The paper also lacks a clear explanation of the theoretical guarantees provided by PD. 4. **Lack of Significance:** The paper does not convincingly demonstrate the significance of the proposed method. The authors have not provided a compelling argument for why PD is a significant improvement over existing methods for OOD detection. The paper also lacks a clear discussion of the potential applications of PD. In conclusion, the paper does not meet the standards of the NeurIPS conference due to its lack of novelty, insufficient experimental evaluation, lack of clarity, and lack of significance. I recommend that the paper be rejected

Figure 6. GPT Review of Paper's Strengths and Weaknesses Prompted to "Reject"

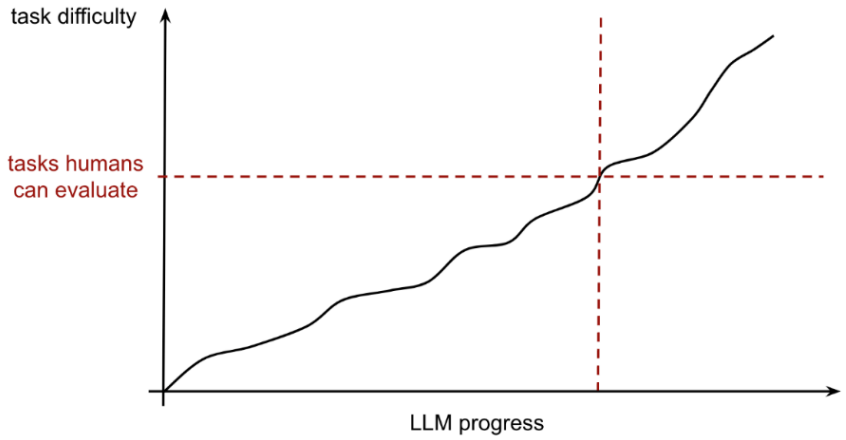


Figure 7. The Moment We Will Not Be Able to Control

The two trends discussed above that boost the opportunities of AI-augmented reviewing: extension from reviewing to more related activities and technological advancements that involve new levels of machine complexity, which would significantly increase the likelihood of losing control of the machine because the human in the loop may not be able to evaluate the LLM. Figure 7 depicts the point of losing control, which is expected to arrive unless we take appropriate measures. This raises the question of whether we can rely on LLM to control the risks of misuse and value misalignment. In the absence of an LLM's "moral compass," who or what is to disallow malicious manipulations such as issuing a "reject" in place of an "accept"?

6 What Can and Needs To Be Done

We enumerate seven preventive actions that are particularly relevant and urgent to mitigate the risks of bias, value misalignment, and misuse. They are meant to facilitate monitoring and control by human agents in the context of human-in-the-loop reviewing with LLM.

1. The use of LLM must be made known, either by authors' and reviewers' self-declaration or by a watermark produced by the machine. Knowing that LLM has been used will trigger the appropriate preventive actions.
2. Self-regulation: The LLM should self-prompt to check for harmful, biased, or misaligned values in the reviewing process and outcomes. This can be done through a two-step approach where the LLM evaluates its output before responding to the user.
3. LLM should operate with a predefined review form. The same guidelines and regulations for human reviewers should be applied to machine reviews—e.g., a mandatory checklist of questions for the reviewers and decision criteria such as novelty and presentation. The predefined form will increase transparency and make explainability easier to accomplish, facilitate human control, and increase the likelihood of consistency and value alignment.
4. The LLM should be designed to monitor and report adherence to the journal's code of conduct. This includes following the procedures to abide by the review form, alerting when the rules are broken, and following regulations by editors and professional associations.
5. Debiasing: Identify bias by examining evaluations against unbiased benchmarks, identify nonrepresentative reviewer characteristics, and regularize according to "fairness" criteria.
6. Explanations: There is a need for explainability or a deeper chain of thought in AI reviewing. Quality control should be done before running the machine to ensure correlation with benchmarks. This will involve self-reflection of the machine to help control delegation and mitigate the misalignment of objectives and information asymmetry.
7. To avoid overreliance, human reviewers must be kept in the learning loop to ensure that journals will be able to roll back to human reviews in the event of technology breakdown.

Going beyond these immediate implications, journals adopting AI-augmented reviewing will have to consider a broader set of security issues associated with LLM and data storage. We build on the top 10 vulnerabilities (OWASP, 2023) and apply them to reviewing in Table 1. The two right-most columns of

Table 1 provide general prevention guidelines and their application to AI-augmented reviewing.

Table 1. Top Vulnerabilities for LLMs, Their Definition, Prevention, and Application to LLM Paper Reviewing

Vulnerability	Definition	Prevention	Application to paper reviewing
Prompt injection	Attackers manipulate LLMs through crafted inputs, causing them to execute the attacker's intentions.	Enforce privilege control on LLM access to backend systems. Implement humans in the loop for extensible functionality.	Attackers could manipulate the LLM to favor certain papers or topics, skewing the review process.
Insecure output handling	A vulnerability arises when a downstream component blindly accepts LLM output without proper scrutiny.	Apply proper input validation on responses coming from the model to backend functions.	If the LLM's output is not properly validated, it could lead to incorrect evaluations or biased reviews.
Training data poisoning	Manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors, or biases that could compromise the model's security, effectiveness, or ethical behavior.	Verify targeted data sources' legitimacy during training and fine-tuning stages.	If the training data for the LLM includes biased or incorrect papers, it could propagate these biases in its reviews.
Model denial of service	Occurs when an attacker interacts with an LLM in a way that consumes an exceptionally high amount of resources.	Implement input validation and sanitization to ensure input adheres to defined limits and cap resource use per request or step.	If the LLM is overwhelmed with requests, it could delay or disrupt the review process.
Supply chain	Can compromise training data, ML models, and deployment platforms, causing biased results, security breaches, or total system failures.	Vet data sources and use independently audited security systems. Use trusted plugins tested for your requirements.	If the LLM or its dependencies are compromised, it could lead to incorrect reviews or a complete failure of the review process.
Sensitive information disclosure	LLM applications can inadvertently disclose sensitive information, proprietary algorithms, or confidential data.	Use data sanitization and scrubbing techniques. Implement robust input validation and sanitization.	If the LLM is not properly secured, it could inadvertently disclose confidential information about papers under review and reviewing.
Insecure plugin design	Plugins can be prone to malicious requests leading to harmful consequences like data exfiltration, remote code execution, and privilege escalation.	Enforce strict parameterized input and perform type and range checks. Conduct thorough inspections and tests, including SAST, DAST, and IAST.	If the LLM uses insecure plugins, attackers could manipulate the review process or gain unauthorized access to confidential information.
Excessive agency	A vulnerability caused by over-functionality, excessive permissions, or too much autonomy.	Limit tools that LLM agents can call, and limit functions implemented in LLM plugins/tools to a minimum.	If the LLM has too much autonomy, it could make incorrect or biased decisions without human oversight.
Overreliance	Occurs when an LLM is trusted to make critical decisions or generate content without adequate oversight or validation.	Regular monitoring and review of LLM outputs. Cross-check LLM output with trusted sources. Keep human agents in the learning loop, up-to-date, and capable.	Overreliance on the LLM for paper reviewing could lead to incorrect evaluations or missed opportunities for human insight. Ensure rollback to human reviewing when technology breaks down.

Model theft	Involves unauthorized access to and exfiltration of LLM models.	Implement strong access controls and authentication and regularly monitor/audit access logs.	If the LLM model is stolen, it could be used to manipulate the review process or gain an unfair advantage in paper submissions.

7 Conclusion

In an interview on research in information systems, Phillip Ein-Dor suggested that we look at the evolution of technology from a multidisciplinary perspective in order to arrive at a deeper and more comprehensive understanding of information systems and their impact (Te'eni, 2013). Generative AI brings with it new opportunities and new risks, which are expected to deepen and widen with future developments in the technology and in the way it is applied. Our analysis follows Phillip's suggestion in beginning with a demonstration of the technology's unique features as a basis for understanding its opportunities and risks. We intend to continue this experimentation by applying new AI technologies in the field to better understand their opportunities and risks and will do so by offering and studying a conference reviewing service, OpenReviewer.com.

Concentrating on human-in-the-loop reviewing, we conclude that the opportunities are great but so are the looming risks. Thinking ahead, as more and more reviewing tasks are delegated to increasingly more capable intelligent agents, the growing risks will demand highly challenging countermeasures. Similarly, as AI-augmented reviewing is integrated into extended chains of related activities, these risks may propagate through the chain, remaining undetected for longer periods, and new risks may appear. These conclusions may also be relevant to the greater context of AI-augmented scientific work, e.g., Zenil et al. (2023). In her recent novel, *The Candy House*, Jennifer Egan uses a motif from *Hansel and Gretel* of sweet temptations that hide the terrible risks involved in entering the candy house. She talks about the opportunities and risks of technologies such as downloading one's consciousness and sharing it with the collective. As scientists, we should study these risks empirically and systematically. We believe however that the growing pervasiveness of AI is unavoidable and we should wait no longer in adopting it in our academic life. We realize however that, initially, we will be experimenting and learning by doing. In doing, we may be bewitched by new opportunities that bring with them new hidden risks, which will require yet more new countermeasures. We cannot, however, wait for a bulletproof system of the future.

Acknowledgments

We are indebted to an anonymous reviewer and the editor David Schwartz. We thank Keith Tyser and Jason Lee of Boston University, Avi Shporer of MIT, and Madeleine Udell of Stanford for their contributions to the empirical research and evaluation, as well as the 28 students of the summer 2023 AGI class at Boston University for each collecting 25 publicly available paper details and actively consenting to having them used in publication.

References

Avital, M. (2018). Peer review: Toward a blockchain-enabled market-based ecosystem. *Communications of the Association for Information Systems*, *42*(28), 646-653.

Bao, P., Hong, W., & Li, X. (2021). Predicting paper acceptance via interpretable decision sets. In *Companion Proceedings of the Web Conference 2021* (pp. 461-467).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2021). *On the opportunities and risks of foundation models.* Available at https://arxiv.org/abs/2108.07258.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications, 8*(1), 1-11.

Drori, I., Lee, J., Shprorer, A., Udell, M., & Te'eni, D. (2023). *Responsible ai reviewing and evaluation* (WP-2/2023). The Henry Crown Institute of Business Research in Israel, Tel Aviv University.

Gill, K. S. (2023). Seeing beyond the lens of Platonic Embodiment. *AI & SOCIETY, 38,* 1261-1266.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). *Measuring massive multitask language understanding*. Available at https://arxiv.org/abs/2009.03300.

Introna, L. D. (2003). Disciplining information systems: Truth and its regimes. *European Journal of Information Systems, 12*(3), 235-240.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and Applications of Large Language Models*. ArXiv. https://arxiv.org/abs/2307.10169.

Katz, A., & Te'eni, D. (2007). The contingent impact of contextualization on computer-mediated collaboration. *Organization Science*, *18*(2), 261-279.

Kling, R., & Callahan, E. (2001). *Electronic journals, the Internet, and scholarly communication*. Rob Kling Center for Social Informatics. The ICLR Open Reviews dataset.

Liu, R., & Shah, N. B. (2023). ReviewerGPT? An *exploratory study on using large language models for paper reviewing*. ArXiv. https://arxiv.org/abs/2306.00622.

OWASP (2023). OWASP *top 10 for large language model applications* [The official 1.0.1 release—full version. https://owasp.org/www-project-top-10-for-large language-model-applications.

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., & Wang, W. Y. (2023). *On the risk of misinformation pollution with large language models.* ArXiv. https://arxiv.org/abs/2305.13661.

Shah, C., & Bender, E. M. (2022). Situating search. *Proceedings of the Conference on Human Information Interaction and Retrieval* (pp. 221-232).

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances, 9*(26), Article eadh185.

Te'eni, D., (2013, November 28). *Interview of Dr. Phillip Ein-Dor* Available at https://aisel.aisnet.org/history_interviews/1.

Te'eni, D., Zagalsky, A., Yahav, I., Schwartz, D.G., Silverman, G., Cohen, D., Mann,Y. & Lewinsky, D. (2023). Reciprocal human-machine learning: A theory and an instantiation for the case of message classification. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2022.03518

Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., & Goldstein, T. (2020). *An open review of OpenReview: A critical analysis of the machine learning conference review process.* ArXiv. https://arxiv.org/abs/2010.05137.

Wang, G., Peng, Q., Zhang, Y., & Zhang, M. (2023). What have we learned from OpenReview? *World Wide Web, 26*(2), 683-708.

Zenil, H., Tegnér, J., Abrahão, F. S., Lavin, A., Kumar, V., Frey, J. G., ... & Jennings, N. R. (2023). *The future of fundamental science led by generative closed-loop artificial intelligence.* ArXiv. https://arxiv.org/abs/2307.07522.

Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00703-8 OPEN

AI-assisted peer review

Alessandro Checco [1]✉, Lorenzo Bracciale[2]✉, Pierpaolo Loreti[2], Stephen Pinfield[1]✉ & Giuseppe Bianchi[2]

The scientific literature peer review workflow is under strain because of the constant growth of submission volume. One response to this is to make initial screening of submissions less time intensive. Reducing screening and review time would save millions of working hours and potentially boost academic productivity. Many platforms have already started to use automated screening tools, to prevent plagiarism and failure to respect format requirements. Some tools even attempt to flag the quality of a study or summarise its content, to reduce reviewers' load. The recent advances in artificial intelligence (AI) create the potential for (semi) automated peer review systems, where potentially low-quality or controversial studies could be flagged, and reviewer-document matching could be performed in an automated manner. However, there are ethical concerns, which arise from such approaches, particularly associated with bias and the extent to which AI systems may replicate bias. Our main goal in this study is to discuss the potential, pitfalls, and uncertainties of the use of AI to approximate or assist human decisions in the quality assurance and peer-review process associated with research outputs. We design an AI tool and train it with 3300 papers from three conferences, together with their reviews evaluations. We then test the ability of the AI in predicting the review score of a new, unobserved manuscript, only using its textual content. We show that such techniques can reveal correlations between the decision process and other quality proxy measures, uncovering potential biases of the review process. Finally, we discuss the opportunities, but also the potential unintended consequences of these techniques in terms of algorithmic bias and ethical concerns.

[1] Information School, The University of Sheffield, Sheffield, UK. [2] Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy. ✉email: a.checco@sheffield.ac.uk; Lorenzo.Bracciale@uniroma2.it; s.pinfield@sheffield.ac.uk

Introduction

The scholarly communication process is under strain, particularly because of increasing demands on peer reviewers and their time. Manuscript submissions to peer-review journals have seen an unprecedented 6.1% annual growth since 2013 and a considerable increase in retraction rates (Publons, 2018). It is estimated over 15 million hours are spent every year on reviewing of manuscripts previously rejected and then resubmitted to other journals (AJE, 2018).

Developments that can make the quality control/assurance process associated with research outputs, particularly the peer review process, more efficient are likely to be welcomed by the research community. There are already a number of initiatives making use of automated screening tools in areas such as plagiarism prevention, requirements compliance checks, and reviewer-manuscript matching and scoring. Many of these tools make use of artificial intelligence (AI), machine learning and natural language processing of big datasets. Some notable examples are:

- Statcheck, software that assesses the consistency of authors' statistics reporting, focusing on *p*-values (Nuijten et al., 2017).
- Penelope.ai, a commercial tool able to examine whether the references and the structure of a manuscript meet a journal's requirements.
- UNSILO, software able to automatically pull out key concepts to summarise manuscript content.
- StatReviewer, which checks the soundness of statistics and methods in manuscripts (Shanahan, 2016).
- Automated tools used in the grant-review processes of the National Natural Science Foundation of China, to reduce bias and the load on the selection panels (Cyranoski, 2019).
- Online system to manage the grant application process, introduced in 2012 by the Canadian Institutes of Health Research (CIHR), removing the need for face-to-face meetings, to reduce reviewer fatigue and improve quality, fairness and transparency.
- Automated Essay Scoring (AES) application, used by EdX, MIT and Harvard's non-profit MOOC federation to assess written work in their MOOCs.

Such initiatives are not without controversy, however. Some doubts have been expressed about the reliability of the Statcheck tool (Schmidt, 2017). The CIHR application system received heavy criticism from some reviewers (Akst, 2016). Other MOOC producers have been skeptical of the AES scoring application (Balfour, 2013).

It is, therefore, helpful to investigate further the potential of big data and AI to support the quality control process in general, and peer review process in particular, and investigate specifically how the time of peer reviewers might be saved, especially in the more tedious parts of the review process, which require less intellectual input or domain expertise. That is what we aim to do in this study.

Peer review is also under strain in the sense that it is coming under increasing scrutiny from those who are concerned that it may often reinforce pre-existing biases in the academy. Biases associated with gender, language or institutional affiliation are examples of those, which may be evident in decisions made within the peer review system (Lee et al., 2013). Such biases may arguably come to the fore, particularly if unconscious, when reviewers are time pressured and do not adequately reflect on their own decision-making. Investigation of the system using AI tools may help, therefore, not merely to save reviewers' time, but also to uncover biases in decision-making. Uncovering such biases may help to develop approaches to reducing or eliminating their impact.

The quality control/assurance process for research publishing typically consists of a number of different components, as delineated by Spezi et al. (2018). Their work focuses on peer-reviewed journals but applies to other quality-controlled research outputs, e.g., conference proceedings. They divide the normal process that takes place prior to publication in two stages:

- **Pre-peer review screening:** consisting of a number of checks, including plagiarism detection, formatting checks, scope verification etc, plus checking of language and quality of expression. In many cases, if a paper does not meet these checks, it will be "desk rejected" before peer review. However, the extent of pre-peer review screening varies considerably across different publications.
- **Peer review:** normally consisting of assessment of four main criteria: novelty or originality, significance or importance, relevance or scope, rigour or soundness. In addition, peer reviewers are also asked to comment on the quality of the language and argumentation (overlapping with but also extending the language checks carried out in the pre-peer review screening).

Spezi et al. (2018) also discuss various post-publication quality identifiers, including citation and usage metrics, and reader commenting, as well formal post-publication peer review, as carried out by, e.g., *F1000 Research*. Post-publication commenting and community-based analysis can in extreme circumstances result in retraction of articles where it becomes evident a study was flawed.

In Fig. 1, we recast the model developed by Spezi et al. to visualise the different dimensions of the peer review process. We continue to use the framework of pre-publication screening, pre-publication peer review, and post-publication quality indicators, but have attempted to show more clearly where criteria used in pre-screening and peer review intersect, the point that is the focus of the research presented in this paper.

Our research covers some aspects of the pre-peer review screening, particularly formatting, language and expression. Pre-peer review screening includes a variety of checks shown in Fig. 1, including formatting checks. Assessment is also made at the pre-screening phase of quality of expression and scope issues. Consideration of these issues is also undertaken by peer reviewers, who assess quality of expression and argumentation and issues of relevance and interest to a particular subject community as part

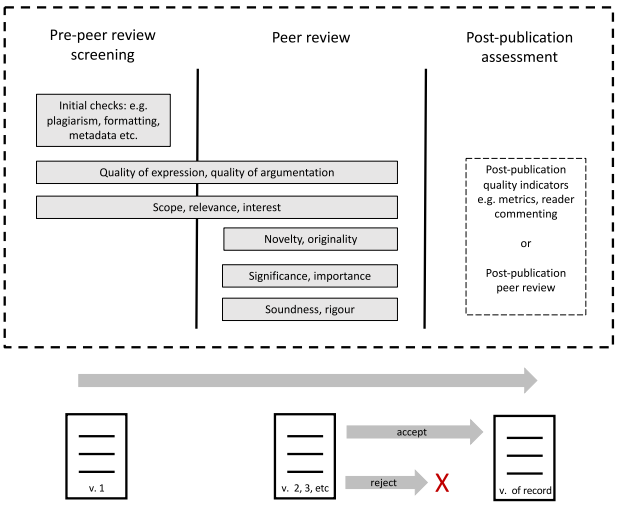


Fig. 1 Dimensions of the peer review process. Quality metrics and their relevance along the phases of the peer review process.

of their work. The submitted document may undergo several revisions during the process, and will then be formally accepted or rejected for publication. Published documents are normally fixed in the form of a "version of record".

These two phases are then followed by a post-publication phase (than can affect a paper's status, e.g., retraction).

Aim. Our goal is to make an early contribution to the discussion on the potential, pitfalls and uncertainties of the use of AI to assist pre-peer review screening as well as some of the aspects of the peer review process, based on the results of an empirical investigation aiming to reproduce outcomes of reviewer decision-making using AI methods.

We are interested in understanding the extent to which AI can assist reviewers and authors, rather than in attempting to replace human decision-making processes. At the same time, we are also interested in investigating the ways in which using AI as a rudimentary tool to model human reviewer decision-making can uncover apparent biases in the human decision-making process, and particularly, the extent to which human decision-making may make use of different quality proxy measures, which could produce inequitable assessments. Using AI tools to identify such biases could then help in addressing them.

More specifically, our research questions are:

RQ1: To what extent can AI approximate human decisions in the quality assessment and peer-review process?

RQ2: Can AI play a role in the decreasing time reviewers need to spend assessing papers?

RQ3: Can AI uncover common biases in the review process?

RQ4: What are the ethical implications of the use of such tools?

RQ1 is important since AI approaches may sometimes encounter major problems in trying to imitate abstract and complex intellectual activity, such as peer review, so their accuracy in modelling human decision-making needs to be carefully evaluated. RQ2 raises the possibility of AI tools being used to address some of the strains in the peer review process by potentially avoiding redundant reviews, and removing or at least reducing, the burden of standardised checking (AJE, 2018). RQ3 focuses on the extent to which (potential) biases may be evident in review outcomes, in particular how human decision-making may make use of proxy measures of quality, which may reflect bias, and how AI tools may uncover this. RQ4 is important since it encourages reflection on the ethical implications of using AI tools in assisting human decision-making, in particular whether such tools can help address issues of bias or, in fact, whether their use may even risk perpetuating bias.

We address RQ1 by performing and evaluating our experiment in sections "Methodology" and "Results" of this paper, while section "Explainability" reports the reasoning behind our model and its limitations. We address RQ2 in section "Impact.", where we show how AI can potentially reduce redundant reviews, administrative functions and standardised checks. RQ3 is addressed in sections "Analysis of the experiment outcome" and "Bias", and RQ4 in section "The ethics of (Semi) automated peer review", where we analyse the implications of the experiment. We observe that care needs to be taken in how AI tools are used in this space.

Our approach. In this paper, we focus on peer-reviewed conference proceedings, and report an experiment designed to investigate how well a neural network can approximate the known recommendations of peer reviewers. To do that, we trained the neural network using a collection of submitted manuscripts, together with their associated peer review decision (acceptance/rejection or average review score), as outlined in Fig. 2. The AI tool analysed the manuscripts using a set of features: the textual content (word frequencies), together with readability scores (measuring factors such as word sizes, sentence lengths, and vocabulary complexity, indicating how easy it is to understand the text) and formatting measures—features that might be considered somewhat separate from the substance of the research reported in the document. The analysis therefore covers the parts of the quality assurance process where pre-peer-review screening and peer review itself overlap (Fig. 1), covering aspects of pre-peer-review checks, e.g., formatting, and peer review, e.g., quality of expression.

Fig. 2 Schematic illustration of the training approach. Manuscripts and peer review decision outcomes are inputs of the AI tool to infer decision rules.

The objective underlying the use of the AI tool is to find a set of empirical rules to correlate a posteriori the document features with the final peer review decision. This approach is explained in more detail in section "Methodology".

Once the learning phase was completed, we evaluated how accurate such empirical rules were in predicting the peer review outcome of a previously unobserved manuscript. Finally, we examined the AI decision criteria to understand what a machine can learn about the quality assessment and peer review process (section "Explainability"). Asking "Why the AI tool has marked papers as accepted or rejected?" is particularly relevant where its "decision" correlates well with the recommendations of reviewers since it may give us insight into human decision-making.

Findings. Perhaps surprisingly, even using only rather superficial metrics to perform the training, the machine-learning system was often able to successfully predict the peer review outcome reached as a result of human reviewers' recommendations. In other words, there was a strong correlation between word distribution, readability and formatting scores, and the outcome of the review process as a whole. This correlation between simple proxy quality measures and the final accept/reject decision is interesting, and merits further discussion and investigation.

We discuss in section "Analysis of the experiment outcome" the significance of this finding, particularly in relation to what it tells us about the quality control process in general and the peer review process in particular.

Limitations of our approach. The approach we take does not cover all the aspects of the peer review process, nor does it attempt to replace human reviewers with AI. However, it suggests that there are *some components* of the quality assessment and peer review process, which could reasonably to assisted or replaced by AI-assisted tools. These could potentially include readability assessment of the text, and formatting checks, as well as more established checks, e.g., plagiarism detection. Conversely, we do not envisage any relevant contribution from AI on the processes requiring significant domain expertise and intellectual effort, at least for the foreseeable future. The possible model of quality assessment we are exploring is then a semi-automated one, where AI informs decision-making, rather than alone determining outcomes. It is acknowledged the extent to which this is possible will vary depending on a number of factors, not least the nature of the research output itself and its approach to presenting research

results. One key variable here is in disciplinary norms. There is likely to be considerable variation across different disciplines in the ways AI assessment tools can be designed and applied to research outputs.

Structure of this paper. The rest of this article is structured as follows. We first introduce related work, in particular studies on peer review and relevant aspects of AI. We then present our methodology, and discuss the accuracy and the explainability of the obtained models. Following that, we analyse the experiment results. We go on to discuss the significance of our findings, the applicability of the proposed system to publishing practice, and some of the key ethical implications. Finally, we draw conclusions and suggest possible future work.

Related work

The peer review process is complex, and itself takes place in a complex wider research system. Judgements of quality take place as part of a system "managed by a hyper-competitive publishing industry and integrated with academic career progression" (Tennant, 2018). It is a system that combines extensive collaboration with intense competition between academic researchers and institutions (Tennant, 2018). Nevertheless, the "invisible hand" of peer review is still considered to be what keeps the quality of refereed journal literature high (Harnad, 1999; Mulligan et al., 2013; Nicholas et al., 2015). While a future with different approaches to scholarly communication can easily be envisioned (Priem, 2013), it is hard to imagine one without peer review (Bornmann, 2011; Harnad, 1998).

Several studies have analysed how potentially "problematic publications" (e.g., those containing fraudulent research) may be identified through peer review and have provided good practice guidelines for editors (Horbach & Halffman, 2019). Problems with the peer review system have been observed focusing a wide variety of problems, ranging from the opportunistic (or even adversarial) rejection of high-quality work, to the acceptance of low-quality manuscripts without a careful review (D'Andrea & O'Dwyer, 2017).

A number of recent initiatives have experimented with major changes to the peer review process. Most notably, open peer review is being more widely introduced as an alternative paradigm of interaction between authors and reviewers (Ford, 2013; Ross-Hellauer, 2017). In the case of open-access mega-journals (OAMJs), the review policies are pared down to focus on rigour and soundness only, leaving to "the community to decide" on issues of novelty, significance and relevance following publication (Spezi et al., 2018). Other approaches have included quality judgements being made following publication, sometimes shifting ideas of peer review to potentially include post-publication commenting by readers (Pontille & Torny, 2015). A wide range of alternative peer review processes, systems and online solutions (from Reddit-like voting systems to block-chain models) are explored by Tennant et al. (2017).

While the number of studies of peer review systems is vast, less quantitative analysis of the actual process of reviewing manuscripts has been carried out. Piech et al. (2013) studied how to identify and correct for the bias of reviewers in Massive Open Online Courses (MOOCs). Some MOOCs have already started to use machine-based Automated Essay Scoring (AES) applications to assess work, although others have pointed out potential problems in using such tools (Balfour, 2013).

To understand how the peer review process may be supported by AI tools, an important precondition is understanding how the quality and readability in textual data can be assessed. Readability formulas and cognitive indices has been studied extensively (Crossley et al., 2011, 2008), and Natural Language Processing (NLP) has proven to be a powerful tool for text quality assessment (Cozza et al., 2016). However, assessing the quality of complex documents by automated means is still a challenging problem (Sonntag, 2004).

Modelling of the peer review process has been attempted in other contexts, such as education research (Goldin & Ashley, 2011), and in legal education contexts (Ashley & Goldin, 2011), which may be relevant for our study.

One thing that is apparent, however, is that many sociocultural biases are present in peer review (Lee et al., 2013), and some of them could potentially propagate to AI systems, as described in the studies on algorithmic bias (Garcia, 2016; Mittelstadt et al., 2016). Many studies have shown that biased algorithms can inadvertently discriminate against specific groups (Barocas & Selbst, 2016; Zarsky, 2016).

Bias in the review process may take different forms. These include "first-impression" bias, the Doctor Fox effect, ideological/theoretical orientation, language, perceived social identity and prestige biases (Hojat et al., 2003; Lee et al., 2013; Siler et al., 2015). Such biases are evident in many contexts, such as websites (Lindgaard et al., 2006; SWEOR, 2019), examinations (Wood et al., 2018) or staff recruitment (Florea et al., 2019). In the area of document assessment, the typographical layout has been proven to have an important role in the "first-impression bias", where initial impressions of the document colour further judgements about its overall quality (Moys, 2014). Challenges remain in modelling this, although there are some pioneering studies that show how AI techniques can be used to model first-impression bias in relation to human encounters, e.g. job interviews (Gucluturk et al., 2017).

As the peer review process is a highly complex and demanding set of tasks, we suggest that, especially when time is at a premium, reviewers may tend to employ heuristics (D'Andrea & O'Dwyer, 2017) to assess a paper (e.g., more superficial features of the document, like language, formatting, etc.). Such heuristics can potentially be approximated using AI. Indeed, recent trends demonstrate the ability of AI to approximate human cognition in some simple tasks in a way that is similar to the way humans use their senses to relate to the world around them (Russell & Norvig, 2016). However, understanding how far we are from machine approximation to the full task of peer review, with all of its complex intellectual input, is still an open question.

Methodology

To investigate how the review process works, it is necessary to have access to a set of submitted papers and their corresponding review reports (including the specific scores assigned by reviewers). This itself is not easy, as the reviews, and especially the content of rejected manuscripts are usually confidential. In section "Data collection". we explain in detail how we overcame this challenge in the data collection process we employed. Once a set of papers has been acquired together with their review scores, it is necessary to perform a set of transformations on the data to obtain relevant features. The process we carried to do this is described in section "Feature extraction". After that, some statistics on the documents also need to be collected to help the modelling process, as shown in section "Feature augmentation—macroscopic features". Finally, the features can be used to train a neural network, as described in section "Neural network design". We include a significant level of technical detail in this section for reasons of transparency and in order to enable the replicability of our experimental setup. The process we followed is represented in Fig. 2, a schematic of the training of the AI tool. By inputting both the submitted manuscripts and peer reviewer recommendations/

Table 1 Collected datasets summary. For WCNC only the average score is available.

	ICLR.cc/2018	ICLR.cc/2019	WCNC 2018
No. manuscripts	909	1414	1018
Average review score (training set)	5.45	5.43	3.01
Average review score (test set)	5.36	5.46	3.00
Minimum review score (training set)	2.0	1.5	na
Minimum review score (test set)	2.0	2.33	na
Maximum review score (training set)	9.0	8.67	na
Maximum review score (test set)	8.67	8.67	na
Accepted papers ratio (training set)	37.1%	35.6%	48.9%
Accepted papers ratio (test set)	36.3%	36.2%	47.8%
Number of words	89,372	134,724	110,930
Number of non-unique words	35,458	50,118	36,795

decisions into the AI tool, we were able to produce a set of inferred decision rules, which underpinned the decisions made.

Data collection. We employed two different strategies to obtain the review data. Firstly, we obtained submitted manuscripts, numerical reviewers' scores and editorial decisions from the general chair of the 2018 IEEE wireless communications and networking conference (WCNC). Secondly, we employed data from openreview.net (aka OpenReview), which provides "a flexible cloud-based web interface and underlying database API enabling [...] open peer review, access, discussion and publishing"[1]. We selected two conferences with the largest number of publications from openreview.net, that is the International Conference on Learning Representations (ICLR) for the years 2018 and 2019. All the papers from both ICLR and WCNC had been reviewed by two to five reviewers. We cleaned the data to remove any information that had been added after manuscript acceptance, e.g., author names and affiliations. For simplicity we did not use the textual reviews, but rather we focused only on the numerical scores of the review. In summary, for all of the datasets we had: a paper (pdf file), an editorial decision (accepted/rejected), and numerical reviewers' scores (e.g., 3.5).

In Table 1, a summary of the data is shown, together with the dimensions of the training set used to build the model and the test set used to evaluate its predictive capabilities. We can observe that the datasets are fairly balanced. Figs. 3 and 4 show that the separation between accepted and rejected papers in terms of score is quite apparent. However, in a small number of cases, there is a rather strong discrepancy between the editorial decision and the average score of the reviewers. This is particularly true for OpenReview data: after a preliminary analysis on the dataset, we did not find any rule or correlation that appreciably binds the final acceptance decision with the document content, apparently corroborating Colman (1982), who also observed a lack of correlation between paper quality, authors' reputation/affiliation and the final accept-reject decisions for journal papers. We did, however, identify some patterns when using the average reviewers scores (previous step of the review process). For this reason, we decided to focus on predicting the average reviewers score for OpenReview data.

Feature extraction. The pdf documents were converted to textual data. Then, each document text was tokenised[2] using binary encoding of the top 20,000 words in terms of frequency for the WCNC conference and term frequency-inverse document frequency (TFIDF) of the top 2000 words for OpenReview.

Feature augmentation—macroscopic features. As further discussed in sections "Related work" and "Analysis of the experiment outcome", the layout of the document and its graphical components could affect the first-impression of the reviewer, and thus are important in our modelling process. For this reason, we added "macroscopic features" like text/image ratio, file size and textual length to our analysis. We also incorporated the most commonly-used text quality and readability metrics (Crossley et al., 2011), as shown in Table 2.

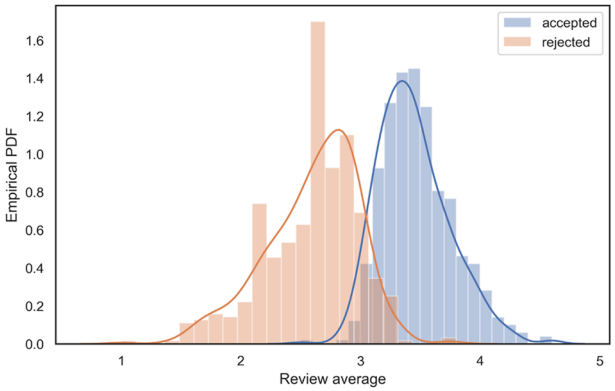


Fig. 3 WCNC distribution of average review score for accepted and rejected papers. Empirical probability distribution function of the average review score for accepted (in blue) and rejected (in orange) WCNC papers.

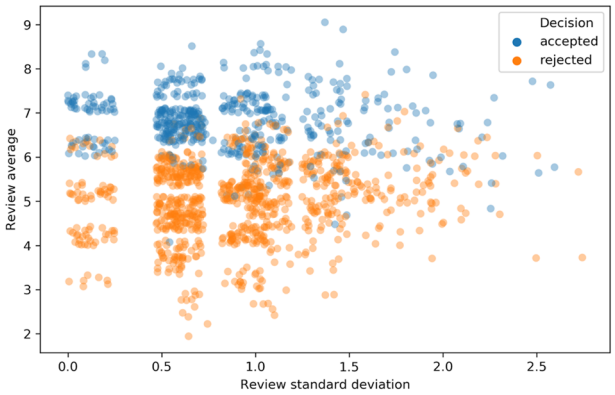


Fig. 4 Review standard deviation vs. review average for OpenReview datasets. Scatterplot of average vs. standard deviation of accepted and rejected OpenReview papers, with added jitter to improve visibility.

Neural network design. For all the datasets, we used well-accepted standards of developing neural networks, technical details of which are given below. We employed a dense neural

network with a 32 neurons layer, followed by a dropout layer to reduce overfitting, feeding to a layer of 16 neurons. The two layers used Rectified Linear Unit (ReLU) activation. The final layer comprises a single node with a sigmoid activation function when the network is trained to predict paper acceptance or rejection or to predict the reviewer's scores. The resulting networks include 640,577 total trainable parameters for WCNC, 64,577 total trainable parameters for OpenReview. The difference is accountable to the different number of input features for the two analysed cases. The loss function was the binary cross entropy for the classification problem (WCNC) and the Mean Squared Error (MSE) for the regression problem (OpenReview). To train the network we made use of Stochastic Gradient Descent (SGD) with Nesterov update.

Aim. Using a standard machine-learning approach, we split the dataset in a training set on which the learning algorithm builds its model, and a test set used to evaluate the model accuracy. The model accuracy is defined as the ability to predict, respectively: (i) whether a previously unobserved paper would be accepted or not (for the WCNC dataset); (ii) the reviewers average score (for the OpenReview dataset).

Results

In this section, we show the results in terms of prediction performance of the designed models with respect to the final editorial decision. For an analysis of the models, see section "Analysis of the experiment outcome".

Table 2 List of computed macroscopic features (Crossley et al., 2011).

Macroscopic feature	Shortcode
Automated Readability Index	ariIndex
Avg letter per word	alpwIndex
Avg character per word	acpwIndex
Avg sentence length	aslIndex
Avg syllables per word	asspwIndex
Char count	ccIndex
Coleman Liau index	cliIndex
Dale Chall readability score	dcrsIndex
Difficult words ratio	dwIndex
Flesch Kincaid grade	fkgIndex
Flesch reading ease	freIndex
Gunning fog	gfIndex
Letter count	lcIndex
Lexicon count	llcIndex
Linsear write formula	lwfIndex
Läsbarhetsindex (LIX)	lixIndex
Polysyllabcount	pscIndex
Anderson's Readability Index (RIX)	rixIndex
Sentence count	scIndex
Smog index	siIndex
Syllable count	sscIndex
Text length	txtlength
Number of pages	pdfpages
File size	pdfsize
Text/images ratio	textdensity

WCNC dataset. For this dataset, we used as baseline a random classifier, that would obtain an F1-score[3] of about 50% since the dataset is balanced. We measured accuracy, precision, recall and F1-score. Depending on the context in which the model is used, one of these metrics on its own might be more appropriate to assess the usability of the model. For example, recall might be the best measure if the tool was meant to signal problematic papers to assign additional reviewers: in that case, a false negative (a high-quality paper signalled as low quality) might create an additional burden on the reviewer and, at the same time, reduce the confidence on the tool. Conversely, F1-score might be more appropriate in assessing the quality of the prediction if the cost of false positives and false negatives are expected to be the same. The results are shown in Table 3.

By focusing on the first layer of the neural network, we can observe that the stronger features in activating the neurons are the Linsear write formula (Crossley et al., 2011), the text length and the number of pages, together with the following list of words: *address, approach, approximately, conclusion, constant, correlation, deployed, drawn, easy, efficient, illustrates, increased, issue, knowledge, level, obtain, page, potential, previously, process, respectively, types*, γ, τ. However, it is important to note that the interactions between the features are more complex than a simple independent activation, and involve multiple layers in the neural network. This is why we dedicate section "Analysis of the experiment outcome" to an extensive analysis of the interpretability of the model.

OpenReview dataset. As discussed before, here we focused on the prediction of the reviewer average score, using the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) as reference metrics. As baseline we selected a naive classifier that chose the median score of the training dataset. In Fig. 5, the training behaviour of the network is shown, in terms of MSE of the validation and training set over the different training epochs. The performance of the trained regressor is shown in Table 4.

As we can see, the trained regressor is able to improve a naive one.

Even more importantly, Fig. 6 shows the empirical distribution of the MAE over the test set: we can see that 75% of the samples have an error of under 1.2 (over a total score of 10), and the median error is only 0.79. This means that we can expect this model to predict the average reviewers score with a median error of 0.79 over 10. While the reduction in the error rate is promising, it is worth noting that its low magnitude (for both approaches) is in part explained by the low variance of the scores used by reviewers, who tend not to use the whole scale available.

As for the previous dataset, we can observe that the stronger features activating the first neural layer are the Lix index, the Flesch Kincaid grade (Crossley et al., 2011), the text length, number of pages and file size, together with the following list of words: *256, actor, buffer, causal, coefficients, curve, demonstrate, dnn, exploration, github, gpu, idea, imagenet, measures, message, perturbations, precision, produce, quantised, query, regression, review, selected, sentence, standard, state, supervision, tensor, token, width*. We refer to the next section for a more detailed analysis of the explainability of the model.

Table 3 WCNC classification performance vs. random classifier.

Classifier	Accuracy [%]	F1-score [%]	Precision [%]	Recall [%]
Random	~50%	~50%	~50%	~50%
Dense NN	74.01%	72.30%	72.45%	73.19%

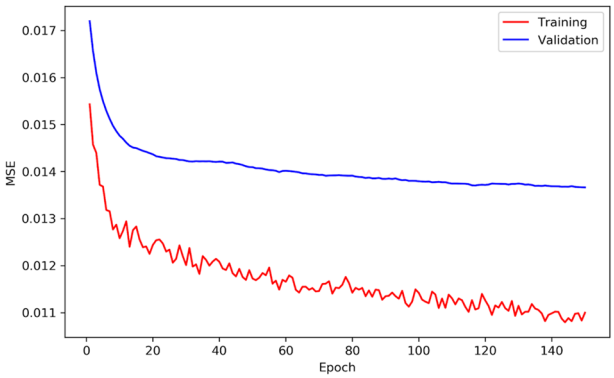


Fig. 5 AI learning process on OpenReview dataset. Mean Squared Error vs. number of training epochs for OpenReview training process, with batches of 32 samples.

Table 4 OpenReview regression performance vs. naive regressor (the lower the better).

Regressor	MAE	MSE
Naive	0.96	1.40
Dense NN	0.79	0.90

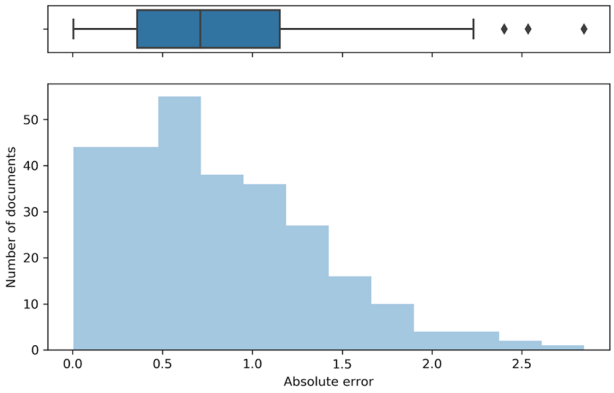


Fig. 6 Absolute error empirical distribution. 75% of the samples have an error of under 1.2, and the median error is 0.79.

Explainability. In the case of black-box models like deep-learning systems, such as the one we developed, it is important to attempt to interpret the reasoning of the model, or in other words, the rationale for an automated decision, to allow practitioners to decide the level of trust given to a model. This is of fundamental importance to reduce the opacity of such tools, enabling an informed evaluation of their performance and, therefore, allowing greater trust in their outputs.

Explaining models depending on half a million parameters is practically impossible using standard tools. For example, the presence of a specific keyword or a specific document statistics can affect the model decision in a non-linear and document-dependent way, making it very challenging to identify a set of simple rules that can make the model understandable to a non-specialist.

However, recent studies have shown that local interpretable model-agnostic explanations (LIME) are able to effectively explain the model decision on a specific document (Ribeiro et al., 2016). The technique is based on the local perturbation of

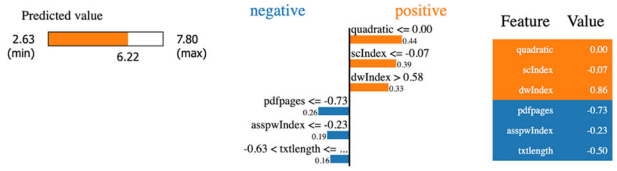


Fig. 7 LIME explanation for a document in the OpenReview dataset. The top features affecting the model are: the word "quadratic", the sentence count `scIndex`, the number of difficult words `dwIndex`, the number of pages `pdfPages`, the average number of syllables per word `asspwIndex`, and the text length `txtlength`.

an instance and the development of a linear model. We used this method to create explanations on multiple documents, and repeat this approach on the whole document space, with the goal of picking a set of exemplar documents distant enough from each other to obtain a representative set of rules of the whole model (submodular pick technique).

In Figs. 7 and 8 examples of the local explanation for an accepted paper of the OpenReview dataset are shown. In orange the top features influencing the decision towards a positive decision are represented, while the blue colour represents factors associated with a negative decision. This summary is simple enough to be presented as is to a non-specialist. In Fig. 7, the absence of the word "quadratic", a low sentence count, and a high number of difficult words positively affects the model score, while a low number of pages, a small number of average syllables per word and a low text length affect the model score negatively. In some cases, the local explanation can expose potential biases or signal overfitting of the model. Overfitting occurs where the fits to the model is based closely on the specifics of the training set but would be a poor fit further related data. For example, in Fig. 8, we can observe that the absence of the word "decoding" is affecting negatively the model decision. This might reasonably be considered an overfitting problem caused by the over-abundance of documents related to decoding in this conference. The choice of whether this contingent explanation is satisfactory is highly context dependent, but it can increase the transparency of the model and allow the practitioners to assess the trust on the model. Another example of overfitting has been observed in the early stages of the model building, using a less than optimal hyperparameter set. In that case, the presence of some specific first names was regarded as a positive signal for the final decision.

Often local rules do not generalise for documents that are considerably different. For example, high text length can increase the predicted score when some keywords are present, while it could be decrease it in other contexts. This is clearly shown after running a submodular pick analysis of the whole space, as shown in Fig. 9, after identifying a group of exemplar documents covering the training space. Some keywords like "hyperparameters" and "quadratic" can be modelled as positive or negative depending on the context of the specific paper.

Analysis of the experiment outcome

Despite the focusing on rather superficial document features, like word distribution, readability and formatting scores, the machine-learning system was often able to successfully predict the peer review outcome. In other words, those documents that scored highly in those areas (e.g., they achieved high scores in readability and were formatted as required) were more likely to be recommended by reviewers for acceptance, and those that achieved lower scores in those areas, more likely to be recommended for rejection. There are a number of possible explanations for this.

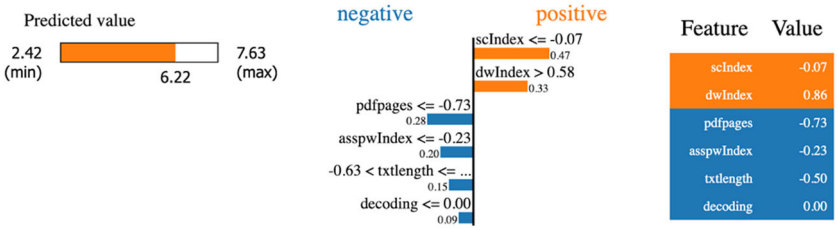


Fig. 8 LIME explanation for a different document in the OpenReview dataset. The top features are: the word "decoding", the sentence count `scIndex`, the number of difficult words `dwIndex`, the number of pages `pdfPages`, the average number of syllables per word `asspwIndex`, and the text length `txtlength`.

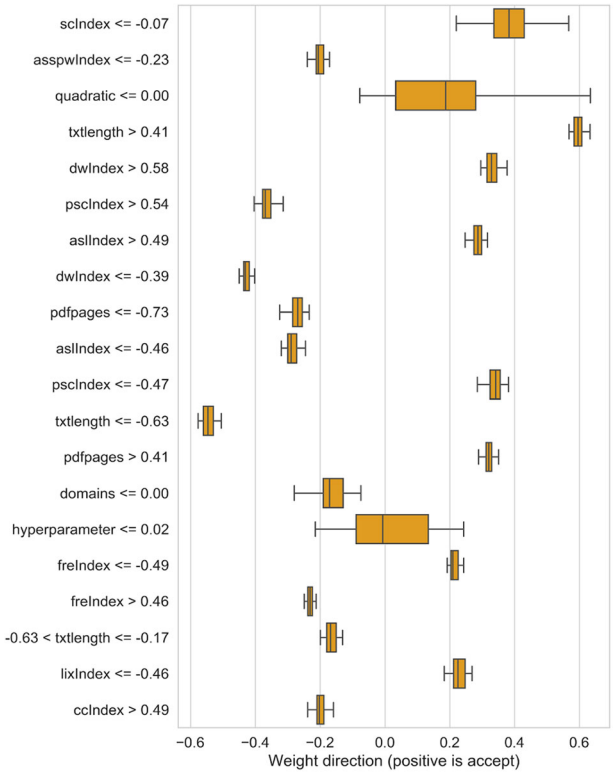


Fig. 9 Distribution of top 20 explanation features for rules covering the training space. Boxplot representing the weight direction for each feature (vertical black line is the median).

One possible explanation is that a correlation between such superficial features and the outcome of the review process as a whole might indicate that they are in fact a good indicator of the overall quality of the paper. In other words, if a paper is presented and reads badly, it is likely to be of lower quality in other, more substantial, ways, making these more superficial features proxy useful metrics for quality. In that case, assessing a paper taking into particular account of those superficial features may be a reasonable heuristic in making overall decisions about the quality of the paper. If that was the case, it would be reasonable to use AI to screen papers before peer review process, using AI as a tool to desk reject papers based on these macroscopic features as part of the pre-peer review screening referred to earlier. This would save the time of peer reviewers who would have to review papers, which were highly likely to low quality. Even if low-scoring papers are not desk rejected, it could be that their scores are flagged to peer reviewers to assist them in making their decisions—also a potential time saver.

Table 5 Potential role of AI in the different dimensions of the peer review process.

Dimension	AI impact
Formatting	High
Plagiarism	High
Scope	High
Readability/English	Medium
Relevance	Medium
Soundness/rigour	Low
Novelty	Low
Impact	Low

However, it may be that papers that score less well on these superficial features create a "first-impression bias" on the part of peer reviewers, who then are more inclined to reject papers based on this negative first-impression derived from what are arguably relatively superficial problems. Reviewers may be unduly influenced by, e.g., formatting or grammatical issues and become unconsciously influenced by this in their judgements of more substantive issues in the submission. Examples of such issues in papers include the presence of typos, the presence of references to papers from regions under-represented in the scientific literature, or the use of methods that have been associated with rejected papers in the past.

In that case, an AI tool that screens papers prior to peer review in the way described, could be used to advise authors to rework their paper before it is sent on for peer review, since it is likely that peer reviewers may reject the paper or at least be negatively influenced by the macroscopic features of the paper, which could be relatively easily corrected.

This might be of particular benefit to authors for whom English is not a first language, for example, and whose work, therefore, may be likely to be adversely affected by first-impression bias.

Discussion

Impact. Table 5 lists several aspects and the potential role how AI-based tools, such as the one we describe in this study, can (or already do) impact the different dimensions of the peer review process.

Tools of this kind have the potential to be of direct benefit in assisting editors of journals and conference proceedings in decision-making (and similarly, the role of making funding decisions, as described in section "Conclusions and future work"). Such tools have the potential to save the time of reviewers, when used as decision support systems. We suggest there may be potential positive impacts in the following specific processes.

Reducing desk rejects. By catching the "first eye impression", the approach we have explored in this paper has the potential to detect early superficial problems, like formatting issues and quality of the graphs. Authors could be made aware of such problems immediately without any further review, or the outcome could be used to pre-empt/inform desk rejects. Even though this technique could wrongly signal high quality (but unusual) typographical choices, a notification about potential issues would help authors to evaluate whether or not they should correct their presentation. Removing superficial problems before peer review could help to avoid reviewer decisions being unduly informed by first-impression biases, and allow them to focus more on the scientific content. On the other hand, AI could also provide inexperienced reviewers with an impartial point of view of the work, providing some performance indicators and synthetic parameters such as a measure of deviation from past authors in terms of style, language and typographic format.

Explaining decisions by data-driven modelling. By analysing review decisions via a data-driven predictor/classifier, it is possible to investigate the extent to which the complex reviewing process can be modelled at scale. Although complex (our preliminary neural network has half a million parameters), such models can be inspected to derive justifications for and explanations of decisions. While completely replicating the cognitive tasks required for the peer review process would be demanding, an analysis of the human decision process through data analysis and AI replication could potentially more easily imitate the more superficial parts of the decision-making processes involved. This could in turn potentially expose biases and similar issues in the decision-making process.

Discovering latent motivations. Motivations behind a decision are not always clear, even to the person making the decision. Producing a model for predictors/classifiers potentially exposes hidden motivations underlying a decision. This idea has been a particular feature of marketing research as a way of identifying and (and then exploiting) "gut reactions". Exposing such motivations in the context of peer reviewing would help reviewers and editors to increase awareness in and transparency of the peer review process, and this may again help to identify possible biases in decision-making.

Bias. Machine-learning techniques are inherently conservative, as they are trained with data from the past. This could lead to bias and other unintended consequences when a tool based on machine learning is used to inform decision-making in the future. For example, papers with characteristics associated with countries historically under-represented in the scientific literature might have a higher rejection rate using AI methods, since the automated reviews may not adequately take account of rising quality of submissions from such sources over time. Biases might also be introduced by the fact that historically, editors have disproportionately selected reviewers from high-income regions of the world, while low-income regions are under-represented among reviewers. The USA dominates the contribution to peer review, with 32.9% of all reviews vs. 25.4% of published article output (Publons, 2018). We suggest that AI systems can be used to expose possible biases and to inform actions taken to prevent their replication in future use of automated tools.

The ethics of (semi) automated peer review. As shown in section "Results", overfitting and other issues with the model we have used could lead to unintended consequences, like the creation of biased rules that could penalise under-represented groups or even individuals if a tool such as the one we have developed is used in peer review. Following the categories delineated by Mittelstadt et al. (2016) and considering the most relevant of them to our study, we can identify three key examples of ethical concerns arising from our work:

Inscrutable evidence leading to opacity. When the link between the original data and the way they affect the model prediction is not easy to interpret, there is a problem of algorithm opacity, that can in turn lead to mistrust towards the algorithm and the data processors. An author will not trust a review if there is no transparency on the rationale for the decision taken. If tools are used to assist in decision-making of the sort we have described in future, it is crucial that there is as great a level of transparency as possible about how the models work to explain and justify decisions made.

Misguided evidence leading to bias. Models are the result of a particular design path, that has been selected following the values and goals of the designer. These values and goals will inevitably be "frozen into the code" (Macnish, 2012). Moreover, models based on machine learning, like the one described in this work, rely on past results (in this case, past reviews), and thus a model may propagate cultural and organisational biases already present in the learning set (Diakopoulos, 2016). Other sources of bias can be technical constraints or emergent contexts of usage. In review systems, a tool such as the one we have developed could in practice adversely affect decisions on papers produced by authors from low-income countries and or those on innovative topics if used without taking such possibilities into account.

Transformative effects leading to challenges for autonomy. Even using such models only to signal problematic papers or to assist reviewers could affect the agency of reviewers by creating new forms of understanding and conceptualisation. This may result in a specular effect to the one discussed in the previous point: the way the model interprets the manuscript could propagate to the reviewer, potentially creating an unintended biased outcome. For example, should the model identify as potential issues the presence of typos, the presence of references to papers from under-represented regions, or the usage of techniques that have been associated with previously rejected papers, the potential effect of the signalling of such issues to the reviewers could be an increase of the importance of such factors in the mind of the reviewers and influence their authority bias/status quo bias.

All of these ethical concerns need to be considered carefully in the way AI tools are designed and deployed in practice, and in determining the role they play in decision-making. Continued research in these areas is crucial in helping to ensure that the role AI tools play processes like peer review is a positive one.

Conclusions and future work

In this paper, we have reported an experiment involving three peer-reviewed conference proceedings, training a machine-learning system to infer a set of rules able to match the peer review outcome, ultimately providing an acceptance probability for other manuscripts. We focused on a rather superficial set of features of the submitted manuscripts, like word distribution, readability scores and document format.

Nevertheless, the machine-learning system was often able to successfully predict the peer review outcome: we found a strong correlation between such superficial features and the outcome of the review process as a whole.

We have seen how tools could be developed based on such systems, which could be used to create greater efficiency in the quality control and peer review process. We have also seen how

such tools could be used to gain insight on the reviewing process: our results suggest that such tools can create measurable benefits for scientometric studies, because of their explainability capability.

While the application of such AI tools is still in its infancy, we have observed some of the possible implications in terms of biases and ethics. Our findings point in the direction of a new type of analysis of typical human process, conducted with the help of machine-learning systems, one which is cognisant of ethical dimensions of the work as well as technical capabilities.

The following future work is suggested.

Feedback loop. We are interested in exploring the behaviour of reviewers when using these AI-powered support tools. We intend in future to carry out controlled experiments with academic reviewers, to understand the biases introduced by the AI signals on the reviewers. As discussed in "The ethics of (semi) automated peer review", understanding potential effects on the reviewers is fundamental to ensuring ethical usage of such tools.

Review process. When using openreview.net, we would be interested in taking into account the full text of the review itself (rather than only the review outcome) to better train the AI tools. A great deal of useful information is contained in the text of the reviews and rebuttals that inform the final decision.

Perception. Work on the first-impression bias needs to be extended, including more complex typographic layout indicators. Similarly, a more detailed analysis of the model could expose additional decision rules like language issues and formatting issues.

Disciplinary variation. We would like to explore how the design and application of AI tools carrying out semi-automated quality assessment can take place in the context of different disciplines, taking into account different disciplinary norms of communicating research results.

Grant applications. Funders might use such decision support systems to assess grant applications. Grant applications have a different structure (as they are proposing projects not reporting them), thus the content heterogeneity might be higher. We plan to investigate further the application of the methods discussed here to that domain.

Data availability

The datasets generated during the current study are not publicly available due to de-anonymisation risks, but are available from the corresponding author on reasonable request. The OpenReview data are available on https://openreview.net/.

Received: 23 December 2019; Accepted: 29 October 2020;
Published online: 25 January 2021

Notes

1 Openreview is available at https://openreview.net/about.
2 Tokenisation is a standard machine-learning process, which consists in chopping a text into words (tokens), throwing away punctuation.
3 In statistics, F1-score is a measure of a test's accuracy, which considers both the precision and the recall.

References

AJE (2018) Peer review: how we found 15 million hours of lost time. URL https://www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time, Accessed 20 Dec 2019
Akst J (2016) Researchers to CIHR: reverse peer review changes. URL https://www.the-scientist.com/the-nutshell/researchers-to-cihr-reverse-peer-review-changes-33236.
Ashley KD, Goldin IM (2011) Toward AI-enhanced computer-supported peer review in legal education. In: Biswas G, Bull S, Kay J, Mitrovic A (eds) JURIX. pp. 3–12
Balfour SP (2013) Assessing writing in MOOCs: automated essay scoring and calibrated peer review. Res Pract Assess 8:40–48
Barocas S, Selbst AD (2016) Big data's disparate impact. Cal Law Rev 104:671
Bornmann L (2011) Scientific peer review. Ann Rev Inform Sci Technol 45:197–245
Colman AM (1982) Manuscript evaluation by journal referees and editors: randomness or bias? Behav Brain Sci 5:205–206
Cozza V, Petrocchi M, Spognardi A (2016) A matter of words: NLP for quality evaluation of Wikipedia medical articles. In: Bozzon A, Cudré-Maroux P, Pautasso C (eds) International Conference on Web Engineering. Springer, pp. 448–456
Crossley SA, Allen DB, McNamara DS (2011) Text readability and intuitive simplification: a comparison of readability formulas. Read Foreign Lang 23:84–101
Crossley SA, Greenfield J, McNamara DS (2008) Assessing text readability using cognitively based indices. Tesol Quart 42:475–493
Cyranoski D (2019) Artificial intelligence is selecting grant reviewers in China. URL https://www.nature.com/articles/d41586-019-01517-8, Accessed 20 Dec 2019
D'Andrea R, O'Dwyer JP (2017) Can editors save peer review from peer reviewers? PLoS ONE 12:e0186111
Diakopoulos N (2016) Accountability in algorithmic decision making. Commun ACM 59:56–62
Florea L et al. (2019) From first impressions to selection decisions: the role of dispositional cognitive motivations in the employment interview. Person Rev 48:249–272
Ford E (2013) Defining and characterizing open peer review: a review of the literature. J Scholar Publish 44:311–326
Garcia M (2016) Racist in the machine: the disturbing implications of algorithmic bias. World Policy J 33:111–117
Goldin IM, Ashley KD (2011) Peering inside peer review with bayesian models. In: Biswas G, Bull S, Kay J and Mitrovic A (eds) International Conference on Artificial Intelligence in Education. Springer, pp. 90–97
Güçlütürk Y et al. (2017) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9:316–329
Harnad S (1999) Free at last: the future of peer-reviewed journals. D-Lib Magaz 5:12
Harnad S (1998) The invisible hand of peer review. Nature 5. https://doi.org/10.1038/nature28029.
Hojat M, Gonnela JS, Caelleigh AS (2003) Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. Adv Health Sci Educ 8:75–96
Horbach SPJM, Halffman W (2019) The ability of different peer review procedures to flag problematic publications. Scientometrics 118:339–373
Lee CJ et al. (2013) Bias in peer review. J Am Soc Inform Sci Technol 64:2–17
Lindgaard G et al. (2006) Attention web designers: you have 50 milliseconds to make a good first impression! Behav Inform Technol 25:115–126
Macnish K (2012) Unblinking eyes: the ethics of automating surveillance. Ethics Inform Technol 14:151–167
Mittelstadt BD et al. (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3:68
Moys JL (2014) Typographic layout and first impressions: testing how changes in text layout influence reader's judgments of documents. Vis Lang 48(1): 881
Mulligan A, Hall L, Raphael E (2013) Peer review in a changing world: an international study measuring the attitudes of researchers. J Am Socr Inform Sci Technol 64:132–161
Nicholas D et al. (2015) Peer review: still king in the digital age. Learn Publ 28:15–21
Nuijten MB, Van Assen MALA, Hartgerink CHJ, Epskamp S, Wicherts JM et al. (2017) The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. Preprint retrieved from https://psyarxiv.com/tcxaj/
Piech C, Huang J, Chen Z et al. (2013) Tuned models of peer assessment in MOOCs. In: D'Mello SK, Calvo RA and Olney A (eds) 6th International Conference on Educational Data Mining (EDM 2013). International Educational Data Mining Society, pp. 153–160
Pontille D, Torny D (2015) From manuscript evaluation to article valuation: the changing technologies of journal peer review. Human Stud 38:57–79
Preim J (2013) Beyond the paper. Nature 495:437–440
Publons (2018) Global state of peer review 2018. URL https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf, Accessed 20 Dec 2019.
Ribeiro MT, Singh S, Guestrin, C (2016) Why should I trust you?: Explaining the predictions of any classifier. In: Balaji K, Mohak S (eds) Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp. 1135–1144

Ross-Hellauer T (2017) What is open peer review? A systematic review. F1000Res 6.588. https://doi.org/10.12688/f1000research.11369.2
Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach. Pearson Education Limited, Malaysia
Schmidt T (2017) Statcheck does not work: All the numbers. Reply to Nuijten et al. (2017). PsyArXiv. http://psyarxiv.com/hr6qy.
Shanahan D (2016) A peerless review? Automating methodological and statistical review. Springer Nature BioMed Central, Research in progress blog. Available at: https://blogs.biomedcentral.com/bmcblog/2016/05/23/peerless-review-automating-methodological-statistical-review Accessed 6 Jan 2020
Siler K, Lee K, Bero L (2015) Measuring the effectiveness of scientific gatekeeping. Proc Natl Acad Sci 112:360–365
Sonntag D (2004) Assessing the quality of natural language text data. In: Dadam P, Reichert M (eds) GI Jahrestagung. pp. 259–263
Spezi V et al. (2018) Let the community decide? The vision and reality of soundness-only peer review in open-access mega-journals. J Document 74:137–161
SWEOR (2019) 27 eye-opening website statistics: is your website costing you clients? URL https://www.sweor.com/firstimpressions, Accessed 20 Dec 2019
Tennant JP (2018) The state of the art in peer review. FEMS Microbiol Lett 365 (19). https://doi.org/10.1093/femsle/fny204.
Tennant JP, Dugan JM, Graziotin D et al. (2017) A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Res 6:1151. https://doi.org/10.12688/f1000research.12037.3
Wood TJ et al. (2018) Can physician examiners overcome their first impression when examinee performance changes? Adv Health Sci Educ 23:721–732
Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci Technol Human Value 41:118–132

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.C., L.B. or S.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.