E-ISSN: 2964-6219, P-ISSN: 2964-6200
Volume 4 Number 2 August 2026.
pp: 238-248. DOI: https://doi.org/10.58477/cj.v4i2.462.

Computer Journal

RESEARCH ARTICLE Open Access

Fine-Tuning Model *Generative Pre-Trained Transformer 2* (GPT-2) untuk Pembuatan Soal-Jawab Otomatis pada Materi Ilmu Pengetahuan Alam dan Sosial (IPAS) di SD Negeri 1 Wawoangi

Anisa Agustina [1*], **Muhammad Fuad** [2]

[1*,2] Program Studi Sistem Dan Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Muslim Buton, Kota Baubau, Provinsi Sulawesi Tenggara, Indonesia.

*Correspondence email: anisaagustina868@gmail.com.

Received: 31 July 2026.
Revised: 1 August 2026.
Accepted: 25 August 2026.
Published: 30 August 2026.

Full list of author information is available at the end of the article.

Abstract

The increasing administrative workload of teachers makes manual preparation of assessment questions time-consuming and may lead to the reuse of questions from previous years. This study examines the application of the transformer-based Generative Pre-trained Transformer 2 (GPT-2) model for automatic question-and-answer generation in elementary school Natural and Social Sciences (IPAS). The model was fine-tuned using 239 Indonesian context-question-answer triples, with 90% used for training and 10% for testing. Fine-tuning was conducted using four dataset sizes: 50, 100, 150, and 239 samples, to examine the relationship between training data volume and generation quality. Model performance was evaluated by examining parameter changes between the pre-trained and fine-tuned models and using the ROUGE metric to measure textual similarity between generated and reference questions and answers. The 239-sample dataset produced the most coherent and contextually appropriate questions, with ROUGE-1, ROUGE-2, and ROUGE-L scores of 1.0 for questions and 0.41, 0.24, and 0.36 for answers. These findings suggest that GPT-2 fine-tuning can support automatic question-generation tools when sufficient training data are available.

Keywords: Question Creation; GPT-2; Transformer; Fine-Tuning; ROUGE.

Abstrak

Beban administratif guru membuat penyusunan soal evaluasi secara manual menjadi kurang efisien dan dapat mendorong penggunaan ulang soal dari tahun sebelumnya. Penelitian ini mengkaji penerapan model Generative Pre-trained Transformer 2 (GPT-2) untuk pembuatan soal dan jawaban otomatis pada materi Ilmu Pengetahuan Alam dan Sosial (IPAS) sekolah dasar. Model GPT-2 di-fine-tuning menggunakan 239 triplet konteks-pertanyaan-jawaban berbahasa Indonesia, dengan 90% data digunakan sebagai data latih dan 10% sebagai data uji. Fine-tuning dilakukan pada empat ukuran dataset, yaitu 50, 100, 150, dan 239 data, untuk mengamati hubungan antara volume data latih dan kualitas keluaran. Kinerja model dievaluasi melalui perubahan parameter antara model pra-latih dan model hasil fine-tuning serta metrik ROUGE untuk mengukur kemiripan tekstual antara hasil generasi dan referensi. Dataset 239 data menghasilkan soal yang paling koheren dan sesuai konteks, dengan skor ROUGE-1, ROUGE-2, dan ROUGE-L sebesar 1,0, sedangkan jawaban memperoleh skor 0,41, 0,24, dan 0,36. Hasil ini menunjukkan bahwa fine-tuning GPT-2 berpotensi mendukung pembuatan soal otomatis dengan data latih yang memadai.

Kata Kunci: Pembuatan Soal; GPT-2; Transformer; Fine-Tuning; ROUGE.

1. Pendahuluan

Pemanfaatan teknologi digital dalam pendidikan terus berkembang seiring dengan kebutuhan terhadap proses pembelajaran dan evaluasi yang lebih efektif. Evaluasi pembelajaran dilakukan untuk mengetahui tingkat pencapaian siswa melalui penilaian terhadap jawaban yang diberikan berdasarkan soal yang disusun oleh guru (Owan *et al*., 2023; Zhang & Aslan, 2021). Perkembangan teknologi juga membuka peluang untuk mendukung berbagai aktivitas pendidikan, termasuk pemanfaatan teknologi berbasis kecerdasan buatan pada bidang pendidikan kedokteran (Kung *et al*., 2023). Dalam pelaksanaannya, penyusunan soal membutuhkan pemahaman terhadap materi, tujuan pembelajaran, dan tingkat kemampuan siswa. Proses tersebut juga membutuhkan waktu, terutama ketika guru harus menyusun sejumlah soal yang bervariasi untuk digunakan dalam evaluasi. Tingginya beban administratif dapat membatasi waktu guru dalam mengembangkan soal baru sehingga penggunaan kembali soal dari tahun sebelumnya masih dapat terjadi (Wijaya & Jati, 2022). Kondisi tersebut menunjukkan adanya kebutuhan terhadap teknologi yang dapat membantu guru dalam menghasilkan soal secara lebih efisien tanpa menghilangkan kesesuaian soal dengan materi pembelajaran. Pembangkitan soal secara otomatis (*automatic question generation*) telah dikembangkan melalui berbagai pendekatan pemrosesan bahasa alami dan pembelajaran mesin. Pendekatan awal memanfaatkan teknik NLP konvensional, seperti *Term Frequency-Inverse Document Frequency* (TF-IDF) dan N-gram untuk mengidentifikasi informasi yang dapat digunakan dalam pembentukan pertanyaan (Nwafor & Onyenwe, 2021). Perkembangan berikutnya memanfaatkan pendekatan *Recurrent Neural Network* (RNN) untuk menghasilkan pertanyaan berdasarkan informasi atau konteks tertentu (Wijaya & Jati, 2022). Model berbasis *transformer* kemudian digunakan karena mampu memanfaatkan mekanisme *attention* untuk mempelajari hubungan antarkata dalam suatu rangkaian teks. Salah satu model yang berkembang dari arsitektur tersebut adalah *Text-to-Text Transfer Transformer* (T5), yang telah digunakan untuk pembangkitan soal dalam bahasa Inggris (Chomphooyod *et al*., 2023). Dalam konteks bahasa Indonesia, penelitian pembangkitan soal otomatis juga telah berkembang melalui pendekatan *sequence-to-sequence*. Vincentio dan Suhartono (2022) membandingkan beberapa model generatif, yaitu IndoBART, IndoGPT, mBART, dan mT5, dalam konteks pembangkitan soal berbahasa Indonesia. Perkembangan tersebut menunjukkan bahwa model bahasa berbasis transformer memiliki potensi untuk digunakan dalam menghasilkan teks secara otomatis pada bahasa Indonesia.

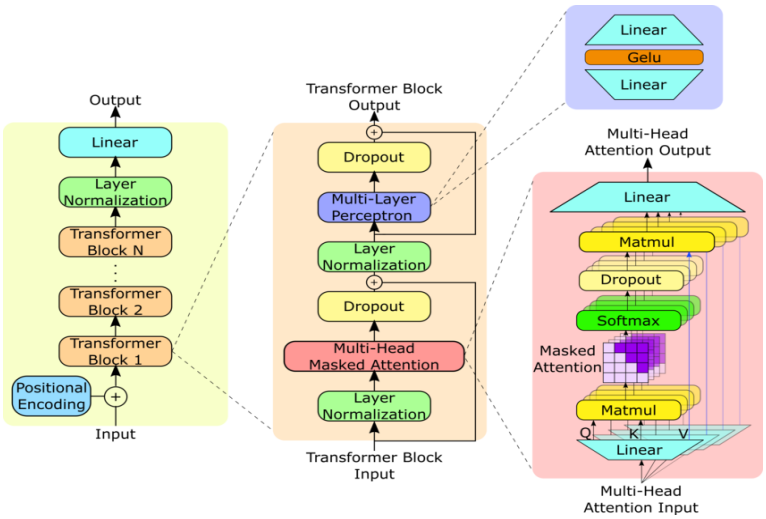
Salah satu model generatif yang dapat digunakan adalah *Generative Pre-trained Transformer 2* (GPT-2), yaitu model bahasa autoregresif yang memanfaatkan arsitektur transformer dan dilatih melalui pendekatan *unsupervised learning* pada data teks dalam jumlah besar (Radford *et al*., 2019). Karakteristik GPT-2 memungkinkan model yang telah melalui tahap *pre-training* untuk kemudian disesuaikan melalui *fine-tuning* dengan dataset yang lebih spesifik sesuai dengan tujuan aplikasi. Dalam penelitian ini, karakteristik tersebut dipertimbangkan karena dataset yang digunakan berjumlah 239 triplet konteks-pertanyaan-jawaban. Penelitian mengenai pembangkitan soal pada konteks pendidikan di Indonesia juga telah dilakukan dengan pendekatan yang berbeda. Flores *et al*. (2021), misalnya, mengembangkan pembangkitan soal IPA sekolah dasar dengan memanfaatkan taksonomi Bloom sebagai dasar dalam menentukan tingkat kognitif soal. Pendekatan tersebut menunjukkan bahwa pembangkitan soal tidak hanya berkaitan dengan kemampuan menghasilkan teks, tetapi juga perlu mempertimbangkan kesesuaian soal dengan materi pembelajaran dan karakteristik peserta didik. Di sisi lain, penelitian terkait evaluasi model generatif dan penggunaan metrik ROUGE menunjukkan pentingnya pengukuran kemiripan antara teks yang dihasilkan dan teks referensi (Naufal & Kusuma, 2023). Meskipun demikian, penerapan model transformer generatif untuk pembangkitan soal sekaligus jawaban esai berbahasa Indonesia pada materi IPAS sekolah dasar masih terbatas. Berdasarkan kondisi tersebut, penelitian ini menerapkan GPT-2 yang di-*fine-tuning* menggunakan dataset berbahasa Indonesia yang terdiri atas 239 triplet konteks-pertanyaan-jawaban dari materi IPAS sekolah dasar. Penelitian dilakukan dengan membandingkan beberapa ukuran dataset, yaitu 50, 100, 150, dan 239 data, untuk mengamati perbedaan kualitas keluaran berdasarkan volume data latih. Selain mengevaluasi hasil pembangkitan soal dan jawaban menggunakan metrik ROUGE, penelitian ini juga membandingkan perubahan parameter antara model GPT-2 pra-latih dan model setelah *fine-tuning*. Dengan demikian, penelitian ini berfokus pada penerapan dan evaluasi GPT-2 untuk mendukung pembangkitan soal dan jawaban otomatis pada materi IPAS sekolah dasar, khususnya pada kondisi dataset yang relatif terbatas.

2. Tinjauan Pustaka

Pembangkitan soal otomatis (*automatic question generation*) telah dikembangkan melalui berbagai pendekatan, mulai dari teknik pemrosesan bahasa alami konvensional hingga model berbasis jaringan saraf dan *transformer*. Pendekatan berbasis NLP konvensional, seperti TF-IDF dan N-gram, digunakan untuk mengidentifikasi kata atau informasi penting dari teks sumber yang kemudian dimanfaatkan dalam pembentukan pertanyaan berbasis aturan (Nwafor & Onyenwe, 2021). Pendekatan tersebut relatif sederhana dari sisi komputasi, tetapi kualitas pertanyaan bergantung pada aturan dan pola bahasa yang telah ditentukan. Perkembangan berikutnya menggunakan arsitektur *Recurrent Neural Network* (RNN) yang memungkinkan model mempelajari pola bahasa berdasarkan data pelatihan tanpa sepenuhnya bergantung pada aturan yang ditentukan secara manual (Wijaya & Jati, 2022). Meskipun demikian, RNN memiliki keterbatasan dalam mempertahankan informasi pada ketergantungan yang panjang dalam suatu rangkaian teks. Selanjutnya, model berbasis *transformer* digunakan dalam berbagai tugas pembangkitan teks dengan memanfaatkan mekanisme *attention* untuk mempelajari hubungan antarbagian teks. Model *transformer* mencakup arsitektur *encoder-decoder*, seperti *Text-to-Text Transfer Transformer* (T5), maupun arsitektur *decoder-only*, seperti GPT-2. Penelitian Chomphooyod *et al*. (2023) menunjukkan bahwa T5 dapat digunakan untuk pembangkitan soal pada konteks tata bahasa Inggris dan menghasilkan soal yang dapat diterima berdasarkan evaluasi yang dilakukan dalam penelitian tersebut. Dalam konteks bahasa Indonesia, Vincentio dan Suhartono (2022) membandingkan beberapa pendekatan pembangkitan soal, termasuk model berbasis RNN seperti Bi-LSTM dan Bi-GRU serta model *transformer* pra-latih seperti mBART, mT5, IndoBART, dan IndoGPT. Eksperimen tersebut menggunakan dataset TyDiQA dan SQuAD yang diterjemahkan ke dalam bahasa Indonesia. Hasil penelitian menunjukkan bahwa model *transformer* yang diuji menghasilkan kinerja yang lebih baik dibandingkan model RNN pada konfigurasi dan dataset yang digunakan, dengan IndoBART memperoleh hasil terbaik di antara model yang dibandingkan. Temuan tersebut menunjukkan bahwa pemilihan arsitektur dan model pra-latih merupakan salah satu aspek penting dalam pengembangan sistem pembangkitan soal berbahasa Indonesia. GPT-2 merupakan model bahasa generatif berbasis *transformer* dengan arsitektur *decoder-only* yang dikembangkan untuk menghasilkan teks berdasarkan konteks sebelumnya (Radford *et al*., 2019). Model ini dapat disesuaikan dengan tugas tertentu melalui proses *fine-tuning* menggunakan dataset yang relevan dengan tujuan aplikasi. Dalam penelitian ini, GPT-2 dipilih untuk dieksplorasi pada dataset yang relatif terbatas, yaitu 239 triplet konteks-pertanyaan-jawaban. Pemilihan tersebut merupakan pertimbangan eksperimental yang disesuaikan dengan karakteristik dataset dan tujuan penelitian, sehingga tidak dimaksudkan untuk menyatakan bahwa GPT-2 secara umum lebih unggul dibandingkan model generatif berbahasa Indonesia lainnya. Kemampuan model generatif berbasis GPT juga telah diterapkan pada berbagai domain. Paik dan Wang (2021) menerapkan pendekatan GPT-2 untuk pembangkitan kode program, sedangkan Sufi (2024) membahas pemanfaatan GPT-2 dalam pembangkitan teks. Pada domain kimia, Bagal *et al*. (2022) menggunakan model generatif berbasis GPT untuk menghasilkan struktur molekul menggunakan representasi SMILES. Berbagai penerapan tersebut menunjukkan bahwa model generatif berbasis *transformer* dapat diadaptasi untuk menghasilkan teks atau representasi terstruktur pada domain yang berbeda. Namun, penerapan GPT-2 untuk pembangkitan soal dan jawaban berbahasa Indonesia pada materi IPAS sekolah dasar memiliki karakteristik data dan kebutuhan yang berbeda sehingga perlu dievaluasi secara khusus. Evaluasi hasil pembangkitan teks dapat dilakukan menggunakan metrik otomatis, salah satunya ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). ROUGE digunakan untuk mengukur kemiripan antara teks hasil generasi dan teks referensi. Beberapa varian yang umum digunakan adalah ROUGE-1 yang mengukur kesamaan *unigram*, ROUGE-2 yang mengukur kesamaan *bigram*, dan ROUGE-L yang didasarkan pada *longest common subsequence*. Metrik tersebut memberikan ukuran kuantitatif terhadap tingkat kemiripan teks yang dihasilkan model dengan teks referensi dan dapat digunakan pada tugas pembangkitan teks berbasis *transformer*. Penelitian terkait pembangkitan soal menggunakan T5 juga menunjukkan penggunaan evaluasi berbasis teks selain penilaian manusia (Chomphooyod *et al*., 2023). Sementara itu, Naufal dan Kusuma (2023) membahas perkembangan otomatisasi pembangkitan pertanyaan berbahasa Indonesia dengan berbagai pendekatan, termasuk model berbasis *transformer*. Hal tersebut menunjukkan bahwa evaluasi hasil pembangkitan perlu disesuaikan dengan karakteristik teks dan tujuan tugas yang dilakukan. Dalam penelitian ini, ROUGE-1, ROUGE-2, dan ROUGE-L digunakan untuk mengukur kemiripan tekstual antara pertanyaan dan jawaban yang dihasilkan oleh GPT-2 dengan teks referensi. Penggunaan ketiga varian tersebut memungkinkan evaluasi pada tingkat kesamaan kata, pasangan kata, dan kesamaan urutan teks. Meskipun demikian, skor ROUGE merepresentasikan kemiripan tekstual dan tidak secara langsung menunjukkan kualitas pedagogis, kesesuaian tingkat kognitif, atau validitas suatu soal. Oleh karena itu, hasil ROUGE dalam penelitian ini digunakan sebagai ukuran kuantitatif terhadap kesamaan teks hasil generasi dengan teks referensi, sedangkan kualitas soal tetap perlu dipertimbangkan berdasarkan konteks materi IPAS dan karakteristik dataset yang digunakan.

3. Metode

Penelitian ini menggunakan pendekatan eksperimental dengan menerapkan model *transformer* GPT-2 untuk pembuatan soal dan jawaban secara otomatis. Sistem dirancang melalui beberapa tahapan, yaitu guru melakukan autentikasi ke dalam sistem, memasukkan teks konteks materi pembelajaran, kemudian sistem melakukan pemrosesan terhadap teks tersebut menggunakan pendekatan pemrosesan bahasa alami yang umum diterapkan pada sistem pembelajaran berbasis NLP (Fanni *et al*., 2023). Pendekatan tersebut sejalan dengan pemanfaatan model berbasis *transformer* untuk pemrosesan teks berbahasa Indonesia (Baharuddin & Naufal, 2023). Selanjutnya, model GPT-2 hasil *fine-tuning* digunakan untuk menghasilkan pasangan soal dan jawaban berdasarkan konteks yang diberikan. Hasil pembangkitan kemudian dapat ditinjau, disimpan, dan diunduh oleh pengguna. GPT-2 merupakan model bahasa generatif berbasis *transformer* dengan arsitektur *decoder-only*. Arsitektur tersebut memanfaatkan mekanisme *masked self-attention* untuk menghasilkan token secara berurutan berdasarkan token sebelumnya, kemudian dilanjutkan dengan jaringan *feed-forward* dan normalisasi pada setiap blok *transformer* (Sajun *et al*., 2024). Kemampuan generatif model berbasis GPT-2 telah diterapkan pada berbagai tugas, termasuk pembangkitan kode program (Paik & Wang, 2021) dan pembangkitan teks (Sufi, 2024). Gambar 1 menunjukkan skema arsitektur model GPT-2 yang digunakan dalam penelitian ini.



Gambar 1. Arsitektur model GPT-2 (diadaptasi dari Radford *et al*., 2019)

Dataset penelitian terdiri atas 239 triplet konteks-pertanyaan-jawaban yang mencakup materi mata pelajaran IPAS di SD Negeri 1 Wawoangi. Dataset bersumber dari buku soal, bank soal daring, dan materi yang telah disetujui oleh dinas pendidikan setempat. Untuk menjaga kesesuaian isi dengan kurikulum, seluruh triplet ditinjau oleh satu guru kelas berpengalaman di SD Negeri 1 Wawoangi. Peninjauan dilakukan untuk memverifikasi kesesuaian soal dan jawaban dengan kurikulum IPAS sekolah dasar, kebenaran konsep sains dan sosial, kesesuaian tingkat kesulitan dengan jenjang kelas sasaran, serta konsistensi penggunaan bahasa Indonesia pada pasangan konteks-pertanyaan-jawaban. Data yang tidak memenuhi kriteria tersebut direvisi oleh peneliti berdasarkan masukan guru atau dikeluarkan dari dataset apabila tidak dapat diperbaiki. Dataset yang telah melalui proses tersebut kemudian digunakan pada tahap *fine-tuning*. Contoh dataset penelitian disajikan pada Tabel 1.

Tabel 1. Contoh dataset penelitian

No	Konteks	Soal	Jawaban
1	Cahaya merupakan bentuk energi yang bergerak dalam garis lurus dan dapat dipantulkan, dibiaskan, serta diuraikan.	Sebutkan sifat-sifat cahaya!	Cahaya merambat lurus, dapat dipantulkan, menembus benda bening, dibiaskan, dan diuraikan.
2	Cahaya merupakan bentuk energi yang bergerak dalam garis lurus dan dapat dipantulkan, dibiaskan, serta diuraikan.	Bagaimana cahaya merambat?	Cahaya merambat lurus.

...	...	...	...
239	Individu berperan penting dalam menjaga keberlanjutan Bumi melalui tindakan sehari-hari yang mengurangi dampak negatif terhadap lingkungan.	Apa peran individu dalam menjaga keberlanjutan Bumi?	Individu dapat menghemat energi, mendaur ulang, dan mengurangi sampah plastik untuk melindungi lingkungan.

Sumber: Data penelitian, diolah oleh penulis.

Dataset dibagi menjadi data latih dan data uji dengan rasio 90:10, masing-masing sebanyak 215 data latih dan 24 data uji. Pembagian dilakukan secara acak menggunakan *random state* untuk memastikan proses pengacakan dapat direproduksi. Tahap prapemrosesan menggunakan teknik *Byte Pair Encoding* (BPE) yang merupakan tokenisasi bawaan GPT-2. Teknik tersebut memecah teks menjadi unit token atau subkata sehingga teks dapat direpresentasikan dalam bentuk token yang dapat diproses oleh model GPT-2. Proses *fine-tuning* dilakukan menggunakan pustaka *Hugging Face Transformers*. Konfigurasi *hyperparameter* yang digunakan selama proses pelatihan disajikan pada Tabel 2.

Tabel 2. Konfigurasi hyperparameter fine-tuning

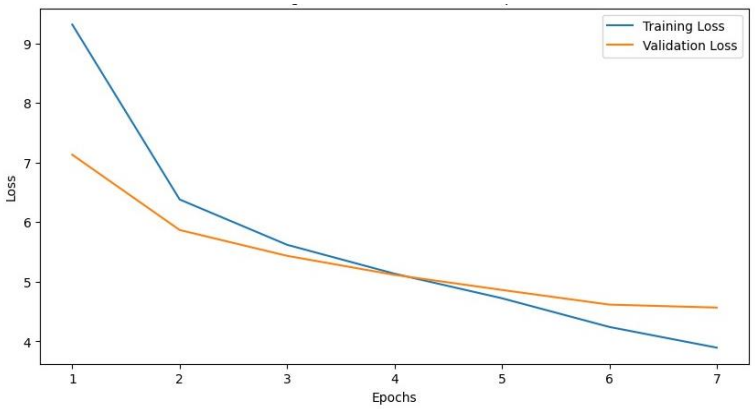
Hyperparameter	Nilai
Learning rate	1×10^{-5}
Batch size	4
Jumlah epoch	7
Weight decay	0,1
Scheduler learning rate	Linear
Warm-up steps	500
Early stopping (toleransi)	2 evaluasi

Untuk menguji pengaruh ukuran dataset terhadap kualitas hasil generasi, proses *fine-tuning* dilakukan menggunakan empat ukuran dataset, yaitu 50, 100, 150, dan 239 data. Evaluasi kinerja model dilakukan melalui dua pendekatan, yaitu membandingkan distribusi rata-rata bobot parameter antara model *pre-trained* dan model hasil *fine-tuning*, serta mengukur kemiripan tekstual antara soal dan jawaban hasil generasi dengan teks referensi menggunakan metrik ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), yang meliputi ROUGE-1, ROUGE-2, dan ROUGE-L. Prototipe sistem dikembangkan menggunakan PHP dan Python dengan framework Laravel, basis data MySQL, serta lingkungan pengembangan XAMPP dan Visual Studio Code. Penelitian dilaksanakan selama tiga bulan, yaitu April hingga Juni 2026, dengan lokasi penelitian di SD Negeri 1 Wawoangi, Kecamatan Sampolawa, Kabupaten Buton Selatan.

4. Hasil dan Pembahasan

4.1 Hasil

Fine-tuning model GPT-2 dilakukan menggunakan empat variasi ukuran dataset, yaitu 50, 100, 150, dan 239 data, untuk mengamati perubahan hasil pembangkitan soal dan jawaban berdasarkan jumlah data yang digunakan dalam pelatihan. Pada setiap skema pelatihan, nilai *training loss* dan *validation loss* menunjukkan tren penurunan dari *epoch* pertama hingga *epoch* ketujuh sebagaimana ditunjukkan pada Gambar 2. Berdasarkan kurva tersebut, tidak terlihat indikasi *overfitting* yang jelas pada rentang *epoch* yang diamati.



Gambar 2. Kurva training loss dan validation loss selama fine-tuning
Sumber: Data penelitian, diolah oleh penulis.

Hasil pembangkitan menunjukkan adanya perbedaan keluaran pada setiap ukuran dataset. Pada penggunaan 50 dan 100 data, model belum menghasilkan soal dan jawaban yang koheren. Keluaran yang dihasilkan masih menunjukkan pengulangan kata dan belum membentuk kalimat tanya yang bermakna. Pada penggunaan 239 data, keluaran yang dihasilkan menunjukkan bentuk pertanyaan yang lebih jelas dan sesuai dengan konteks. Salah satu contoh keluaran adalah pertanyaan "Jelaskan apa yang dimaksud dengan rabun jauh!" yang dihasilkan berdasarkan konteks mengenai gangguan penglihatan. Perbandingan distribusi rata-rata parameter antara model *pre-trained* dan model hasil *fine-tuning* menggunakan 239 data disajikan pada Tabel 3.

Tabel 3. Perbandingan parameter model pre-trained dan fine-tuned

Parameter	Pre-trained (mean)	Fine-tuned (mean)	Selisih
transformer.h.0.attn.c_attn.bias	0	0,00072	0,0585
transformer.h.1.ln_1.weight	1	0,9157	0,0843
transformer.h.6.ln_1.weight	1	1,1044	0,1050
transformer.h.9.ln_2.weight	1	1,1964	0,1969
transformer.ln_f.weight	1	2,5157	1,5157

Sumber: Data penelitian, diolah oleh penulis.

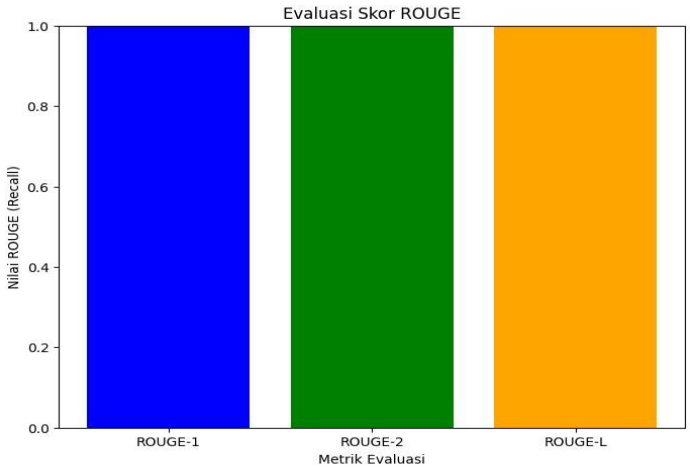
Perubahan paling signifikan terjadi pada parameter di lapisan-lapisan akhir arsitektur, khususnya pada bobot normalisasi akhir (*transformer.ln_f.weight*), yang bergeser dari nilai rata-rata 1,0 pada kondisi *pre-trained* menjadi 2,52 setelah *fine-tuning*, dengan selisih 1,52. Sebaliknya, parameter pada lapisan-lapisan awal, seperti bias pada lapisan *attention* pertama, mengalami perubahan yang lebih kecil. Pola perubahan parameter tersebut menunjukkan adanya penyesuaian bobot model setelah proses *fine-tuning* terhadap dataset yang digunakan.

Tabel 4. Hasil evaluasi ROUGE untuk soal hasil generasi

Metrik Evaluasi	Nilai	Keterangan
ROUGE-1	1,0	Kesamaan *unigram* antara teks hasil generasi dan referensi
ROUGE-2	1,0	Kesamaan *bigram* antara teks hasil generasi dan referensi
ROUGE-L	1,0	Kesamaan berdasarkan *longest common subsequence* antara teks hasil generasi dan referensi

Sumber: Data penelitian, diolah oleh penulis.

Hasil pada Tabel 4 menunjukkan bahwa ketiga metrik ROUGE, yaitu ROUGE-1, ROUGE-2, dan ROUGE-L, memperoleh nilai 1,00 pada komponen soal. Nilai tersebut menunjukkan tingkat kemiripan tekstual yang sangat tinggi antara soal hasil generasi model dan soal referensi pada data uji. Hasil ini menunjukkan bahwa model mampu menghasilkan soal dengan susunan kata yang memiliki kemiripan tinggi terhadap referensi yang digunakan dalam evaluasi.



Gambar 3. Grafik hasil evaluasi ROUGE terhadap soal hasil generasi

Sumber: Data penelitian, diolah oleh penulis.

Visualisasi pada Gambar 3 memperlihatkan bahwa nilai ROUGE pada komponen soal berada pada nilai maksimum untuk seluruh metrik yang digunakan. Kesamaan nilai ROUGE-1, ROUGE-2, dan ROUGE-L menunjukkan bahwa hasil generasi memiliki kemiripan yang tinggi terhadap teks referensi, baik pada tingkat unigram, bigram, maupun kesamaan urutan teks. Temuan ini memperkuat hasil pada Tabel 4, tetapi interpretasinya tetap perlu mempertimbangkan karakteristik data uji dan bentuk pertanyaan yang digunakan.

Tabel 5. Hasil evaluasi ROUGE untuk jawaban hasil generasi

Metrik Evaluasi	Nilai	Keterangan
ROUGE-1	0,41	41% kata individu dari teks referensi muncul kembali pada teks prediksi
ROUGE-2	0,24	24% pasangan dua kata berurutan pada referensi terdapat pada teks prediksi
ROUGE-L	0,36	36% kesamaan urutan kata/struktur kalimat antara referensi dan prediksi

Berbeda dengan komponen soal, hasil pada Tabel 5 menunjukkan nilai ROUGE yang lebih rendah pada komponen jawaban, yaitu 0,41 untuk ROUGE-1, 0,24 untuk ROUGE-2, dan 0,36 untuk ROUGE-L. Perbedaan nilai tersebut menunjukkan bahwa jawaban hasil generasi memiliki kemiripan tekstual yang lebih rendah terhadap jawaban referensi dibandingkan dengan hasil pembangkitan soal. Perbedaan ini juga menunjukkan bahwa pembangkitan jawaban memiliki variasi susunan kata yang lebih besar, sehingga kesamaan teks antara keluaran model dan referensi tidak setinggi pada komponen soal.

Tabel 6. Contoh perbandingan soal dan jawaban hasil generasi dengan referensi

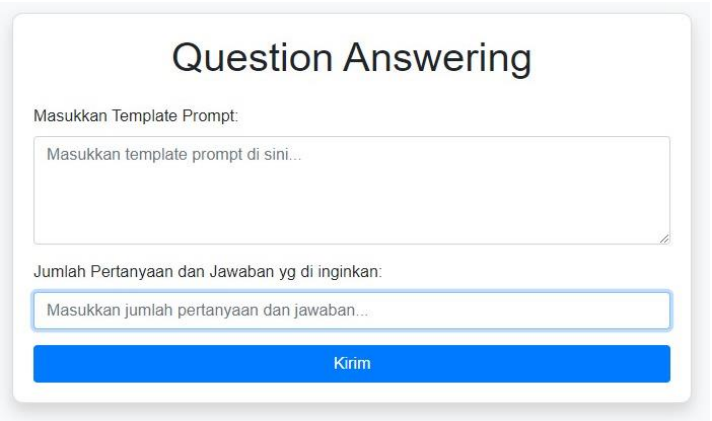
No	Hasil Generate	Referensi	Catatan
1	Jelaskan apa yang dimaksud dengan rabun jauh!	Apa yang dimaksud dengan rabun jauh?	Komponen: Soal. Kesamaan kata kunci penuh ("rabun jauh", "apa yang dimaksud") dan struktur kalimat tanya identik → ROUGE-1 = 1.0, ROUGE-2 = 1.0, ROUGE-L = 1.0. Mengindikasikan pola kalimat tanya yang pendek dan seragam pada data uji, bukan semata kemampuan generatif model.
2	Rabun jauh adalah penglihatan yang sulit melihat benda jauh dengan jelas.	Rabun jauh adalah kondisi di mana seseorang kesulitan melihat benda yang jauh dengan jelas, tetapi bisa melihat benda yang dekat dengan baik.	Komponen: Jawaban. Kesamaan sebagian pada kata kunci inti ("rabun jauh", "melihat benda jauh") namun struktur kalimat dan panjang teks berbeda → ROUGE-1 = 0.409, ROUGE-2 = 0.238, ROUGE-L = 0.364. Menunjukkan jawaban lebih panjang dan bervariasi sehingga kemiripan kata demi kata lebih sulit dicapai dibanding soal.

Sumber: Data penelitian, diolah oleh penulis.

Contoh pada Tabel 6 memperlihatkan perbedaan karakteristik antara hasil pembangkitan soal dan jawaban. Pada contoh pertama, soal hasil generasi menggunakan bentuk "Jelaskan apa yang dimaksud dengan rabun

jauh!", sedangkan referensinya menggunakan bentuk "Apa yang dimaksud dengan rabun jauh?". Meskipun terdapat perbedaan pada susunan kalimat, kedua soal memiliki kata kunci utama yang sama sehingga menghasilkan nilai ROUGE yang tinggi. Pada contoh jawaban, hasil generasi dan referensi juga menyampaikan informasi mengenai rabun jauh, tetapi referensi memiliki penjelasan yang lebih panjang dan susunan kalimat yang berbeda. Kondisi tersebut menghasilkan nilai ROUGE yang lebih rendah dibandingkan komponen soal. Contoh ini menunjukkan bahwa perbedaan formulasi bahasa dapat memengaruhi nilai ROUGE meskipun kedua teks membahas konteks yang sama.

Model hasil *fine-tuning* selanjutnya diintegrasikan ke dalam prototipe aplikasi berbasis web sebagaimana ditunjukkan pada Gambar 4. Prototipe memungkinkan guru memasukkan teks konteks materi pembelajaran, menentukan jumlah soal yang diinginkan, dan menghasilkan soal serta jawaban secara otomatis berdasarkan *prompt* yang diberikan. Sistem juga menyediakan fitur autentikasi pengguna, input teks, peninjauan hasil generasi, penyimpanan ke basis data, serta pengunduhan soal dan jawaban.



Gambar 4. Tampilan antarmuka prototipe sistem pembangkit soal-jawab otomatis
Sumber: Dokumentasi penelitian.

Gambar 4 menunjukkan penerapan model GPT-2 hasil *fine-tuning* ke dalam prototipe sistem berbasis web. Pada alur tersebut, guru dapat memasukkan konteks materi yang akan digunakan sebagai dasar pembangkitan, kemudian sistem memproses input tersebut untuk menghasilkan soal dan jawaban. Hasil pembangkitan selanjutnya dapat ditinjau sebelum disimpan atau diunduh. Integrasi ini menunjukkan bahwa model yang telah melalui proses *fine-tuning* tidak hanya diuji pada tingkat keluaran teks, tetapi juga diimplementasikan dalam alur penggunaan yang mendukung proses pembuatan soal. Namun, prototipe yang ditampilkan pada penelitian ini belum digunakan untuk mengukur tingkat penerimaan pengguna, efektivitas pembelajaran, atau pengurangan beban kerja guru.

4.2 Pembahasan

Hasil eksperimen menunjukkan bahwa peningkatan ukuran dataset diikuti oleh perubahan kualitas keluaran model. Pada dataset berukuran 50 dan 100 data, model masih menghasilkan keluaran yang kurang koheren, sedangkan pada penggunaan 239 data, model menghasilkan pertanyaan yang lebih jelas dan berkaitan dengan konteks yang diberikan. Temuan ini menunjukkan bahwa jumlah data yang digunakan dalam proses *fine-tuning* berperan terhadap kemampuan model dalam mempelajari pola pembangkitan soal dan jawaban pada domain IPAS yang digunakan dalam penelitian. Namun, hubungan tersebut perlu dipahami dalam batas eksperimen yang dilakukan karena penelitian tidak menguji seluruh faktor yang dapat memengaruhi kualitas model. Penurunan *training loss* dan *validation loss* pada seluruh skema pelatihan menunjukkan bahwa model dapat mempelajari data selama proses *fine-tuning*. Akan tetapi, perubahan nilai *loss* tidak secara langsung menunjukkan bahwa kualitas soal dan jawaban yang dihasilkan selalu meningkat. Perbedaan antara tren *loss* dan kualitas keluaran yang diamati pada dataset 50, 100, 150, dan 239 data menunjukkan bahwa evaluasi model generatif tidak cukup dilakukan hanya berdasarkan nilai *loss*. Pemeriksaan terhadap keluaran aktual tetap diperlukan untuk mengetahui koherensi dan kesesuaian hasil generasi dengan konteks. Perbandingan parameter menunjukkan adanya perubahan nilai bobot setelah proses *fine-tuning*, terutama pada parameter transformer.ln_f.weight berdasarkan nilai rata-rata yang dilaporkan pada Tabel 3. Perubahan tersebut menunjukkan bahwa proses *fine-tuning* mengubah parameter model untuk menyesuaikannya dengan dataset yang digunakan. Namun, perubahan rata-rata bobot parameter belum cukup untuk menyimpulkan secara langsung bahwa model mengalami perubahan

representasi tingkat tinggi atau bahwa lapisan akhir memiliki peran paling dominan dalam pembentukan soal. Interpretasi tersebut memerlukan analisis parameter atau representasi model yang lebih komprehensif. Evaluasi menggunakan ROUGE menunjukkan perbedaan antara komponen soal dan jawaban. Nilai ROUGE pada komponen soal dilaporkan mencapai 1,0 untuk ROUGE-1, ROUGE-2, dan ROUGE-L, sedangkan komponen jawaban memperoleh nilai masing-masing 0,41, 0,24, dan 0,36. Perbedaan tersebut menunjukkan bahwa teks soal memiliki tingkat kemiripan yang lebih tinggi terhadap referensi dibandingkan teks jawaban. Salah satu kemungkinan penyebabnya adalah soal pada dataset memiliki pola yang relatif pendek dan seragam, sedangkan jawaban cenderung lebih panjang dan memiliki variasi susunan kalimat yang lebih besar.

Meskipun demikian, skor ROUGE sebesar 1,0 pada seluruh metrik soal perlu ditafsirkan secara hati-hati. Contoh pada Tabel 6 menunjukkan adanya perbedaan antara soal hasil generasi "Jelaskan apa yang dimaksud dengan rabun jauh!" dan referensi "Apa yang dimaksud dengan rabun jauh?". Tingginya skor tersebut dapat berkaitan dengan karakteristik pertanyaan pada data uji yang relatif pendek dan memiliki pola kalimat yang seragam. Pola pertanyaan yang serupa dapat menghasilkan tingkat kemiripan tekstual yang tinggi meskipun terdapat perbedaan pada beberapa kata dalam kalimat. Oleh karena itu, skor ROUGE yang tinggi pada komponen soal tidak serta-merta menunjukkan bahwa model menghasilkan soal yang identik dengan referensi atau memiliki kualitas generatif yang sempurna. Pada komponen jawaban, nilai ROUGE yang lebih rendah menunjukkan bahwa hasil generasi tidak selalu menggunakan susunan kata yang sama dengan referensi meskipun dapat memiliki informasi atau kata kunci yang serupa. Kondisi tersebut merupakan karakteristik yang perlu diperhatikan dalam evaluasi pembangkitan jawaban karena satu konteks dapat menghasilkan beberapa formulasi jawaban yang berbeda tetapi tetap memiliki makna yang sama. Dengan demikian, ROUGE dapat digunakan untuk mengukur kemiripan tekstual, tetapi tidak dapat digunakan sebagai satu-satunya dasar untuk menentukan kualitas atau kebenaran pedagogis jawaban. Integrasi model ke dalam prototipe menunjukkan bahwa model hasil *fine-tuning* dapat diterapkan dalam alur sistem yang menyediakan input konteks, pembangkitan soal dan jawaban, peninjauan hasil, penyimpanan, dan pengunduhan. Prototipe tersebut menunjukkan penerapan hasil eksperimen model ke dalam sistem berbasis web. Namun, keberadaan fitur prototipe tidak secara langsung membuktikan peningkatan efektivitas pembelajaran atau pengurangan beban kerja guru karena aspek tersebut belum dievaluasi menggunakan pengukuran pengguna atau pengujian sebelum dan sesudah penggunaan sistem.

5. Kesimpulan

Berdasarkan hasil penelitian, model GPT-2 yang melalui proses *fine-tuning* dapat digunakan untuk menghasilkan soal dan jawaban berbahasa Indonesia berdasarkan konteks materi pembelajaran IPAS. Hasil eksperimen pada variasi ukuran dataset 50, 100, 150, dan 239 data menunjukkan bahwa peningkatan jumlah data pelatihan memberikan perbedaan pada hasil pembangkitan yang diperoleh. Dataset berukuran 50–100 data menghasilkan keluaran yang masih terbatas, sedangkan penggunaan 239 data memberikan hasil yang lebih baik. Pada dataset 239 data, komponen soal memperoleh skor ROUGE-1, ROUGE-2, dan ROUGE-L sebesar 1,00. Namun, skor tersebut perlu ditafsirkan secara hati-hati karena kemiripan tekstual yang tinggi tidak secara langsung menunjukkan kualitas soal secara keseluruhan. Pada komponen jawaban, skor ROUGE-1 sebesar 0,41, ROUGE-2 sebesar 0,24, dan ROUGE-L sebesar 0,36 menunjukkan bahwa kemiripan tekstual antara jawaban hasil generasi dan referensi masih lebih rendah dibandingkan komponen soal. Perubahan nilai parameter antara model *pre-trained* dan model hasil *fine-tuning* juga menunjukkan adanya penyesuaian pada parameter model, dengan perubahan yang lebih besar pada beberapa lapisan akhir dibandingkan parameter yang diamati pada lapisan awal. Penelitian ini memiliki beberapa keterbatasan. Pertama, dataset yang digunakan hanya terdiri atas 239 triplet konteks-pertanyaan-jawaban sehingga kemampuan model untuk menghasilkan keluaran pada dataset dengan volume dan variasi yang lebih besar belum dapat dipastikan. Kedua, data penelitian bersumber dari satu sekolah dan berfokus pada materi IPAS sekolah dasar, sehingga penerapan model pada sekolah, mata pelajaran, atau jenjang pendidikan lain masih memerlukan pengujian lebih lanjut. Ketiga, evaluasi ROUGE menggunakan referensi yang dihasilkan dengan bantuan satu model pembanding, yaitu ChatGPT, dan belum melibatkan validasi oleh penilai manusia. Oleh karena itu, hasil evaluasi ROUGE perlu dipertimbangkan bersama dengan pemeriksaan kualitatif terhadap relevansi, ketepatan, dan kesesuaian jawaban dengan konteks pembelajaran. Penelitian selanjutnya dapat diarahkan pada penambahan jumlah dan variasi dataset untuk menguji pengaruh volume data terhadap konsistensi hasil pembangkitan, khususnya pada jawaban esai. Pengujian menggunakan data dari sekolah, mata pelajaran, dan jenjang pendidikan yang berbeda juga

diperlukan untuk mengetahui kemampuan generalisasi model. Selain itu, penelitian berikutnya dapat membandingkan GPT-2 dengan model generatif lain, termasuk model yang dikembangkan atau disesuaikan untuk bahasa Indonesia seperti IndoGPT, serta mengeksplorasi teknik augmentasi data dan strategi *decoding* yang berbeda. Evaluasi dengan melibatkan penilai manusia juga dapat digunakan untuk melengkapi metrik ROUGE dalam menilai relevansi, ketepatan, dan kualitas pedagogis soal serta jawaban yang dihasilkan.

Referensi

Bagal, V., Aggarwal, R., Vinod, P. K., & Priyakumar, U. D. (2022). MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling, 62*(9), 2064–2076. https://doi.org/10.1021/acs.jcim.1c00600

Baharuddin, F., & Naufal, M. F. (2023). Fine-tuning IndoBERT for Indonesian exam question classification based on Bloom's Taxonomy. *Journal of Information Systems Engineering and Business Intelligence, 9*(2), 253–263. https://doi.org/10.20473/jisebi.9.2.253-263

Chomphooyod, P., Suchato, A., Tuaycharoen, N., & Punyabukkana, P. (2023). English grammar multiple-choice question generation using text-to-text transfer transformer. *Computers and Education: Artificial Intelligence, 5*, 100158. https://doi.org/10.1016/j.caeai.2023.100158

Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In M. E. Klontzas, S. C. Fanni, & E. Neri (Eds.), *Introduction to artificial intelligence* (pp. 87–99). Springer. https://doi.org/10.1007/978-3-031-25928-9_5

Flores, V. A., Jasa, L., & Hartati, R. S. (2021). Pembangkit pertanyaan otomatis pada materi pelajaran Ilmu Pengetahuan Alam berbahasa Indonesia di tingkat sekolah dasar berdasarkan revisi taksonomi Bloom. *Majalah Ilmiah Teknologi Elektro, 20*(2), 343–350. https://doi.org/10.24843/mite.2021.v20i02.p19

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Naufal, M. F., & Kusuma, S. F. (2023). Otomatisasi pembangkitan pertanyaan untuk bahasa Indonesia (Systematic literature review). *Jurnal Teknologi Informasi dan Ilmu Komputer, 10*(1), 185–192. https://doi.org/10.25126/jtiik.2023106455

Nwafor, C. A., & Onyenwe, I. E. (2021). An automated multiple-choice question generation using natural language processing techniques. *International Journal on Natural Language Computing, 10*(2), 1–10. https://doi.org/10.5121/ijnlc.2021.10201

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Paik, I., & Wang, J.-W. (2021). Improving text-to-code generation with features of code graph on GPT-2. *Electronics, 10*(21), 2706. https://doi.org/10.3390/electronics10212706

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Roshini, R., A, P., C, A., & P, S. (2023). Text generator using GPT-2 model. *International Journal of Scientific Research in Engineering and Management, 7*(2), 1–5. https://doi.org/10.55041/IJSREM17751

Sajun, A. R., Zualkernan, I., & Sankalpa, D. (2024). A historical survey of advances in transformer architectures. *Applied Sciences, 14*(10), 4316. https://doi.org/10.3390/app14104316

Sufi, F. (2024). Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation. *Information, 15*(2), 99. https://doi.org/10.3390/info15020099

Vincentio, K., & Suhartono, D. (2022). Automatic question generation monolingual multilingual pre-trained models using RNN and transformer in low resource Indonesian language. *Informatica, 46*(7). https://doi.org/10.31449/inf.v46i7.4236

Wijaya, D., & Jati, H. (2022). Pengembangan model deep learning untuk pembangkit soal otomatis menggunakan recurrent neural network. *E-JPTI (Jurnal Elektronik Pendidikan Teknik Informatika), 10*(1), 33–39. https://doi.org/10.21831/e-jpti.v10i1.19101

Yang, S. D., Ali, Z. A., & Wong, B. M. (2023). FLUID-GPT (Fast Learning to Understand and Investigate Dynamics with a Generative Pre-Trained Transformer): Efficient predictions of particle trajectories and erosion. *Industrial & Engineering Chemistry Research, 62*(37), 15278–15289. https://doi.org/10.1021/acs.iecr.3c01639

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence, 2*, 100025. https://doi.org/10.1016/j.caeai.2021.100025

How Cites

Agustina, A., & Fuad, M. (2026). Fine-Tuning Model Generative Pre-Trained Transformer 2 (GPT-2) untuk Pembuatan Soal-Jawab Otomatis pada Materi Ilmu Pengetahuan Alam dan Sosial (IPAS) di SD Negeri 1 Wawoangi. *Computer Journal*, *4*(2), 238-248. https://doi.org/10.58477/cj.v4i2.462.

Publisher's Note