Rengganis Jurnal Pengabdian Masyarakat
Volume 6 Nomor 2, November 2026
e-ISSN 2797-1694 | | p-ISSN 2797-1708
https://mathjournal.unram.ac.id/index.php/Rengganis/index

Pemanfaatan *Artificial Intelligence* (AI) dalam penyusunan dan validasi asesmen pembelajaran di MA Muhammadiyah 1 Malang

Mochammad Hafiizh[1], Vita Kusumasari[1], Ety Tejo Dwi Cahyowati[1], Sudirman[1], Fazrina Saumi[2,3], Nourma Pramestie Wulandari[2,4*], Rosmaiyadi[2,5]

[1] *Departemen Matematika, Universitas Negeri Malang, Malang*
[2] *Mahasiswa S3 Pendidikan Matematika, Departemen Matematika, Universitas Negeri Malang, Malang*
[3] *Matematika, Universitas Samudra, Langsa*
[4] *Pendidikan Matematika, Universitas Mataram, Mataram*
[5] *PGSD, Institut Sains dan Bisnis Internasional Singkawang, Singkawang*

nourma.pramestie.2503119@students.um.ac.id; nourmapw@unram.ac.id

Abstract
This community service program assisted teachers in using generative artificial intelligence (AI) to develop and review learning assessments at MA Muhammadiyah 1 Malang. Twenty-two teachers from different subjects participated in a face-to-face workshop and online mentoring from 24 April to 1 May 2026. Participants practised developing assessment items with ChatGPT, designing visual presentations using Canva AI, checking AI-generated outputs for alignment with learning objectives and cognitive demands, and revising their assessment products. Evaluation included 20-item pre- and post-tests, a product assessment rubric, and a 20-item participant response questionnaire. The mean test score increased descriptively from 15.95 to 18.41 out of 20. The mean product score was 86.08, with 11 products rated Good and 11 rated Very Good, while the questionnaire score reached 79.27% of the maximum possible score. Based on these evaluations, the mentoring program achieved its intended objectives during the training period. However, some assessment products still require greater cognitive depth in higher-order thinking tasks. These findings reflect outcomes observed during the program and do not establish the empirical validity of the assessments in classroom use or the long-term sustainability of teachers' skills.

Keywords: Artificial Intelligence; learning assessment; ChatGPT; teacher training; assessment validation

Abstrak
Kegiatan pengabdian ini mendampingi guru dalam memanfaatkan kecerdasan buatan generatif untuk menyusun dan menelaah asesmen pembelajaran di MA Muhammadiyah 1 Malang. Sebanyak 22 guru lintas mata pelajaran mengikuti workshop tatap muka dan pendampingan daring pada 24 April–1 Mei 2026. Peserta berlatih menyusun butir dengan ChatGPT, mengembangkan tampilan melalui Canva AI, memeriksa kesesuaian keluaran AI terhadap tujuan pembelajaran dan tuntutan kognitif, serta merevisi produk asesmen. Evaluasi menggunakan pre-test dan post-test masing-masing 20 butir, rubrik penilaian produk, serta angket respons 20 pernyataan. Rata-rata skor tes secara deskriptif meningkat dari 15,95 menjadi 18,41 dari skor maksimum 20. Rata-rata penilaian produk mencapai 86,08 dengan 11 produk berkategori Baik dan 11 berkategori Sangat Baik, sedangkan capaian angket sebesar 79,27% dari skor maksimum. Berdasarkan hasil evaluasi yang dilakukan, dapat disimpulkan bahwa kegiatan pendampingan penyusunan dan validasi asesmen pembelajaran di MA Muhammadiyah 1 Malang tercapai dengan baik. Namun, sebagian produk asesmen masih memerlukan pendalaman tuntutan berpikir tingkat tinggi. Temuan ini menggambarkan capaian selama pelatihan dan belum membuktikan validitas empiris produk dalam pembelajaran maupun keberlanjutan keterampilan guru.

Kata Kunci: *Artificial Intelligence*; asesmen pembelajaran; ChatGPT; pelatihan guru; validasi asesmen

1. PENDAHULUAN

Kecerdasan buatan atau *artificial intelligence* (AI), khususnya *generative* AI, semakin banyak digunakan untuk membantu penyusunan butir soal, penyajian stimulus, dan pemberian umpan balik dalam pendidikan; perkembangan tersebut membuka peluang efisiensi penyusunan asesmen sekaligus menuntut pengawasan terhadap kualitas keluaran teknologi (González-Calatayud et al., 2021; Kamalov et al., 2023; Owan et al., 2023). Penggunaan aplikasi yang mudah diakses, seperti ChatGPT dan Canva AI, memungkinkan guru mengembangkan draf asesmen dan tampilan soal melalui serangkaian instruksi atau prompt, tetapi kecepatan produksi tidak dengan sendirinya menunjukkan kecocokan instrumen dengan tujuan pembelajaran.

Masalah penting dalam pemanfaatan AI untuk asesmen terletak pada keandalan isi dan ketepatan keputusan pedagogis karena jawaban yang terlihat meyakinkan tetap dapat mengandung kekeliruan konsep, pilihan jawaban yang kurang tepat, atau tuntutan kognitif yang berbeda dari indikator yang diharapkan (Yan et al., 2024; Gardner et al., 2021). Telaah terhadap kesesuaian indikator, bahasa, tingkat kognitif, kunci jawaban, dan rubrik tetap menjadi tanggung jawab guru sebagai pengambil keputusan profesional; kebutuhan terhadap pemeriksaan tersebut juga dibahas dalam kajian validitas asesmen berbantuan generative AI (Kaldaras et al., 2024).

Berdasarkan observasi, wawancara, dan kuesioner kebutuhan yang dilaporkan tim pengabdian di MA Muhammadiyah 1 Malang, guru telah mengenal sejumlah platform digital dan menggunakan AI terutama untuk administrasi pembelajaran serta penyusunan soal sederhana; diskusi awal juga mengungkap kebutuhan akan contoh penyusunan prompt, penelaahan kesesuaian butir dengan indikator, pemeriksaan ketepatan materi, dan pengembangan soal yang menuntut higher-order thinking skills (HOTS). Informasi kebutuhan tersebut bersifat deskriptif sebagai dasar perancangan kegiatan, sedangkan jumlah informan pada setiap teknik dan frekuensi masing-masing respons tidak tersedia dalam rekap yang digunakan untuk artikel ini. Kegiatan dirancang dengan konteks pendidikan matematika sebagai salah satu contoh utama dan dilaksanakan kepada 22 guru lintas mata pelajaran, sehingga produk yang dinilai berasal dari bidang studi peserta masing-masing.

Pendampingan diarahkan pada penggunaan ChatGPT untuk menghasilkan dan meninjau draf asesmen serta Canva AI untuk mendukung penyajian visual, dengan penekanan pada pemeriksaan hasil, revisi, dan refleksi guru atas kualitas butir. Kegiatan ini bertujuan mendeskripsikan keterlaksanaan workshop dan pendampingan, perubahan skor pemahaman yang teramati selama program, mutu produk berdasarkan rubrik kegiatan, serta respons peserta. Batasan dalam cakupan kegiatan ini ialah penelaahan pedagogis dalam pelatihan dan belum mencakup pengujian validitas empiris instrumen pada peserta didik. Fokus tersebut selaras dengan pentingnya literasi AI guru dan pengawasan manusia dalam pemanfaatan teknologi untuk asesmen (Celik et al., 2022; Perkins et al., 2024).

2. METODE PELAKSANAAN

Kegiatan pengabdian dilaksanakan di MA Muhammadiyah 1 Malang dengan melibatkan 22 guru dari beragam mata pelajaran selama 24 April–1 Mei 2026, melalui workshop tatap muka dan pendampingan daring. Contoh yang diberikan menonjolkan asesmen matematika, sedangkan praktik dan penilaian produk menyesuaikan dengan mata pelajaran setiap peserta. Data kebutuhan sebelum pelatihan dikumpulkan melalui koordinasi dengan sekolah, observasi, wawancara, dan kuesioner kepada guru serta siswa. Data tersebut dimanfaatkan secara deskriptif untuk menentukan materi pelatihan dan bentuk pendampingan.

Persiapan mencakup koordinasi pada 4 Februari 2026, analisis kebutuhan pada 6 April 2026, penyusunan bahan serta instrumen pada 9–17 April 2026, dan pengecekan perangkat pada 22 April 2026. Workshop pada 24 April 2026 memuat prinsip asesmen, HOTS, literasi dan etika AI, risiko kekeliruan keluaran *generative* AI, serta praktik penyusunan prompt, butir, stimulus, dan tampilan soal. Validasi pedagogis dan revisi draf dibahas secara daring melalui Google Meet pada 26–28 April 2026, kemudian dilanjutkan dengan konsultasi dan refleksi pada 29–30 April 2026; evaluasi akhir dilaksanakan secara tatap muka pada 1 Mei 2026.

Dalam praktiknya, guru menyusun draf asesmen dengan bantuan AI lalu menelaah kesesuaiannya terhadap tujuan dan indikator pembelajaran, ketepatan konsep atau kunci jawaban, tingkat kognitif, kejelasan bahasa, serta rubrik penilaian. Kemudian, umpan balik pendamping dan diskusi dengan peserta digunakan sebagai bahan revisi. Proses tersebut dilaporkan sebagai penelaahan atau validasi pedagogis terbimbing dan belum menjadi suatu uji validitas empiris dengan analisis butir pada siswa. Canva AI dimanfaatkan untuk rancangan visual dan kuis interaktif. Oleh karena itu, hasil artikel dibatasi pada penggunaan ChatGPT dan Canva AI.

Evaluasi kuantitatif menggunakan pre-test dan post-test yang masing-masing memuat 20 butir, rubrik analitik untuk menilai produk asesmen peserta, serta angket respons berisi 20 pernyataan berskala Likert 1–5. Rubrik memuat kesesuaian pedagogis, kualitas desain, pemanfaatan AI berbasis bukti, serta validasi dan akurasi materi. Sementara kategori mutu produk mengikuti klasifikasi pada rekap penilaian kegiatan. Penilaian dilakukan secara deskriptif melalui rata-rata, rentang dan distribusi skor, kategori produk, serta skor dan persentase respons.

Pre-test dan post-test masing-masing memiliki 22 respons lengkap. Hasil tes ini kemudian dibandingkan secara deskriptif pada tingkat kelompok, sementara hasil produk dan respons peserta dibaca bersama catatan refleksi untuk mengidentifikasi kekuatan dan kebutuhan pendampingan lanjutan. Rincian kronologi disajikan pada Tabel 1 berikut.

Tabel 1. Kronologi dan fokus kegiatan pengabdian

Tahap	Kegiatan
Persiapan	4 Februari–22 April 2026: koordinasi, analisis kebutuhan, penyusunan bahan serta instrumen dan pengecekan perangkat
Workshop tatap muka	24 April 2026: Pre-test, konsep asesmen, HOTS, literasi AI, praktik ChatGPT dan Canva AI.
Validasi daring	26–28 April 2026: penelaahan draf, diskusi dan revisi pedagogis melalui Google Meet
Refleksi daring	29–30 April 2026: konsultasi kelompok dan individu serta umpan balik produk
Evaluasi akhir	1 Mei 2026: post-test, penilaian produk dan angket respons

3. HASIL DAN PEMBAHASAN

3.1 Hasil

3.1.1 Keterlaksanaan Kegiatan

Persiapan kegiatan menghasilkan rancangan workshop berbasis kebutuhan yang teridentifikasi melalui koordinasi, diskusi awal, dan kuesioner. Guru telah menggunakan teknologi digital untuk administrasi dan pembuatan soal, tetapi membutuhkan panduan yang lebih terarah dalam menelaah keluaran AI terhadap indikator, level kognitif, kejelasan redaksi, dan ketepatan materi. Dokumentasi persiapan serta penyusunan bahan workshop dan pendampingan disajikan pada Gambar 1 dan Gambar 2.



Gambar 1. Diskusi awal bersama sekolah mitra untuk mengidentifikasi kebutuhan pelatihan

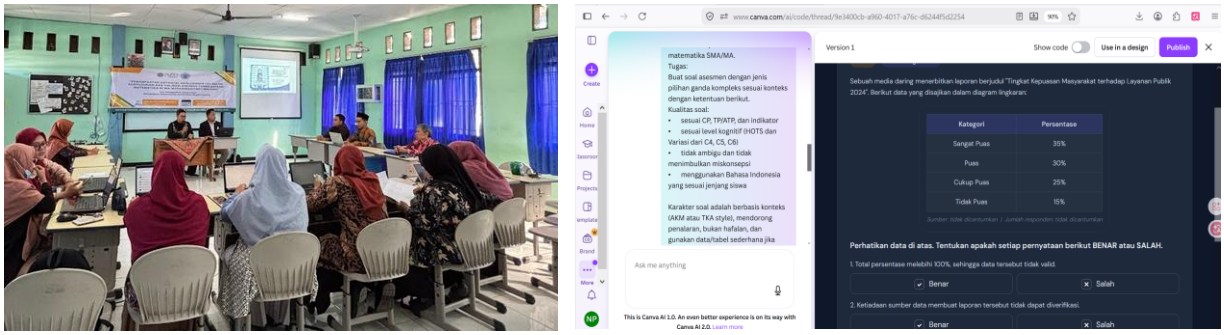


Gambar 2. Dokumentasi penyusunan *handout* dan instrumen pelatihan

Pada workshop 24 April 2026 peserta memperoleh penguatan prinsip asesmen dan berlatih menyusun draf butir serta stimulus menggunakan ChatGPT, kemudian mengembangkan penyajian visual melalui Canva AI. Selama pendampingan daring pada 26–30 April 2026 peserta menelaah kesesuaian indikator, tuntutan kognitif, akurasi materi, dan kejelasan bahasa dari hasil AI lalu mendiskusikan revisinya dengan tim pendamping. Kemudian, pada hasil refleksi disebutkan bahwa ada contoh masalah berupa distraktor yang kurang kuat, instruksi ambigu, serta soal yang masih prosedural. Dokumentasi pada Gambar 3 dan Gambar 4 menunjukkan kegiatan pemaparan dan praktik.



Gambar 3. Dokumentasi Kegiatan Penyampaian Materi Konsep Asesmen dan AI



Gambar 4. Dokumentasi Praktik Penyusunan Asesmen Menggunakan Canva AI dan ChatGPT

Evaluasi pada 1 Mei 2026 mencakup post-test, telaah produk peserta, dan pengisian angket respons. Penilaian produk menggambarkan hasil pekerjaan yang disusun selama kegiatan, sedangkan respons peserta menginformasikan persepsi mereka terhadap materi, pendampingan, dan pelaksanaan. Catatan terbuka memuat permintaan penambahan waktu praktik, contoh prompt yang lebih konkret, dan pendampingan lanjutan. Foto kegiatan akhir tercantum pada Gambar 5 berikut.



Gambar 5. Evaluasi akhir dan dokumentasi penutupan rangkaian pendampingan

3.1.2 Hasil Evaluasi Kegiatan

Data pre-test dan post-test masing-masing berasal dari 22 respons lengkap dengan 20 butir soal. Rata-rata pre-test mencapai 15,95 dengan rentang 15–17 dan standar deviasi 0,72, sedangkan rata-rata post-test mencapai 18,41 dengan rentang 17–20 dan standar deviasi 1,01. Selisih rata-rata sebesar 2,46 poin menggambarkan perbedaan capaian kelompok pada dua waktu pengukuran. Rincian ringkas disajikan pada Tabel 2.

Tabel 2. Ringkasan capaian evaluasi selama pelatihan

Aspek Evaluasi	Hasil
Tes (22 respons per waktu)	Pre-test 15,95 (15–17); post-test 18,41 (17–20), masing-masing 20 soal
Penilaian produk (22)	Rerata 86,08; 11 Baik; 11 Sangat Baik
Angket (22 peserta)	Skor 1.744 dari 2.200 (79,27%); rerata 3,96 dari 5

Persentase jawaban benar pada sejumlah materi terpilih disajikan pada Tabel 3 berdasarkan rekap per butir. Pada pre-test, persentase relatif rendah ditemukan pada pemanfaatan ChatGPT (59,09%) dan pentingnya validasi soal (68,18%), sedangkan pada post-test butir validasi hasil AI dan tahapan penggunaan AI masing-masing mencapai 95,45% serta butir soal HOTS mencapai 86,36%. Perbedaan redaksi dan isi butir pada kedua instrumen membatasi penggunaan tabel ini sebagai bukti peningkatan indikator yang setara karena angka-angka tersebut diperlakukan sebagai gambaran hasil pada dua tes, bukan sebagai analisis perubahan setiap butir.

Tabel 3. Persentase jawaban benar pada indikator terpilih dalam pre-test dan post-test

Aspek yang berdekatan	Butir pada pre-test (%)	Butir pada post-test (%)
Pemanfaatan ChatGPT	59,09 (pemanfaatan ChatGPT)	90,91 (manfaat ChatGPT dalam asesmen)
Validasi soal/hasil AI	68,18 (pentingnya validasi soal)	95,45 (validasi hasil AI)
Fungsi rubrik	77,27 (fungsi rubrik)	95,45 (fungsi rubrik penilaian)
HOTS	77,27 (HOTS)	86,36 (soal HOTS)
Tahapan penggunaan AI	81,82 (tahapan penggunaan AI)	95,45 (tahapan penggunaan AI)
Objektivitas	90,91 (objektivitas asesmen)	100,00 (objektivitas asesmen)

Penilaian proyek memperlihatkan rerata 86,08, skor minimum 80,63, serta skor maksimum 91,69 dari 22 produk. Sebanyak 11 produk berada dalam kategori Baik dan 11 produk dalam kategori Sangat Baik. Hasil telaah menurut rubrik menunjukkan

kategori Baik untuk kesesuaian pedagogis dan desain proyek, serta kategori Sangat Baik untuk pemanfaatan AI berbasis bukti dan validasi atau akurasi materi. Lebih lanjut, sejumlah produk masih menghadapi keterbatasan pada kualitas distraktor, kejelasan instruksi, dan kedalaman tuntutan kognitif. Hasil tersebut menunjukkan mutu menurut rubrik internal kegiatan dan belum merupakan bukti validitas empiris soal ketika diberikan kepada siswa.

Tabel 4. Distribusi kategori dan ringkasan aspek penilaian produk

Komponen	Temuan pada laporan
Jumlah produk dan kategori	22 produk; 11 Baik; 11 Sangat Baik
Rentang dan rerata	80,63–91,69; rerata 86,08
Kesesuaian pedagogis	24,51 — Baik
Kualitas desain proyek	20,53 — Baik
Pemanfaatan AI berbasis bukti	22,62 — Sangat Baik
Validasi dan akurasi materi	18,08 — Sangat Baik

Angket respons diisi oleh 22 guru dengan 20 pernyataan menggunakan skala 1–5. Skor aktual mencapai 1.744 dari maksimum 2.200 atau 79,27% dengan rerata 3,96 dan kategori Baik. Berdasarkan rekap aspek, nilai rerata tertinggi terdapat pada pemanfaatan AI (4,09), sedangkan terendah pada penyampaian narasumber (3,67). Persentase respons positif per aspek merupakan ukuran proporsi jawaban setuju atau sangat setuju yang berbeda dari persentase capaian skor keseluruhan, sehingga kedua ukuran tersebut tidak ditafsirkan sebagai indikator yang sama.

Tabel 5. Angket respons peserta berdasarkan aspek

Aspek	Respons positif (%)	Rerata / 5
Materi pelatihan	72,70	4,08
Penyampaian narasumber	60,60	3,67
Pemanfaatan AI (ChatGPT)	73,90	4,09
Manfaat kegiatan	64,80	3,85
Pelaksanaan kegiatan	71,80	4,04

Berdasarkan jawaban terbuka yang diberikan oleh peserta, terlihat bahwa hasilnya menunjukkan kebutuhan peserta terhadap waktu praktik yang lebih memadai, penjelasan lebih bertahap, dan contoh penggunaan AI yang sesuai dengan mata pelajaran. Temuan respons tersebut memberi masukan bagi perbaikan program, tanpa membuktikan perubahan kemampuan guru dalam praktik kelas setelah kegiatan berakhir.

3.2 Pembahasan

Perbandingan deskriptif pre-test dan post-test menunjukkan bahwa peserta mencapai skor rata-rata yang lebih tinggi setelah workshop dan pendampingan, dengan perubahan dari 15,95 menjadi 18,41 pada instrumen 20 butir; temuan ini relevan

dengan kebutuhan guru untuk menelaah hasil AI secara pedagogis, tetapi interpretasinya perlu memperhitungkan skor awal yang telah relatif tinggi serta ketiadaan data berpasangan yang dapat diaudit. Penelitian Celik et al. (2022) menempatkan pengetahuan pedagogis dan kesiapan profesional guru sebagai unsur penting pemanfaatan AI, sedangkan pengamatan dalam kegiatan ini menunjukkan adanya ruang penguatan praktik penelaahan soal meskipun peserta telah mengenal teknologi digital sebelum pelatihan.

Perbandingan jawaban benar per butir memberi informasi yang lebih terperinci daripada skor total, termasuk lemahnya capaian awal pada pemanfaatan ChatGPT dan pemahaman tentang pentingnya validasi soal serta capaian pascapelatihan yang relatif tinggi pada validasi hasil AI dan penggunaan rubrik. Karena daftar indikator dan rumusan butir pada pre-test serta post-test tidak seluruhnya sama, perubahan dalam tabel perlu dibaca sebagai dua profil hasil tes yang terpisah dan tidak layak digunakan sebagai pengukuran langsung kenaikan setiap kompetensi. Selain itu, risiko kesalahan keluaran generative AI menuntut pemeriksaan manusia terhadap ketepatan konsep dan bahasa sebelum asesmen digunakan (Yan et al., 2024; Kaldaras et al., 2024).

Pada produk asesmen, skor rerata 86,08 serta distribusi 11 produk Baik dan 11 Sangat Baik menunjukkan bahwa peserta dapat menyelesaikan tugas sesuai rubrik kegiatan. Penilaian aspek mencatat kategori Sangat Baik pada pemanfaatan AI berbasis bukti dan validasi atau akurasi materi, sementara kesesuaian pedagogis serta kualitas desain memperoleh kategori Baik. Kombinasi hasil ini mengisyaratkan bahwa dokumentasi *prompt* dan revisi telah lebih kuat daripada kedalaman sebagian rancangan butir. Temuan tersebut sejalan dengan hasil studi Martin Kowal et al. (2025) tentang pentingnya telaah ahli terhadap butir hasil AI.

Catatan evaluasi proyek memperlihatkan bahwa kelemahan yang muncul mencakup distraktor yang kurang kuat, instruksi ambigu, serta soal yang terlalu prosedural. Oleh karena itu, hasil akhir dengan tampilan yang baik tidak dengan sendirinya menandakan kesesuaian level kognitif yang ditargetkan. Persentase jawaban benar 86,36% pada butir HOTS post-test menggambarkan capaian pengetahuan pada satu butir, sedangkan hasil produk tetap menuntut pendalaman, terutama untuk tugas yang melibatkan penalaran dan evaluasi. Klasifikasi C3 sebagai penerapan dan C4 sebagai analisis juga perlu dipisahkan dalam menafsirkan HOTS, sebab pengelompokan keduanya sebagai satu level tidak menggambarkan tuntutan berpikir yang identik.

Prosedur penelaahan pedagogis selama pendampingan menempatkan guru sebagai pihak yang memeriksa dan memutuskan revisi, dengan AI yang dimanfaatkan untuk menghasilkan rancangan awal dan memberikan alternatif perbaikan. Pelaksanaan diskusi serta konsultasi memberi kesempatan kepada peserta untuk meninjau kesesuaian indikator, redaksi, dan ketepatan materi. Temuan mengenai validasi dibatasi pada praktik pemeriksaan dan revisi terbimbing. Penekanan peran manusia

dalam pengambilan keputusan suatu asesmen ini selaras dengan Perkins et al. (2024) dan Gardner et al. (2021).

Respons peserta memberikan gambaran yang lebih bernuansa daripada sekadar kategori Baik karena rerata pemanfaatan ChatGPT (4,09) lebih tinggi daripada aspek penyampaian narasumber (3,67), sedangkan jawaban terbuka menekankan kebutuhan tambahan waktu dan pendampingan. Hal ini menunjukkan bahwa apresiasi peserta terhadap topik AI dapat berjalan bersamaan dengan kebutuhan peningkatan kualitas pelaksanaan dan eksplorasi asesmen. Salinas-Navarro et al. (2024) menyoroti pentingnya pengalaman autentik dalam integrasi generative AI, sedangkan Shi (2025) memandang evaluasi kritis keluaran AI sebagai bagian kompetensi guru yang perlu dikembangkan lebih lanjut.

Kegiatan ini memiliki keterbatasan pada cakupan satu sekolah, periode pelatihan yang singkat, dan belum adanya penerapan produk asesmen dalam kelas. Oleh sebab itu, hasil yang telah disajikan terutama menjelaskan keterlaksanaan program, capaian tes, karakteristik produk pelatihan, dan respons guru, bukan efektivitas jangka panjang atau dampak langsung pada hasil belajar siswa.

4. SIMPULAN

Pendampingan pemanfaatan ChatGPT dan Canva AI di MA Muhammadiyah 1 Malang terlaksana melalui workshop, penelaahan produk, dan refleksi bersama 22 guru lintas mata pelajaran. Evaluasi selama kegiatan menunjukkan rerata tes kelompok yang lebih tinggi pada post-test serta penyelesaian seluruh produk asesmen dalam kategori Baik dan Sangat Baik menurut rubrik internal program. Kontribusi yang dapat diamati terletak pada kesempatan guru mempraktikkan penyusunan dan penelaahan keluaran AI, sementara mutu empiris instrumen di kelas dan keberlanjutan keterampilan setelah pelatihan belum diuji.

Keterbatasan yang masih menonjol mencakup kebutuhan pendalaman soal HOTS dan keterbatasan waktu pendampingan. Temuan tersebut menempatkan kegiatan sebagai tahap penguatan awal yang memerlukan evaluasi lanjutan sebelum disimpulkan berhasil meningkatkan mutu asesmen dalam penggunaan nyata.

5. UCAPAN TERIMA KASIH

Tim pengabdian mengucapkan terima kasih kepada MA Muhammadiyah 1 Malang atas dukungan tempat dan fasilitas serta kepada seluruh guru peserta atas partisipasi dalam workshop, penelaahan produk, dan evaluasi kegiatan.

6. REKOMENDASI

Program berikutnya disarankan menyediakan lebih banyak waktu untuk praktik penyusunan soal HOTS, pemeriksaan kunci atau rubrik, dan dokumentasi perubahan butir dari keluaran awal AI sampai hasil revisi yang ditelaah ulang. Tim juga perlu

menetapkan prosedur penilaian produk secara lebih rinci dan memeriksa keterterapan asesmen melalui uji coba terbatas di kelas sesuai kewenangan dan kebutuhan sekolah.

Pendampingan lanjutan dapat diarahkan pada pengembangan stimulus multimodal yang relevan dengan mata pelajaran, disertai evaluasi berkelanjutan untuk mengetahui penggunaan produk dan perubahan praktik asesmen guru setelah kegiatan selesai.

7. REFERENSI

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*, 616–630. https://doi.org/10.1007/s11528-022-00715-y

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning, 37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences, 11*(12), Article 5467. https://doi.org/10.3390/app11125467

Kaldaras, L., Akaeze, H., & Reckase, M. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education, 9*, Article 1399377. https://doi.org/10.3389/feduc.2024.1399377

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability, 15*(16), Article 12451. https://doi.org/10.3390/su151612451

Martin Kowal, J., Hurley Bryant, K., Segall, D., & Kantrowitz, T. (2025). Harnessing generative AI for assessment item development: Comparing AI-generated and human-authored items. *International Journal of Selection and Assessment, 33*(3), Article e70021. https://doi.org/10.1111/ijsa.70021

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), Article em2307. https://doi.org/10.29333/ejmste/13428

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The artificial intelligence assessment scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice, 21*(6). https://doi.org/10.53761/q3azde36

Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Using generative artificial intelligence tools to explain and enhance experiential learning for authentic assessment. *Education Sciences, 14*(1), Article 83. https://doi.org/10.3390/educsci14010083

Shi, L. (2025). Assessing teachers' generative artificial intelligence competencies: Instrument development and validation. *Education and Information Technologies, 30*, 23365–23384. https://doi.org/10.1007/s10639-025-13684-5

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*, 1839–1850. https://doi.org/10.1038/s41562-024-02004-5