MEDICAL IMAGE ALIGNMENT ASSESSMENT AS A TEST OF GENERALIST VISUAL REASONING IN FRONTIER MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier multimodal large language models (MLLMs) are increasingly positioned as general purpose visual reasoners as part of the quest for artificial general intelligence. A key test of this generality is whether they can perform novel visual judgments that humans can make reliably from visual evidence and task instructions, without task-specific parameter optimisation. We investigate this question through the task of medical image alignment assessment, where the goal is to establish whether there is anatomical correspondence between two images. Assessing spatial alignment is a fundamental component of medical imaging pipelines across within-subject, between-subject, and cross-scanner settings. Human visual assessment of image alignment is still the gold standard and most common approach; however, it requires trained operators and is impractical to scale for large datasets. Existing automated methods using conventional image quality metrics or convolutional neural networks (CNNs) are generally designed for particular modalities, anatomical regions, or alignment tasks, not generalisable readily to new settings. We evaluate recent generations of MLLMs on two exemplar medical image alignment tasks, varying both prompting strategies and image-presentation methods. We compare against a smaller, locally fine-tuned MLLM and a task-specific CNN, allowing us to examine the trade-off between frontier general purpose models that cannot be used locally and smaller models that require specific task optimisation but can be used locally. Our results show that we are reaching an inflection point, where frontier MLLMs can now effectively perform visual assessment of medical image alignment. Models released only a few months ago generalise poorly and, in some settings, perform barely above chance, whereas GPT-6 achieves over 85% across almost all scenarios tested, including under zero-shot evaluation. Fine-tuned local models can match or exceed frontier-model performance on the tasks on which they are trained, but transfer substantially less effectively to unseen settings. These findings identify medical image alignment as a useful test bed for generalist visual reasoning. They also suggest that frontier multimodal models are beginning to acquire transferable visual assessment capabilities that could support a common quality-control mechanism across heterogeneous medical-imaging pipelines.

1 INTRODUCTION

Generalist visual reasoning is a key aspect of the push for artificial general intelligence in multimodal large language models (MLLMs). More precisely, this means that MLLMs should be able to perform novel visual tasks that humans can do reliably from visual evidence and task instructions, without task-specific parameter optimisation. We investigate whether this is the case for frontier MLLMs through the task of medical image alignment assessment.

Medical imaging is critically important in modern healthcare, helping to guide diagnostic, prognostic and treatment decisions. Establishment of spatial alignment between medical images is a crucial part of many image analysis pipelines, ensuring anatomical correspondence between images across within-subject, between-subject and cross-scanner settings. This correspondence is key to almost all aspects of medical image analysis; for instance, we need to check raw data for motion between

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

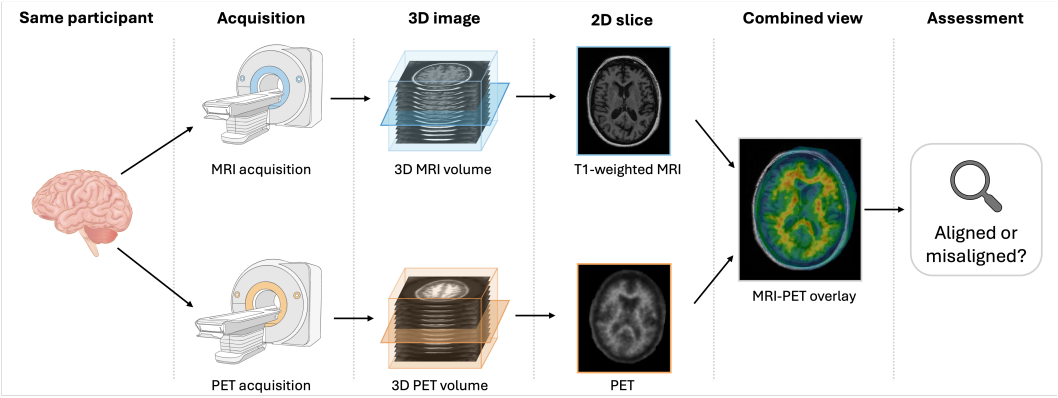

Figure 1: Illustration of the medical image alignment assessment problem.

volumes (Andersson & Sotiropoulos, 2016; Legouhy et al., 2026a; Friston et al., 1996; Picard & Thompson, 1997; Ye et al., 2014) and ensure that images are aligned for multimodal fusion (Greve et al., 2014; 2016; Doré et al., 2019; Kim et al., 2024) and longitudinal analysis (Reuter et al., 2012; Fox & Freeborough, 1997; Bosc et al., 2003).

Since image alignment is a cornerstone of so many image analysis pipelines, it is critical to assess alignment and ensure it is sufficient to ensure the accuracy of downstream processing. Indeed, Dadar et al. (2018) found typical failure rates of around 5-10% when investigating methods for alignment of T1-weighted (T1w) MRI, a type of MRI that is used to visualise anatomical details, to a template space. While image alignment assessment is essential to ensure that any alignment errors do not propagate to downstream analyses, it is inherently difficult, particularly in the multimodal setting (Rohlfing, 2012; Bierbrier et al., 2022).

In practice, the gold standard for image alignment assessment is manual inspection. Typically this will be achieved by overlaying the two images using an image viewing tool to visually assess the accuracy or plausibility of the alignment. This can include directly overlaying the images with translucency and overlaying anatomical cues, such as segmentation outlines from one image onto another. Figure 1 shows an example of this process. Visual assessment, however, requires trained assessors and is time consuming. This makes it impractical for large datasets.

Efforts have been made to automate the image alignment assessment process, attempting to overcome many of the limitations of manual visual assessment; however, automating this process is complex. Some approaches use traditional image quality or segmentation overlap metrics (Rohlfing, 2012; Bosma et al., 2023; Esteban et al., 2017), while deep-learning approaches have recently been proposed for image alignment assessment for quality control of registration (Denis De Senneville et al., 2020; Fonov et al., 2022; Lallement et al., 2023; Salari et al., 2023), a task aiming to place a pair of images into alignment. These methods, however, are generally designed for particular modalities, anatomical regions, or registration tasks and do not generalise readily to new settings.

The task of MRI-PET alignment exemplifies the challenges inherent in developing automated image alignment assessment approaches. MRI and PET depict different physical and biological properties, and therefore lack a simple relationship between corresponding voxel intensities. Moreover, this relationship may differ across radioactive chemical tracers used in a PET exam and disease states, for instance different contrast patterns of amyloid PET (a type of PET exams using tracers that preferentially bind to amyloid plaques, the hallmark of Alzheimer's) between Alzheimer's disease patients and controls. Segmentation or landmark based approaches are equally limited in this context because the comparatively low spatial resolution and tracer-dependent contrast of PET make these features difficult to obtain reliably in both modalities. Importantly, these limitations do not prevent visual assessment of MRI-PET alignment. Human readers can often fairly easily detect anatomical misalignment by checking the correspondence of anatomical cues derived from the structural MRI with the PET image, such as in Figure 1.

Recently, multimodal large-language models (MLLMs) have demonstrated increasingly strong generalised image understanding (Yue et al., 2023; 2024; Gemini Team et al., 2023; OpenAI et al., 2024), including one recent study utilising MLLMs for visual quality control in a wider agentic

framework (Han et al., 2026). We test how well this generalist visual reasoning can be used to replicate a human-like image alignment assessment process in two exemplar medical imaging settings as outlined in Figure 2. We further examine the tradeoff between frontier API-based general purpose models and local models that require specific task optimisation by comparing frontier MLLMs to a smaller, locally fine-tuned MLLM and a task-specific CNN.

2 EXPERIMENTAL DESIGN

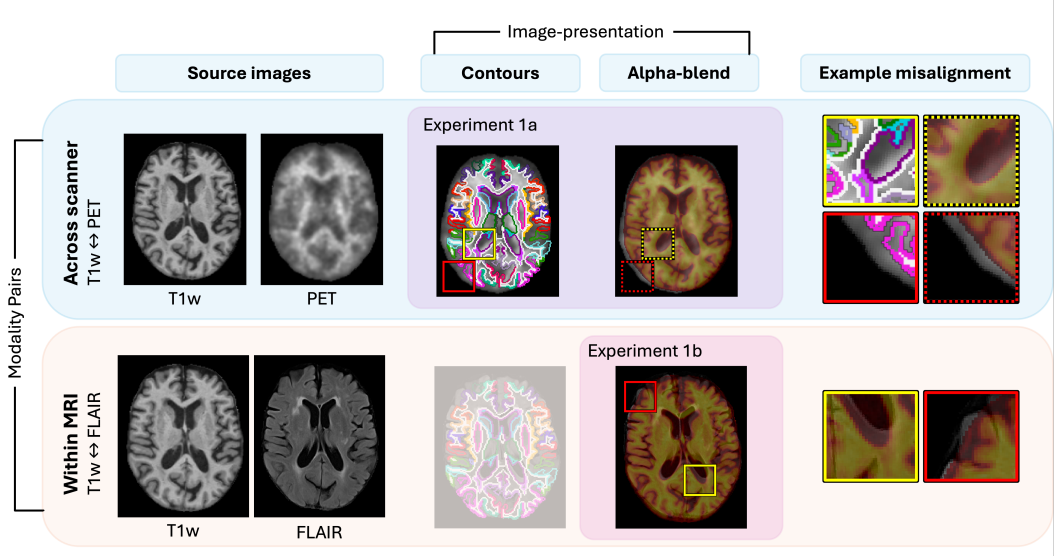

Figure 2: Overview of the medical image alignment assessment method. We test image alignment assessment on two modality pairs: cross-scanner T1w MRI-PET and within MRI T1w-FLAIR alignment. We highlight how different modality pairs and image-presentation techniques can affect the alignment assessment task. Yellow zoom boxes show misalignment of ventricles, something which is easy to spot for MRI-PET contours and T1w-FLAIR alpha-blend (solid yellow boxes) but harder to see in MRI-PET alpha-blend (dashed yellow box). Red zoom boxes show brain-edge misalignment which is easy to spot in MRI-PET alpha-blend (dashed red box) but harder to definitively see in MRI-PET contours (solid red box).

To test the generalist visual reasoning of frontier MLLMs, our first experiment involves testing a range of models in two medical image alignment settings to establish how well they can perform this task. We then investigate whether frontier MLLMs can now outperform task-specific locally trained models in our second experiment:

Experiment 1 Testing MLLMs as medical image alignment assessors

- **Experiment 1a: MRI-PET alignment** a challenging setting where although the two images come from the same subject, they convey different information. We use this as a test bed to probe the impact of various prompting/input choices on model performance.
- **Experiment 1b: T1w-FLAIR (fluid attenuated inversion recovery) MRI alignment** a theoretically simpler task where the two images both show aspects of brain structure. We test how well best performing settings from the MRI-PET experiment transfer to this dataset.

Experiment 2 Comparison to local models

We test a smaller locally fine-tuned MLLM and a task-specific CNN to establish whether, and in which scenarios, API-based frontier MLLMs outperform local task-specific models.

2.1 EXPERIMENT 1A: MRI-PET ALIGNMENT

To benchmark the performance of MLLMs for image alignment assessment, we use the task of detecting misalignment between T1w MRI and PET pairs, a task which is relatively easy for human visual assessors but difficult to robustly implement with traditional metrics. We start from well-aligned MRI-PET pairs that have been manually assessed, introduce synthetic misalignments through random rigid transformations and test the MLLMs' ability to detect this misalignment.

There are a number of axes of variation that can be explored when it comes to utilising MLLMs for image alignment. Here, we explore three: MLLM model choice and their performance over time, zero-shot vs. few-shot prompting, and image presentation method.

MLLM model choice We test a number of frontier MLLM models released over recent months and years as detailed in Table 1. We additionally test a local MLLM that is trained for performance on medical text and image comprehension, MedGemma (Sellergren et al., 2026).

Model Family	Model	Release Date
GPT	GPT-5.2	Dec '25
	GPT-5.4	Mar '26
	GPT-5.6 Luna	Jul '26
	GPT-5.6 Terra	Jul '26
	GPT-5.6 Sol	Jul '26
	GPT-6 Astra	Sep '26
Gemini	Gemini 3 Flash	Dec '25
	Gemini 3.1 Pro	Feb '26
	Gemini 3.6 Flash	July '26
	Gemini 3.7 Flash	Aug '26
	Gemini 3.8 Flash	Sep '26
Claude	Claude Opus 4.6	May '26
	Claude Sonnet 5	Jun '26
	Claude Opus 5	Jul '26
	Claude Fable 5.1	Sep '26
MedGemma[1]	MedGemma 1.5 4B	May '26
ResNet[2]	ResNet-18 11M	'16

Table 1: Models tested in this work. [1] MedGemma tested both as zero/few-shot MLLM in Experiment 1 and locally fintuned model in Experiment 2. [2] ResNet tested as locally finetuned model in Experiment 2.

Zero-shot vs. few-shot prompting We test the zero-shot performance of MLLMs by simply prompting them with a question to determine whether the images are aligned, followed by the actual test image (see Figure 3 for example input images). Additionally, we test few-shot performance using an in-context learning approach similar to Han et al. (2026) where we embed a few example images with labels into the prompt. Example images come from an additional dataset including subjects not used for the other experiments. See Appendix A for examples of the prompts used.

Image presentation method We test two different methods of presenting the images to the MLLMs inspired by how humans might visually assess alignment to see how it affects performance. Firstly, simply overlaying the T1w MRI on the PET with partial transparency. We term this approach "alpha-blend". Secondly, overlaying outlines of anatomical segmentation on the T1w MRI on top of the PET. We term this approach "contours". In each case, we present the models with three views of the brain, a central slice from each of the axial, sagittal and coronal views. Figure 3 demonstrates these two approaches.

2.1.1 DATA

For experiments, we use data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). For the main MRI-PET test set, we use ^{18}F-AV-45 (florbetapir) PET (Camus et al., 2012) and T1w MRI acquired from at the same session for 52 sessions, from 29 subjects from the ADNI database. The PET images acquired had undergone the standard ADNI preprocessing. We obtain the raw T1w MRI scan acquired at the same session for each PET image.

2.1.2 GROUND TRUTH ALIGNMENT

We align each subject's PET image to the corresponding T1w image using an in-house pipeline that runs three different rigid alignment algorithms and attempts to choose the best aligned method. The attempted alignment algorithms are: ANTs (Avants et al., 2008; 2011), `mri_coreg` from FreeSurfer (Fischl, 2012) and Polaffini (Legouhy et al., 2026b). ANTs and FreeSurfer are image-similarity based, using mutual information to optimise alignment, while Polaffini is segmentation-

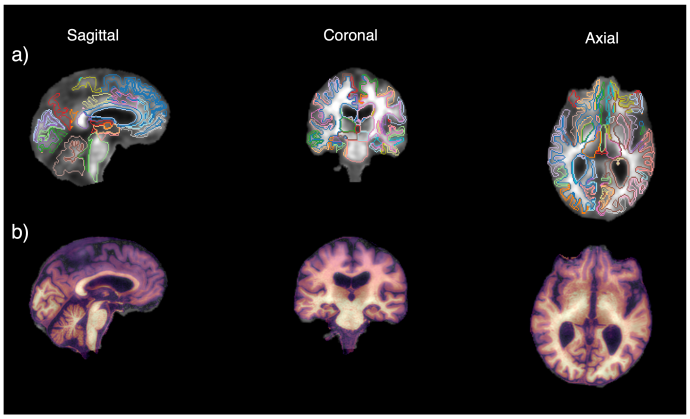

Figure 3: Example input images in an aligned case for a) contours presentation with anatomical segmentation contours from T1w MRI overlaid on the PET and b) alpha-blend presentation with the T1w image overlaid directly on the pet with 50% opacity.

based, using SynthSeg (Billot et al., 2023a;b) to segment both the T1w and PET. To select the best aligned method, we use mutual information multiplied by the dice similarity coefficient between SynthStrip (Hoopes et al., 2022) brain masks derived from both the MRI and PET. As discussed above, image-similarity and segmentation-based metrics are not entirely reliable measures of image alignment, particularly in the context of MRI-PET. SynthSeg and SynthStrip, although developed to be robust, are not completely reliable in PET imaging, while mutual information is only a surrogate measure of image alignment. We manually visually inspect each output to ensure alignment before moving on with the experiments.

2.1.3 Introduced misalignment

For each manually verified aligned pair, we introduce synthetic misalignment by introducing a random rigid transformation to the PET image. We sample random translations in x,y,z $\in [2.5, 10]$ mm and rotations about x,y,z $\in [3.5, 10]^\circ$, i.e. implicitly setting a misalignment classification threshold at 2.5mm, 3.5°. For each pair, we randomly select whether to introduce a translation-only, rotation-only or combined misalignment and then further randomly sample along which axes the transform(s) should apply.

2.1.4 Performance evaluation

The primary metric that we use to evaluate performance is accuracy of misalignment prediction from each MLLM against the ground truth label. For each MRI-PET pair, each MLLM is tested on the ground-truth aligned case and one random misaligned case, meaning that we have 104 total cases, 52 aligned and 52 misaligned.

2.2 Experiment 1b: T1w-FLAIR MRI Alignment

After benchmarking the MLLMs on MRI-PET alignment, we test a subset of MLLMs in the T1w-FLAIR setting. For this experiment, we simplify our model selection by taking two models per model family: one that is positioned as the 'frontier' model for that family, and one that is positioned as a 'cost-effective' model for that family. Furthermore, based on the results of Experiment 1a, we limit this evaluation to the best performing settings: alpha-blend image presentation with either zero-shot or one aligned and one misaligned example few-shot prompting. The same approach to introducing misalignment and performance evaluation is used as in Section 2.1.

2.2.1 Data

To test the robustness of methods across different scanners and MRI acquisition parameters, we use data from three studies to test T1w-FLAIR alignment assessment. We use FLAIR and T1w MRI acquired at the same session for 11 sessions from 10 subjects from the ADNI database, as well as 13

sessions from 13 subjects from the A4 study (Sperling et al., 2023) and 16 sessions from 16 subjects from the MICCAI 2016 multiple sclerosis lesion segmentation challenge (Commowick et al., 2021).

2.2.2 Ground Truth Alignment

We follow a similar approach as in Section 2.2.2, however since T1w-FLAIR registration is generally more reliable, we use a single alignment pipeline based on ANTs to align skull-stripped images using SynthStrip.

2.3 Experiment 2: MLLM comparison against local models

To assess the benefits of supposed generalist MLLMs over task-specific models, we finetune a local MLLM, MedGemma, on the specific task of MRI-PET misalignment detection. Additionally, we finetune an ImageNet Deng et al. (2009) pretrained ResNet-18, following the approach of DARQ (Fonov et al., 2022) as a standard CNN baseline. Each local model is trained on the MRI-PET task and with contour based image presentation, since this is the most challenging setting identified in Experiment 1a. We test how much data is required to achieve optimal performance by varying the training dataset size from 1%-100% of the training dataset outlined in Section 2.3.1. We then test generalisability at each training set size on T1w-FLAIR task with contour based image presentation to test generalisability across modalities and the MRI-PET task with alpha-blend image presentation to test generalisability across image presentation methods. The same approach to introducing misalignment and performance evaluation is used as in Section 2.1.

2.3.1 Data

For training the local models, we use ^{18}F-AV-45 PET and T1w MRI from ADNI the same as described in Section 2.1.1. We use 2143 sessions from 1006 subjects as the total training set and 518 sessions from 125 subjects for validation. All subjects used for training and validation are distinct from the 29 used for testing. The same approach for ground-truth alignment is used as in Section 2.2.2.

3 Results

3.1 Experiment 1a: MRI-PET alignment

Performance varies greatly across both models, zero/few shot promtping and image presentation method as detailed in Table 2. Across the board, GPT-6 Astra performs strongly, while Claude Fable 5.1 is best or second best in four of the six scenarios.

Broadly speaking, model performance has improved over time, as demonstrated by Figure 4. Across most model families and tasks, the best performance was achieved by the most recent frontier models, with the exception of Gemini 3 Flash which shows particularly strong zero-shot performance on the alpha-blend task.

3.2 Experiment 1b: T1w-FLAIR MRI

Performance is similar across the T1w–FLAIR and MRI–PET experiments as shown in Table 3. Notably, Gemini shows comparatively strong performance in T1w-FLAIR compared to MRI-PET. Furthermore, we see some variance in performance across data subsets, though generally models that perform well perform well in all subsets.

3.3 Experiment 2: Comparison to local models

Both MedGemma and ResNet-18 can achieve performance on-par with or exceeding frontier MLLMs on the task they are tuned for (MRI-PET) as shown in Figure 5. Both models show reasonable generalisability to the new modality (T1w-FLAIR) with a similar trend that as dataset size increases, generalisability falls. However, neither model generalises well to the new image-presentation mode (MRI-PET alpha-blend).

	Image Presentation	Contours			Alpha-blend		
	No. Shot Examples (a/m)	0/0	1/1	1/2	0	1/1	1/2
Model Family	Model						
GPT	GPT-5.2	57.7	63.4	58.7	67.3	88.5	<u>91.3</u>
	GPT-5.4	52.9	63.5	58.7	54.8	<u>92.3</u>	86.5
	GPT-5.6 Luna	51.0	48.1	50.0	57.7	81.7	79.8
	GPT-5.6 Terra	51.0	50.0	53.8	71.2	83.7	86.5
	GPT-5.6 Sol	50.0	54.8	54.8	82.7	82.7	83.7
	GPT-6 Astra	<u>74.0</u>	**87.5**	**85.6**	**87.5**	**97.1**	**97.1**
Gemini	Gemini 3 Flash	59.0	71.2	59.6	83.7	60.6	78.8
	Gemini 3.1 Pro	62.5	66.3	58.7	<u>85.6</u>	83.7	83.7
	Gemini 3.6 Flash	71.2	76.9	72.1	71.2	60.6	66.3
	Gemini 3.7 Flash	69.2	74.0	71.2	76.0	77.9	77.9
	Gemini 3.8 Flash	68.3	74.0	67.3	70.2	81.7	77.9
Claude	Claude Opus 4.8	55.8	66.3	72.1	59.6	77.0	70.2
	Claude Sonnet 5	52.9	56.7	58.7	51.0	72.1	74.0
	Claude Opus 5	58.7	73.1	74.0	75.0	85.6	88.5
	Claude Fable 5.1	**75.0**	<u>79.8</u>	<u>78.8</u>	72.1	89.0	<u>91.3</u>
MedGemma	MedGemma 1.5 4B	50.0	53.8	51.9	50.0	52.9	48.0

Table 2: Misalignment detection accuracy across MLLM models, zero/few shot prompting and image presentation method. a/m refers to number of aligned/misaligned examples used in the few-shot case. Values reported are accuracy as a percentage. Boldface numbers denote the best performing model per-column, underlined numbers the second-best.

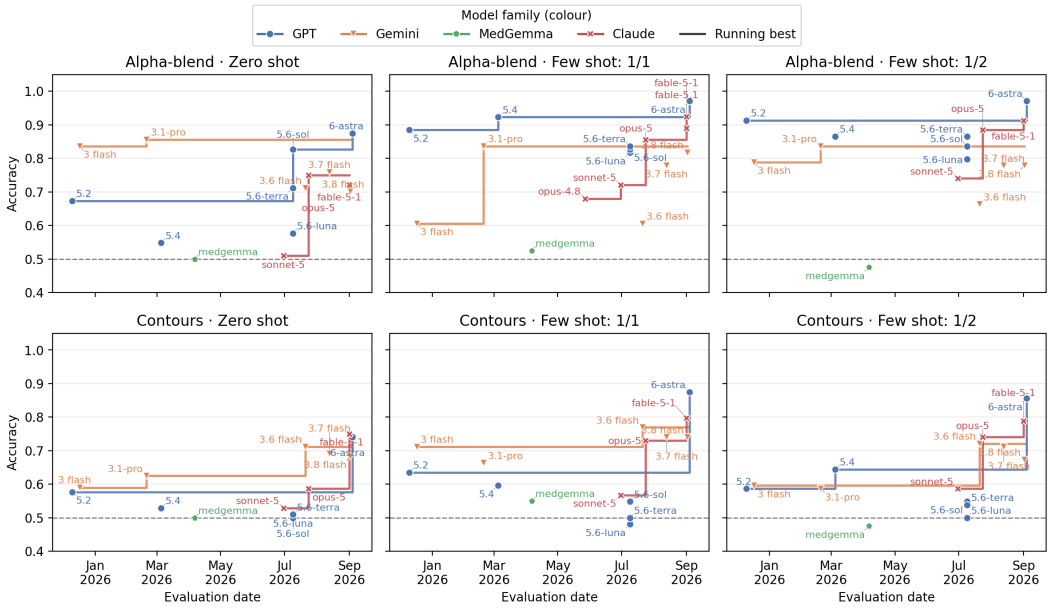

Figure 4: MLLM misalignment detection performance over time across all six scenarios tested.

4 Discussion

In this work, we have demonstrated that frontier MLLMs can be effective medical image alignment assessors. The most recent frontier models, Claude Fable 5.1 and GPT-6 Astra show strong performance in almost all settings. This is notably true of GPT-6 Astra which performs best in 7 of the 8 settings and second best in the remaining one.

Model performance over time As seen in Figure 4, MLLM misalignment assessment has generally improved over time as more and more capable models are released. In the alpha-blend image presentation case, model performance has been strong even with relatively old models such as Gemini 3 Flash and GPT-5.2, however it is only recently that frontier MLLMs can perform well in the contours setting.

		No. Shot Examples (a/m)			
		0/0		1/1	
Model Family	Model	T1w-FLAIR	MRI-PET	T1w-FLAIR	MRI-PET
GPT	GPT-5.6 Luna	65.0 (73/59/63)	57.7	91.3 (88/86/97)	81.7
	GPT-6 Astra	**91.3** (81/95/97)	87.5	**96.3** (88/100/100)	97.1
Gemini	Gemini 3.1 Pro	90.0 (96/91/84)	85.6	95.0 (92/95/97)	83.7
	Gemini 3.8 Flash	86.3 (88/100/75)	70.2	91.3 (88/95/91)	81.7
Claude	Claude Sonnet 5	53.8 (62/54/47)	51.0	66.3 (69/64/66)	72.1
	Claude Fable 5.1	76.3 (73/81/75)	72.1	86.5 (81/86/91)	89.0
MedGemma	MedGemma 1.5 4B	50.0 (50/50/50)	50.0	46.3 (42/50/47)	52.5

Table 3: Misalignment detection in the T1w-FLAIR experiment, compared to equivalent performance on MRI-PET. Small numbers in brackets represent the performance on the A4/ADNI/MICCAI subsets respectively.

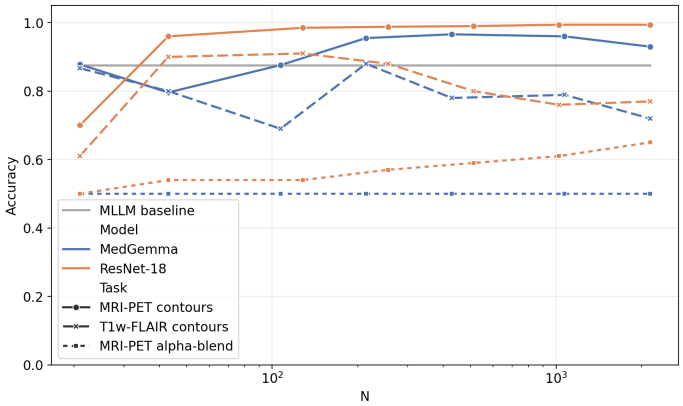

Figure 5: Local model performance with training set size for both the within-distribution test set (MRI-PET) and out-of-distribution test set (T1w-FLAIR). Grey line represents best frontier MLLM performance on the same task (MRI-PET with contour image presentation)

Zero-shot vs. few-shot There is a general trend of few-shot examples helping frontier MLLM performance, particularly when using alpha-blend image presentation as seen in Table 2. Almost universally, showing the model one aligned and one misaligned example brings a performance increase over the zero-shot case, however adding an additional misaligned example brings marginal gains. Interestingly, showing examples is not always guaranteed to improve performance; Gemini 3 Flash presents an unusual case where zero-shot performance in the alpha-blend case is strong, however performance degrades with examples.

Image presentation For almost all models, alpha-blend image presentation enables more accurate detection of misalignment compared to contours. In particular, only the most recent frontier MLLMs achieve strong performance on contours even with examples, while most models achieve strong performance on alpha-blend. One reason for this may that alpha-blend presentation makes edge misalignment apparent, while contours may be more useful in subtle internal brain misalignment as demonstrated by the examples in Figure 2.

T1w-FLAIR assessment MLLMs generally perform slightly better on the T1w-FLAIR alignment task than the MRI-PET task, with some exceptions such as GPT-6 Astra and Claude Fable 5.1 in few-shot mode where they already perform very strongly on the MRI-PET task. This result is not unexpected, since T1w and FLAIR are both MRI modalities that can reveal different aspects of anatomy, while PET measures other physiological and molecular information so the assessment of alignment with MRI is expected to be more inherently difficult.

We see some amount of variance across the data subsets (A4/ADNI/MICCAI) as seen in Table 2. One thing to note is that the single example subject given in the few shot case comes from the MICCAI challenge subset and we see in almost all models that the MICCAI subset is the best performing one. This suggests that although few-shot prompting can generally help MLLMs, we need to carefully consider which examples we select. Indeed for the Gemini models, while overall model performance improves with examples, performance in some subsets (A4 for 3.1 Pro, ADNI for 3 Flash) actually falls when given examples from a different dataset.

Comparison to local models Both local models tested achieve performance at or above the level of frontier MLLMs on the MRI-PET contour task at almost all training dataset sizes, with the exception of 1% (21 subjects) for ResNet-18 and 2% (43 subjects) for MedGemma. Eventually, both models achieve near perfect performance with accuracies >98%. The best frontier MLLM only achieves 85.6% on this task, with examples, demonstrating that task-specific models can outperform generalist ones.

Both models show reasonable performance on the T1w-FLAIR contours task across most training dataset sizes. This relatively strong performance on a modality for which they were not trained suggests that they are learning features that can detect misalignment of contours on top of multiple base image types. The T1w-FLAIR performance falls off as training dataset size grows in both cases, however, suggesting that the models are potentially learning more task-specific features at the cost of reduced generalisability when seeing more data.

Neither model generalises well to different image presentation modes when tested on the MRI-PET alpha-blend task. MedGemma performs no better than chance, always predicting "aligned" while ResNet-18 shows an interesting, inverse behaviour to T1w-FLAIR contour generalisability. As the training dataset size grows, MRI-PET alpha-blend performance grows while T1w-FLAIR contours performance drops.

One important consideration when weighing up API-based frontier models and local models is that although there may be generalisability concerns with local models, API-based models require sending data to third parties. This is of particular relevance for medical imaging data, where patient confidentiality may prevent the use of API-based models. Furthermore, the frontier models that perform strongest are also the most expensive per-token, while local models are cheaper to run after the upfront cost of hardware.

Limitations and Future Work Currently, this work only investigates two exemplar image alignment tasks chosen to be representative of the myriad use cases of image alignment assessment in medical imaging. Future work will extend this approach to other applications including within-subject motion detection and between-subject alignment for group-level analyses.

Furthermore, for practicality and cost reasons while piloting this approach across a wide range of methods, dataset sizes were kept reasonably small. The results of this work, however, will enable us to narrow down model choices moving forward and analyse more data.

A limitation of the approach taken in this study is that the misalignment we are assessing is synthetically introduced through random rigid transformations. This may not realistically reflect the kinds of misalignments we see in practice when using automated alignment techniques. For this study, since we are interested in methods that use rigid transformations to achieve alignment, we believe that the synthetic misalignment approach is reasonable and will result in realistic samples. A more careful approach will be required when considering other alignment assessment tasks such as those involving non-linear deformations.

5 Conclusion

We have demonstrated that medical image alignment assessment can be a useful test bed for generalist visual reasoning. Performance varies substantially across model generations, with only the most recent frontier models performing strongly across most scenarios. However, contour based image presentation remains a challenge, with a maximum of only 75% accuracy in the zero-shot setting.

The comparison with locally finetuned models shows a trade-off. They are able to match or outperform frontier MLLMs on the task that they are tuned for; however, this comes at the expense of reduced generalisability onto different tasks. Although local modals have an advantage in keeping data local, they need to be carefully tuned and validated for each use case.

These results support the use of MLLMs as part of medical imaging pipelines, providing evidence that these models can perform effective visual alignment assessments. Evaluation across more modalities and alignment tasks will be important next steps to establish MLLMs as routine visual image assessors. For now though, the strong performance of frontier MLLMs on these exemplar tasks is evidence of the promise of these methods for visual medical image assessment.

AI USE STATEMENT

In this work, we used generative AI tools for retrieval and discovery of related work, suggestion of modifications to figures, creation and editing of software code and suggestion of changes to text to improve readability. We have not used generative AI tools to generate synthetic data sets, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, and help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation, are not applicable to this work. We have reviewed all AI-assisted work, checking AI proposed changes to text and figures before implement and reviewing AI generated code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We aim to make our work reproducible by providing detailed descriptions of our methods, data processing and experimental design. We provide examples of the prompts used in this work in Section A to aid in reproducibility.

AUTHOR CONTRIBUTIONS

Ross Callaghan: Conceptualisation, Methodology, Software, Investigation, Visualisation, Writing – Original Draft, Writing - Review & Editing

Niannu Gao: Software, Investigation

Hojjat Azadbakht: Writing – Review & Editing, Supervision, Resources, Funding Acquisition

Hui Zhang: Conceptualisation, Methodology, Writing – Review & Editing, Supervision, Resources, Funding Acquisition.

ACKNOWLEDGMENTS

This work is supported by the National Institute for Health and Care Research Grant NIHR207476.

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; BristolMyers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

The A4 Study was a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 Study was funded by a public-private-philanthropic partnership, including funding from the National Institutes of Health-National Institute on Aging, Eli Lilly and Company, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation, and additional private donors, with in-kind support from Avid Radiopharmaceuticals, Cogstate, Albert Einstein College of Medicine and the Foundation for Neurologic Diseases.The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study was funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN Studies were led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School, and Dr.

Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI) at the University of Southern California. The A4 and LEARN Studies were coordinated by ATRI at the University of Southern California, and the data are made available under the auspices of Alzheimer's Clinical Trial Consortium through the Global Research & Imaging Platform (GRIP). The complete A4 Study Team list is available on: https://www.actcinfo.org/a4-study-team-lists/. We would like to acknowledge the dedication of the study participants and their study partners who made the A4 and LEARN Studies possible.

REFERENCES

Jesper L.R. Andersson and Stamatios N. Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, 125:1063–1078, January 2016. ISSN 10538119. doi: 10.1016/j.neuroimage.2015.10.019.

B Avants, C Epstein, M Grossman, and J Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, February 2008. ISSN 13618415. doi: 10.1016/j.media.2007.06.004.

Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, February 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.09.025.

Joshua Bierbrier, Houssem-Eddine Gueziri, and D. Louis Collins. Estimating medical image registration error and confidence: A taxonomy and scoping review. *Medical Image Analysis*, 81:102531, October 2022. ISSN 13618415. doi: 10.1016/j.media.2022.102531.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, May 2023a. ISSN 13618415. doi: 10.1016/j.media.2023.102789.

Benjamin Billot, Colin Magdamo, You Cheng, Steven E. Arnold, Sudeshna Das, and Juan Eugenio Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proceedings of the National Academy of Sciences*, 120(9):e2216399120, February 2023b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2216399120.

Marcel Bosc, Fabrice Heitz, Jean-Paul Armspach, Izzie Namer, Daniel Gounot, and Lucien Rumbach. Automatic change detection in multimodal serial MRI: Application to multiple sclerosis lesion evolution. *NeuroImage*, 20(2):643–656, October 2003. ISSN 10538119. doi: 10.1016/S1053-8119(03)00406-3.

Lando S. Bosma, Cornel Zachiu, Baudouin Denis De Senneville, Bas W. Raaymakers, and Mario Ries. Technical note: Intensity-based quality assurance criteria for deformable image registration in image-guided radiotherapy. *Medical Physics*, 50(9):5715–5722, September 2023. ISSN 0094-2405, 2473-4209. doi: 10.1002/mp.16367.

V. Camus, P. Payoux, L. Barré, B. Desgranges, T. Voisin, C. Tauber, R. La Joie, M. Tafani, C. Hommet, G. Chételat, K. Mondon, V. De La Sayette, J. P. Cottier, E. Beaufils, M. J. Ribeiro, V. Gissot, E. Vierron, J. Vercouillie, B. Vellas, F. Eustache, and D. Guilloteau. Using PET with 18F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment. *European Journal of Nuclear Medicine and Molecular Imaging*, 39(4):621–631, April 2012. ISSN 1619-7070, 1619-7089. doi: 10.1007/s00259-011-2021-8.

Olivier Commowick, Michaël Kain, Romain Casey, Roxana Ameli, Jean-Christophe Ferré, Anne Kerbrat, Thomas Tourdias, Frédéric Cervenansky, Sorina Camarasu-Pop, Tristan Glatard, Sandra Vukusic, Gilles Edan, Christian Barillot, Michel Dojat, and Francois Cotton. Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. *NeuroImage*, 244:118589, December 2021. ISSN 10538119. doi: 10.1016/j.neuroimage.2021.118589.

Mahsa Dadar, Vladimir S. Fonov, and D. Louis Collins. A comparison of publicly available linear MRI stereotaxic registration techniques. *NeuroImage*, 174:191–200, July 2018. ISSN 10538119. doi: 10.1016/j.neuroimage.2018.03.025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848.

Baudouin Denis De Senneville, José V Manjón, and Pierrick Coupé. RegQCNET: Deep quality control for image-to-template brain MRI affine registration. *Physics in Medicine & Biology*, 65 (22):225022, November 2020. ISSN 0031-9155, 1361-6560. doi: 10.1088/1361-6560/abb6be.

Vincent Doré, Santiago Bullich, Christopher C. Rowe, Pierrick Bourgeat, Salamata Konate, Osama Sabri, Andrew W. Stephens, Henryk Barthel, Jurgen Fripp, Colin L. Masters, Ludger Dinkelborg, Olivier Salvado, Victor L. Villemagne, and Susan De Santi. Comparison of ^{18} F-florbetaben quantification results using the standard Centiloid, MR-based, and MR-less CapAIBL ® approaches: Validation against histopathology. *Alzheimer's & Dementia*, 15(6):807–816, June 2019. ISSN 1552-5260, 1552-5279. doi: 10.1016/j.jalz.2019.02.005.

Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O. Koyejo, Russell A. Poldrack, and Krzysztof J. Gorgolewski. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, 12(9):e0184661, September 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0184661.

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, August 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.01.021.

Vladimir S. Fonov, Mahsa Dadar, The Prevent-Ad Research Group Adni, and D. Louis Collins. DARQ: Deep learning of quality control for stereotaxic registration of human brain MRI to the T1w MNI-ICBM 152 template. *NeuroImage*, 257:119266, August 2022. ISSN 10538119. doi: 10.1016/j.neuroimage.2022.119266.

Nick C. Fox and Peter A. Freeborough. Brain atrophy progression measured from registered serial MRI: Validation and application to alzheimer's disease. *Journal of Magnetic Resonance Imaging*, 7(6):1069–1075, November 1997. ISSN 1053-1807, 1522-2586. doi: 10.1002/jmri.1880070620.

Karl J. Friston, Steven Williams, Robert Howard, Richard S. J. Frackowiak, and Robert Turner. Movement-Related effects in fMRI time-series. *Magnetic Resonance in Medicine*, 35(3):346–355, March 1996. ISSN 0740-3194, 1522-2594. doi: 10.1002/mrm.1910350312.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William

Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Has-

san, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke,

Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair,

Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A Family of Highly Capable Multimodal Models, 2023.

Douglas N. Greve, Claus Svarer, Patrick M. Fisher, Ling Feng, Adam E. Hansen, William Baare, Bruce Rosen, Bruce Fischl, and Gitte M. Knudsen. Cortical surface-based analysis reduces bias and variance in kinetic modeling of brain PET data. *NeuroImage*, 92:225–236, May 2014. ISSN 10538119. doi: 10.1016/j.neuroimage.2013.12.021.

Douglas N. Greve, David H. Salat, Spencer L. Bowen, David Izquierdo-Garcia, Aaron P. Schultz, Ciprian Catana, J. Alex Becker, Claus Svarer, Gitte M. Knudsen, Reisa A. Sperling, and Keith A. Johnson. Different partial volume correction methods lead to different conclusions: An 18F-FDG-PET study of aging. *NeuroImage*, 132:334–343, May 2016. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.02.042.

Keqi Han, Songlin Zhao, Yao Su, Xiang Li, Yixuan Yuan, Lifang He, and Carl Yang. Towards a Virtual Neuroscientist: Autonomous Neuroimaging Analysis via Multi-Agent Collaboration, 2026.

Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. SynthStrip: Skull-stripping for any brain image. *NeuroImage*, 260:119474, October 2022. ISSN 10538119. doi: 10.1016/j.neuroimage.2022.119474.

Sunghwan Kim, Sheng-Min Wang, Dong Woo Kang, Yoo Hyun Um, Eun Ji Han, Sonya Youngju Park, Seunggyun Ha, Yeong Sim Choe, Hye Weon Kim, Regina Ey Kim, Donghyeon Kim, Chang Uk Lee, and Hyun Kook Lim. A Comparative Analysis of Two Automated Quantification Methods for Regional Cerebral Amyloid Retention: PET-Only and PET-and-MRI-Based Methods. *International Journal of Molecular Sciences*, 25(14):7649, July 2024. ISSN 1422-0067. doi: 10.3390/ijms25147649.

Alex Lallement, Vincent Noblet, Delphine Antoni, and Philippe Meyer. Detecting and quantifying spatial misalignment between longitudinal kilovoltage computed tomography (kVCT) scans of the head and neck by using convolutional neural networks (CNNs). *Technology and Health Care*, 31(4):1253–1266, June 2023. ISSN 09287329, 18787401. doi: 10.3233/THC-220519.

Antoine Legouhy, Ross Callaghan, Yuchuan Qiao, Whitney Stee, Philippe Peigneux, Hojjat Azadbakht, and Hui Zhang. Eddeep: A deep-learning framework for fast eddy-current distortion correction in diffusion MRI, July 2026a.

Antoine Legouhy, Cosimo Campo, Ross Callaghan, Hojjat Azadbakht, and Hui Zhang. Polaffini: A feature-based approach for robust affine and polyaffine image registration, 2026b.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmat-

ullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. GPT-4o System Card, 2024.

Y. Picard and C.J. Thompson. Motion correction of PET images using multiple acquisition frames. *IEEE Transactions on Medical Imaging*, 16(2):137–144, April 1997. ISSN 02780062. doi: 10.1109/42.563659.

Martin Reuter, Nicholas J. Schmansky, H. Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, July 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.02.084.

T. Rohlfing. Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153–163, February 2012. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2011.2163944.

Soorena Salari, Amirhossein Rasoulian, Hassan Rivaz, and Yiming Xiao. Dense Error Map Estimation for MRI-Ultrasound Registration in Brain Tumor Surgery Using Swin UNETR. In *2023 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, Montreal, QC, Canada, September 2023. IEEE. ISBN 979-8-3503-4645-9. doi: 10.1109/IUS51837.2023.10306435.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, Harshad Purandare, Abhishek Bijay Mishra, Sam Schmidgall, Tao Tu, Anil Palepu, Chunjong Park, Tim Strother, Rahul Thapa, Yong Cheng, Preeti Singh, Kat Black, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Joelle Barral, Tris Warkentin, Shravya Shetty, Dale Webster, Sunny Virmani, David F. Steiner, Can Kirmizibayrak, and Daniel Golden. MedGemma 1.5 Technical Report, May 2026.

Reisa A. Sperling, Michael C. Donohue, Rema Raman, Michael S. Rafii, Keith Johnson, Colin L. Masters, Christopher H. Van Dyck, Takeshi Iwatsubo, Gad A. Marshall, Roy Yaari, Michele Mancini, Karen C. Holdridge, Michael Case, John R. Sims, and Paul S. Aisen. Trial of Solanezumab in Preclinical Alzheimer's Disease. *New England Journal of Medicine*, 389(12): 1096–1107, September 2023. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa2305032.

Hu Ye, Koon-Pong Wong, Mirwais Wardak, Magnus Dahlbom, Vladimir Kepe, Jorge R. Barrio, Linda D. Nelson, Gary W. Small, and Sung-Cheng Huang. Automated Movement Correction for Dynamic PET/CT Images: Evaluation with Phantom and Patient Data. *PLoS ONE*, 9(8):e103745, August 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0103745.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, 2023.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark, 2024.

A APPENDIX

EXAMPLE PROMPTS

Zero-shot contours prompt

```
This image contains three views of the same brain: sagittal, coronal,
    and axial.

The grayscale background is a PET image. The coloured anatomical
    contours come from a T1-weighted MRI segmentation.

Determine whether the PET image and the T1-weighted MRI are spatially
    aligned.

Return exactly one label:
ALIGNED
MISALIGNED

Do not explain your answer.

<test_image>
```

Zero-shot alpha-blend prompt

```
The image contains three views of the same brain: sagittal, coronal,
    and axial.

Each view is an alpha blend: PET is the base image and the coloured
    T1-weighted
MRI is overlaid at 0.5 opacity.

Determine whether the PET image and the T1-weighted MRI are spatially
    aligned.

Return exactly one label:
ALIGNED
MISALIGNED

Do not explain your answer.

<test_image>
```

Few-shot prompt template

```
<zero-shot template>

Example 0 (known classification):
<example_0_img>
Known answer for example 0: <label_0>

Example 1 (known classification):
<example_1>
Known answer for example 1: <label_1>

Test case: classify only the following image(s), using the examples
    above as guidance.
<test_image>
Answer for the test case:
```