Evaluation of General-Purpose AI for Mouse Grimace Scale Assessment and Implications for Animal Welfare Metrics

Gerald Bieber[1]

[1]Fraunhofer IGD Rostock, Joachim-Jungius-Straße 11 18059 Rostock

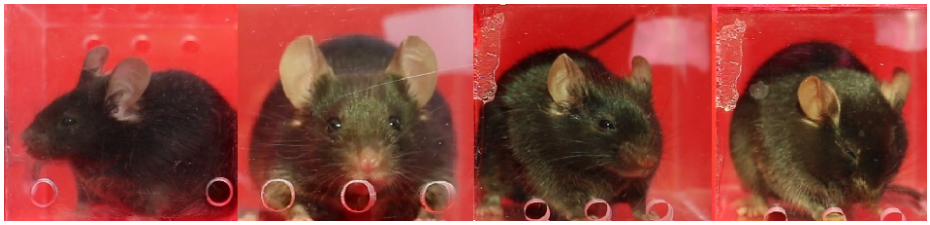

Fig. 1: Variations in stress levels in mice are reflected in the grimace scale.

Abstract. Humans and animals express internal states such as pain, stress, and discomfort through behavioral and facial cues. The Mouse Grimace Scale (MGS) is a widely used tool for assessing such states in laboratory mice, but its manual application is labor-intensive, subjective, and prone to inter-rater variability.

In this work, we introduce AniScale, a modular evaluation framework for the standardized comparison of human and AI-based scoring, and systematically evaluate whether contemporary general-purpose AI models with vision capabilities can support or replace manual MGS assessment. Our results show that large-scale AI models achieve performance comparable to inexperienced human raters while providing full reproducibility. Furthermore, we identify systematic limitations of the MGS itself, including the varying reliability of individual action units and the lack of parameter weighting.

We argue that AI-based approaches should not only be evaluated as replacements for human raters but also as tools to improve the objectivity and design of animal welfare metrics. Our findings suggest that future work should focus on learned weighting schemes and temporal aggregation.

Keywords: MGS, artificial intelligence, contactless, vital signs, facial action units

1 Introduction

The detection of animal welfare and of distress in experimental and farm animals is essential for the legally compliant implementation of reducing pain, suffering, and harm, and for the continuous development of refinement strategies in research and agriculture [19]. Among experimental animals, mice are the most commonly used laboratory species globally. In Germany, mice account for 73% of the animals used in scientific research, reflecting their dominance in laboratory settings [15]. On a broader scale, more than four million mice were utilized across Europe in research and testing in 2022, highlighting their significant role in advancing scientific knowledge [4]. This widespread use of mice in research is reasonable because of the fully genomic identification. The high usage underscores the importance of effective methods for evaluating their welfare and distress. One such tool is the Mouse Grimace Scale (MGS), a widely recognized and validated method for assessing various pain modalities in laboratory mice [38],[6],[17]. To evaluate distress, discomfort and pain according to the MGS, trained medical professionals score distinct facial action units (AUs) [16]. However, manually assessing these action units is time-consuming, labor-intensive, and prone to bias and subjective interpretation by the human scorer [33]. This raises the question of whether current advances in visual computing and artificial intelligence (AI) can enable automated or semi-automated approaches to MGS in order to reduce human workload. This paper makes the following contributions:
We present the first systematic evaluation of general-purpose vision-enabled AI models for Mouse Grimace Scale (MGS) assessment across multiple architectures and scales. Furthermore, we introduce AniScale, a modular and extensible framework for standardized MGS evaluation, enabling reproducible comparisons between human raters and AI systems. In addition, we provide empirical evidence that general-purpose AI models reach performance comparable to untrained human raters, while remaining below expert-level performance. This paper identifies structural limitations of the MGS, including variability across action units and the implicit assumption of equal weighting. Finally, we highlight the potential of AI systems not only as evaluators but as tools to improve the design, objectivity, and reproducibility of animal welfare metrics.

2 Related work

2.1 Mouse Grimace Scale

The Mouse Grimace Scale is a standardized, non-invasive tool for the assessment of pain in mice. It was originally developed and described by Langford et al. [16]. The evaluation of mouse welfare is achieved through the scoring of five facial features, the so-called action units (AU), by trained medical professionals. The AUs, which consist of Orbital tightening, Nose bulge, Cheek bulge, Ear position and Whisker change, are then scored on a scale of 0 (not present), 1 (moderate) or 2 (severe). Furthermore, it is supposed to aggregate or calculate the mean of the

AU scores in order to derive a definitive MGS score [16]. The MGS is widely recognized and utilized, with most applications relying on retrospective scoring of mouse images or videos rather than live assessment [17]. Additionally, the scope of application has also been reevaluated. Contrary to the initial study by Langford et al. [16], which posited that the MGS is mainly suitable for acute pain, subsequent studies have demonstrated that MGS scores can be detected up to one month following the initial injury across a range of different pain types, including neuropathic, visceral, and mixed pain [38]. Furthermore, it must be recognized that the administration of anesthesia and analgesia can elevate the MGS score for a duration of up to 24 hours [38],[9],[10].

Given the pervasive reliance on the MGS, it is crucial to examine the evidence supporting its validity. Numerous studies have demonstrated a discernible impact on MGS scores in response to painful events for subjects [6],[17]. To determine whether the MGS indexes pain rather than other influences, Whittaker et al. conducted a comprehensive review of extant literature to ascertain the construct validity [38]. The researchers identified supporting evidence in the form of concordant changes in mechanical allodynia, general clinical and composite pain-behavior scores, and "luxury" behaviors (e.g., nest building and burrowing), which largely tracked MGS. Nonetheless, it is imperative to acknowledge the potential variability in MGS scores attributable to factors such as observers [38], the circadian cycle [7] and biological variations among mice, encompassing factors such as sex [13] and strain [21]. Separate from validity, the reliability of scoring across raters and sites is a critical concern for pain scales [38]. Despite the assertion in certain publications that inter-rater agreement is generally good to excellent, reporting is inconsistent, there is intra-rater variability [38], and agreement can drift over time [11] and across sites [14].
The manual MGS scoring process is generally characterized by its labor-intensiveness [33], the necessity of highly trained experts [16], [38], and its susceptibility to bias [11]. These limitations highlight the need for automation and have driven the development of automated and semi-automated MGS solutions. Many of these solutions leverage artificial intelligence and machine learning.

2.2 Existing Approaches

Numerous attempts have sought to automate the scoring of MGS, including the development of several automated pipelines [18], [2], [36], [8]. One such approach is the cloud-based platform *PainFace* [18]. The use of machine learning enables the detection of four facial AUs: orbital tightening, nose buldge, whisker change, and ear position. This approach subsequently generates per-frame numeric MGS scores ranging from 0 to 8. The underlying models were trained using a diverse training set consistent of over 70,000 annotated images from 214 C57BL/6 mice. The detection of the AUs has a precision of over 90%, and model-derived MGS scores differ from human ratings by no more than $\pm$ 0.5 MGS units. Furthermore, *PainFace* has been validated across various laboratories and pain models, including laparotomy surgery and carrageenan and formalin injections. However,

cheek bulge was excluded from the scale due to its unreliability in detection. Additionally, videos must meet specific quality standards to ensure reliable scoring, and images cannot be scored [18].

Another proposed solution involves the implementation of a semi-automated pipeline that utilizes a detector based on OpenCV [2]. This detector employs a boosted cascade of simple features to detect mouse faces. The pipeline then uses pretrained deep connected neural networks, based on the *ResNet50* or *InceptionV3* architecture, to classify images into two categories: "post-anesthetic / surgical effect" and "no effect". The dataset employed for model training consisted of 18,273 face crops of 126 freely moving black-furred C57BL/6JRj mice, all of which met the requisite quality standards. The performance demonstrated an accuracy of 98.9% in correctly classifying the mouse following the injection of ketamine/xylazine, 90.1% following the inhalation of isoflurane, and 89.8% following castration when averaging confidence over multiple images per timepoint (cross-validation). However, this approach remains binary and does not calculate a numeric MGS. Moreover, it necessitates manual remediation of detections and is thus merely a semi-automated solution [2].

The *automated Mouse Grimace Scale* (aMGS) [36] retrains the *InceptionV3* Convolutional Neural Network (CNN) on human-scored images to output binary "pain/no pain" classifications with a confidence score. The aMGS utilizes the *Rodent Face Finder* [31] to automatically extract face frames for training and inference. A balanced training set of 6,862 images (derived from 5,771 unique "pain/no pain" images plus augmented variants) achieved internal accuracy of 93.2% when restricting to high-confidence images (>0.75), and 94% accuracy on a held-out validation set [36]. However, akin to the previously described solution [2], the aMGS is merely a binary classifier, lacking the provision of a numeric MGS score.

A different approach [8] automates the quantification of a single MGS action unit – orbital tightening – by training *DeepLabCut* to label the superior/inferior eyelid margins and medial/lateral canthus, then computing eyelid distance and palpebral fissure width from 1920×1080, 24 fps videos of mice positioned in sheltering tubes. The method produces a quantitative, continuous orbital tightening metric, as opposed to a categorical multi facial AUs MGS metric. It necessitates the implementation of lab-specific *DeepLabCut* models and the establishment of stringent image quality criteria [8].

A close examination reveals several notable parallels among the four approaches under consideration. For instance, these systems are designed to reduce subjectivity and labor involved in manual MGS scoring. They rely on the utilization of deep learning to automate pain inferences from images or videos. However, some solutions offer only binary classification [2,**?**], while others focus on scoring only one action unit [8]. Furthermore, these methods are character-

ized by their high degree of application-specificity, with their performance being closely associated with the characteristics of the training data. Therefore, the transferability of these pipelines among diverse laboratories and setups poses significant challenges.

Another approach [33] attempts to resolve this issue by augmenting the software solution with the appropriate hardware. The *GrimACE* is a standardized, portable cage-side system that utilizes front- and top-view infrared cameras in a dark, IR-permeable acrylic arena to automatically assess mouse pain via the MGS. The automated MGS scoring employs three sequential models: a modified *MobileNetV3* functioning as a frame-quality network, a *YOLOv8s* serving as a mouse face detector, and a *ViT*-based classifier to score the five AUs. The training of these models has been conducted on a dataset comprising over 4,400 images in total that have been manually annotated, labeled, and scored. The validation process was executed in three experiments, which demonstrated a high degree of agreement between the automated and expert MGS measurements (Pearson's r=0.87). Automated scoring exhibited minimal overall bias; however, it tended to underestimate nose, cheek, and whisker scores due to the limited number of high-pain training examples [33]. Furthermore, concerns regarding the scalability and transferability of the solution persist, as hardware components are also necessary in addition to software.

The issue of the solutions' strong application-specificity and dependence on the underlying training data prompts the question of whether specialized AI solutions are necessary, or whether general AI can adequately perform MGS.

3 Concept and Methodological Framework

In order to effectively assess the performance of general AI systems in MGS, it is necessary to clarify the concept and methodological approach. The objective of this study is to assess a wide range of AI models by employing a standardized protocol and subsequently comparing their performance. Even the evaluation of current models might be outdated by expecting new upcoming models, the research will provide insights of the current performance of AI for MGS and the needs for further investigations.

Of particular interest are open-source models that can be executed locally, as these models offer the greatest flexibility in terms of subsequent integration or fine-tuning. Nevertheless, state-of-the-art models such as *OpenAI*'s *ChatGPT* series should also be evaluated, as these models exhibit superior computing power in comparison to local models. It is imperative that communication with these models is facilitated through the application programming interface (API) to ensure standardization.

For the optimal approach, an implementation of a defined framework to unify communication, queries, and to exchange models, is required. Another salient topic is the standardization of images. All images must conform to a consis-

tent format, be normalized to a common resolution, and pass quality checks. Cropped images should clearly show the mouse's face and be free of blur, occlusions, or other artifacts. In order to ensure the comparability of responses from multiple AIs, it is imperative that they are required to respond in a uniform format. Moreover, it is necessary to maintain parameters that could influence the AI's response behavior, such as AI-prompt or processing parameters (e.g. AI-temperature). The optimal settings for these parameters should be explored in order to optimize the quality of the results (see 4.4). Moreover, it is imperative to meticulously delineate the metrics employed to assess the efficacy of AI systems, thereby facilitating meaningful comparisons. In order to establish a foundation for comparison, it is recommended that the MGS evaluation is also carried out by humans, individuals with minimal experience and by scorers with extensive training and experience. Furthermore, a randomly generated result can be utilized as a benchmark to assess the lower bound of accuracy of the AI models' predictions. Nevertheless, a pivotal aspect of the methodological approach relates to the underlying test data set, which serves as the ground truth. In order to test the AI models comprehensively, it is necessary that certain characteristics are met by the dataset.

3.1 Ground Truth Dataset

Because an open dataset is not existing, a ground truth dataset was extracted from a dataset consisting of 463 images. These images were contributed by University of Rostock, Germany, and evaluated by two trained, experienced individuals. The dataset was collected during an experiment in which the mice went through several phases. All procedures involving animal were approved by the local regulatory authority, Landesamt für Lebensmittelsicherheit und Fischerei Mecklenburg-Vorpommern, Germany (Az: 7221.3-1-035/20). In the initial phase, designated as the "pre-phase," the mice were recorded prior to the administration of analgesics or the induction of any other form of pain. In the subsequent phase, the subjects were administered analgesics, including metamizole, tramadol, and carprofen, via their drinking water. In the third phase, the mice underwent surgery in which a transmitter was implanted intraperitoneally and the electrodes were sutured to the muscle, followed by a recovery period. The final phase of the experiment consists of bile duct ligation (BDL), a procedure that causes pain and cholestasis. This procedure is also painful and stressful for the mice. During these phases, the mice were filmed in observation cages (Figure 1) similar to those described by Langford et al. [16]. The images were automatically extracted from this video material in which the mouse's face is clearly visible and the quality is sufficient.

The selection of images from the larger dataset to generate a ground truth dataset enabled a fully balanced and uniform distribution. Consequently, the dataset should contain equal numbers of images for each of the 11 potential MGS scores (0-10). This is because the five AUs are represented only by non-negative integer values in the range '0, 1, 2', so the sum lies in the interval from 0 to 10. To enhance the comparability of the images, they were normalized to a resolution of

512x512. This process yielded an image size of approximately 200-400 kilobytes. To ascertain an optimal equilibrium between sample size and memory size of the ground truth dataset, it was determined that a total of 66 images would be utilized. The dataset size was intentionally constrained to enable systematic and cost-efficient evaluation while maintaining a balanced score distribution. This results in a final sample size of approximately 20 MB, which represents an optimal balance between cost-effectiveness when using paid models and the maintenance of a substantial sample size.

To attain a ground truth dataset comprising precisely 66 images, there must be exactly 6 images per possible MGS score. To this end, images were systematically selected from the underlying dataset of 463 images. All images in which each AU was rated identically by both scorers were incorporated into the ground truth dataset. This dataset is regarded as ground truth and employed for the evaluation of the models.

These 66 uniformly distributed images (Figure 2) originate from six male, black C57Bl/6J mice. We acknowledge that the number of animals is limited and may not fully capture biological variability. This aspect will be addressed in future studies. The objective of an evenly distributed dataset is to ascertain whether the AI model systematically overestimates or underestimates the true MGS score, or whether the results are approximately evenly distributed.

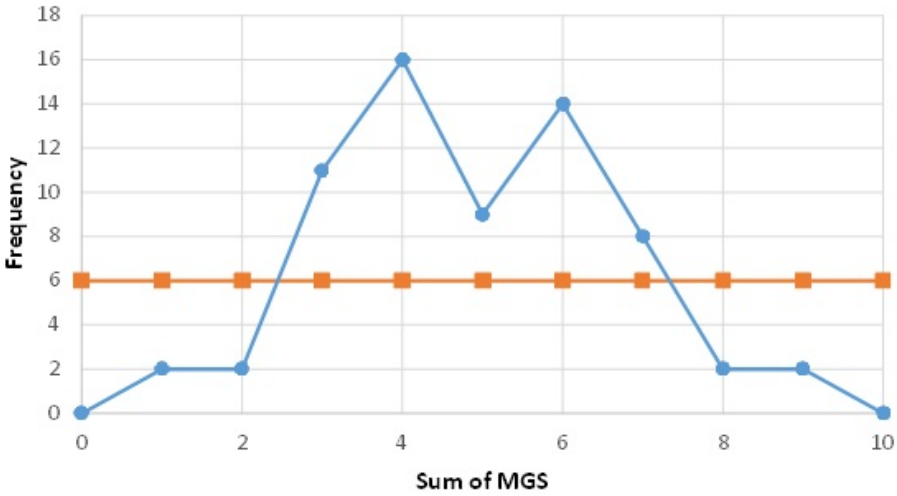

Fig. 2: Histogram of the distribution of MGS in the dataset. Red graphs shows the uniform distribution of the test-set, the blue graphs illustrates the histogram of a random grimace scale test-set.

.

4 Trials and Evaluation Workflow

The trials were conducted under the following hypothesis:

- H1: General-purpose AI systems can achieve performance comparable to human raters under controlled conditions.
- H2: Experts in the field of MGS perform significantly better than untrained and inexperienced humans.

4.1 Evaluation Metrics

In order to test this hypothesis, it is necessary to define which parameters and metrics will be used to measure how good the respective answers are. In this study, performance was evaluated using the following metrics: Spearman's rank correlation [32], mean absolute error (MAE), F1-score, and quadratic weighted kappa (QWK) [3].

Spearman's Correlation The calculation of the Spearman rank correlation was performed by employing a script that utilized the *scipy.stats.spearmanr* function. The following formula is employed:

$$R_i = \text{rank}(y_i), \quad S_i = \text{rank}(\hat{y}_i), \tag{1}$$

$$\rho = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}, \tag{2}$$

and if there are no ties:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \quad d_i = R_i - S_i. \tag{3}$$

For samples with both totals present, y_i is mgs_gold and $\hat{y}_i$ is total_mgs. $R_i = \text{rank}(y_i)$ and $S_i = \text{rank}(\hat{y}_i)$ use average ranks for ties (common with totals in $[0, 10]$); n is the number of paired totals. Spearman's ρ is the Pearson correlation between (R_i) and (S_i); equivalently (without ties), $\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$ with $d_i = R_i - S_i$.

Mean Absolute Error MAE reports the typical magnitude of errors in the original units and is robust to outliers. Spearman's ρ is a measure of monotonic association, focusing on whether predictions maintain the original rank ordering of targets. Hence, these two metrics can be utilized to assess overall performance from multiple perspectives. The mean absolute error (MAE) was calculated using a script that was implemented. This script utilized the *sklearn.mean_absolute_error* function. The following formula is employed:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \,. \tag{4}$$

For each AU, n is the number of evaluated samples, y_i is the gold AU label from the ground truth dataset, and $\hat{y}_i$ is the AI prediction; MAE averages $|y_i - \hat{y}_i|$ over non-missing pairs. For total MGS, $y_i = \text{mgs\_gold} = \sum_{\text{five AUs}} \text{AU}_{\text{gold}} \in [0, 10]$ and $\hat{y}_i = \text{total\_mgs}$; the MAE is computed analogously on samples where both totals are present.

F1-Score The F1-score provides a balanced metric by combining precision and recall into a single value. It is robust to imbalanced datasets, commonly encountered in real-world classification problems. Specifically, the F1-score is calculated as the harmonic mean of precision and recall, emphasizing cases where one metric is notably lower. The scikit-learn library was used to compute the F1-score with the following formula:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{5}$$

where Precision and Recall are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{6}$$

For the calculation, TP refers to the number of true positive cases, FP indicates false positives incorrectly identified, and FN represents false negatives missed by the model. Using the *f1_ score* function from *sklearn.metrics*, multiclass performance was averaged using the macro metric.

Quadratic Weighted Kappa QWK is a statistical method that quantifies agreement on ordinal outcomes. It does so by correcting for chance and penalizing larger ordinal discrepancies more heavily. Consequently, it is particularly advantageous to assess the AI's performance of scoring the AUs. The QWK was calculated with a script using the *sklearn.cohen_ kappa_ score* from the sklearn library. The fundamental formula is expressed as follows:

$$w_{ij} = \frac{(i-j)^2}{(K-1)^2}, \tag{7}$$

$$O_{ij} = \frac{N_{ij}}{\sum_{p=0}^{K-1}\sum_{q=0}^{K-1} N_{pq}}, \tag{8}$$

$$p_i^{(r)} = \sum_{j=0}^{K-1} O_{ij}, \quad p_j^{(c)} = \sum_{i=0}^{K-1} O_{ij}, \tag{9}$$

$$E_{ij} = p_i^{(r)}\, p_j^{(c)}, \tag{10}$$

$$\kappa_w = 1 - \frac{\sum_{i=0}^{K-1}\sum_{j=0}^{K-1} w_{ij} O_{ij}}{\sum_{i=0}^{K-1}\sum_{j=0}^{K-1} w_{ij} E_{ij}}. \tag{11}$$

For a given AU, i and j denote the gold-standard and AI-predicted categories in $\{0, 1, 2\}$ ($K = 3$). N_{ij} counts samples with (gold $= i$, pred $= j$); $n = \sum_{i,j} N_{ij}$ is the number of evaluated samples (after masking missing labels). $O_{ij} = N_{ij}/n$ is the normalized confusion entry; $p_i^{(r)} = \sum_j O_{ij}$ and $p_j^{(c)} = \sum_i O_{ij}$ are its row/column marginals; $E_{ij} = p_i^{(r)} p_j^{(c)}$ is the expected agreement under independence. The quadratic weight is $w_{ij} = \frac{(i-j)^2}{(K-1)^2}$, and κ_w compares weighted observed agreement to chance.

4.2 Evaluation Framework and Experimental Design

The utilization of four metrics enables the capture of complementary aspects of performance, namely ordinal consistency (Spearman), average deviation (MAE), classification balance (F1-score), and categorical agreement (QWK). To collect and evaluate these parameters, we had a selection of AI models evaluate the ground truth dataset. To this end, a comparative analysis was conducted on the AI models listed in Table 1.

In addition to the AI models, three experienced, trained individuals and two inexperienced individuals were also tasked with evaluating the images. Hereby, we define 'inexperienced' as having no or less than 10 minutes of experience [20]. Among the three experienced scorers, the scorer designated as "Experienced 1" (see 5) is one of the individuals who scored the dataset from which the ground truth was collected. The other two experienced scorers had never encountered the dataset nor had they previously scored it. The two novice scorers received no instruction or guidance. They were provided with the 66 images to score, along with the prompt that the AI is also given, in order to ensure conditions and results were as comparable as possible. Another effort to ensure the comparability of the responses was to query all models with identical temperature and prompt, where possible, and to force their responses into JSON format. Regrettably, it was not possible to calibrate the temperature for GPT 5.2 pro [28] and Grok

Model	Reference	Parameters
Gemma 3 12B	[34]	12 Billion
Gemma 3 27B	[34]	27 Billion
MiniCPM-O-2_6	[29]	ca. 2.6 Billion
Magistral Small 1.2	[1]	not published (Small-Model)
InternVL3_5 14B	[30]	14 Billion
Pixtral 12B	[5]	12 Billion
Qwen 2.5 VL 7B	[35]	7 Billion
GPT-4.1 mini	[23]	not published (approx.: < 50B)
GPT-4o	[22]	not published (approx.: $\sim$200B)
GPT-5	[25]	not published (approx.: 1.7–5T)
GPT-5 mini	[24]	not published (Mini-Version)
GPT-5.1	[26]	not published
GPT-5.2	[27]	not published
GPT-5.2 pro	[28]	not published (more than GPT-5.2)
Grok 4	[39]	not published (approx.: $>$300B)

Table 1: Overview about Model and Parameter-Size.

[39]. The AI was also given the option of not evaluating certain AUs if, in its assessment, they were not recognizable or evaluable. The final answers from the various AI models and human evaluators were recorded in comma-separated values (CSV) files. The relevant metrics can then be calculated using a script (see 5). The AI models, calculations, and scripts were executed on a computer equipped with the *Ubuntu 24.04.3 LTS* operating system, an *AMD EPYC™ 7543P x 64* processor, and 256 GiB of memory. To standardize the queries and support the test procedure, a framework was developed: the web application "AniScale".

4.3 Web Application "AniScale"

AniScale is a web application designed for the automated evaluation of MGS. The software's primary functions encompass uploading images and videos, the extraction of suitable frame crops in which the mouse's face is clearly visible, the performance of MGS analysis using an AI model, and the viewing and downloading of the final results. The web application is composed of four primary components: the frontend, the backend, the external model-serving runtime, and an additional mouse detection service, which is responsible for extracting suitable frames from the videos. The frontend was developed using *React*, *TypeScript*, and *Vite*. It is responsible for the user-friendly display of the individual functionalities and serves as an interface to the user. The backend is responsible for

implementing the core functionalities, including the process of uploading files and the communication with the AI models and the mouse detection service. The application utilizes the Python web framework *FastAPI*, which facilitates the construction of an HTTP API that serves as a conduit for communication with the backend. The external model-serving runtime provides the AI models that ultimately perform the MGS analysis via an API. By default, *Ollama* is used via a *REST API*; however, *LM Studio* is also supported. This interface is distinguished by its modular nature, a characteristic that facilitates its replacement by any AI-providing API or AI model. To generate the optimal mouse frame crops, a mouse detector microservice was developed. This can also be accessed via an HTTP API built by *FastAPI*. In general, during the development of AniScale, particular attention was devoted to the creation of a modular and expandable structure. This structure was designed to facilitate the integration of AI models that had been specifically trained for MGS evaluation, as well as the potential for conducting a Grimace Scale analysis for other animals, should such a need arise.

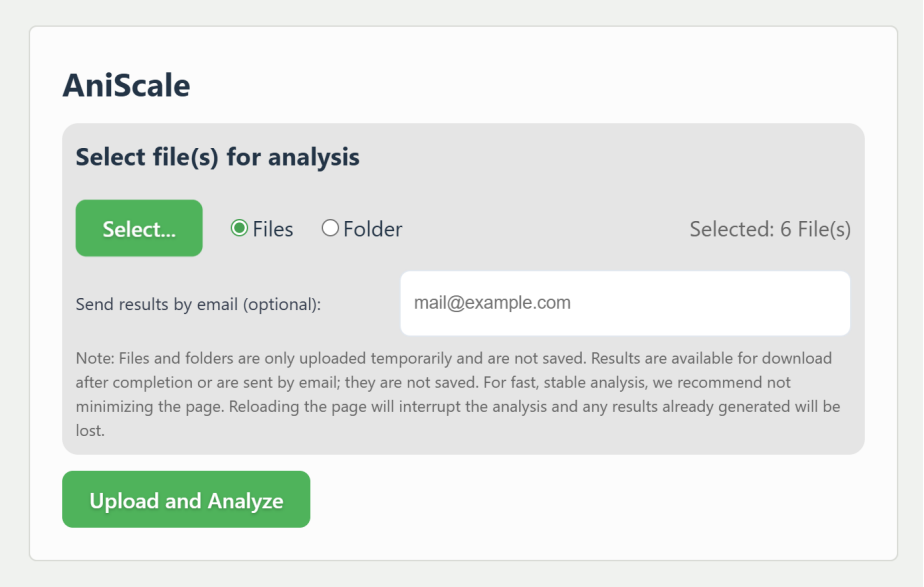

Fig. 3: Webb Application AniScale: Start page after selecting six files

Upon opening the web application, the user is presented with a range of functions (see Figure 3). The radio button serves to select whether an entire folder or individual files in the file browser should be selectable during the file selection process. The "Select..." button opens the file browser, where images in JPG, JPEG, and PNG formats, videos in MP4 and MOV formats, or entire folders containing the corresponding videos and images can be selected. Upon

the successful selection of files, the message "Selected: x File(s)" is displayed, providing the user with the number of files that have been selected. Additionally, users have the option of entering their email address. In that case, the final results will be transmitted to the specified email address in the form of an XLSX and CSV file upon the completion of the analysis. This feature is particularly advantageous for extended analyses, as it eliminates the need for the user to remain in front of the computer. Instead, the analysis can be executed in the background, and the resulting data can be efficiently delivered to the user via email.

Subsequent to the selection of files, the "Upload and Analyze" button will appear at the bottom. Activation of the button initiates the upload and analysis process. Initially, a progress bar for the upload is displayed. This upload process can be canceled at any time by clicking the red X located on the right-hand side of the progress bar. Upon the conclusion of the upload process, the mouse detection model is loaded, and the videos can be analyzed. Mice in the observation cage are detected, and frames with clearly visible faces are selected. Subsequently, the MGS evaluation model is loaded and analyses the extracted frames. The progression can be meticulously monitored by the user via two progress bars. The interface displays the file name of the video currently being processed, and the top-right corner shows the number of videos completed and the total number in the batch. Conversely, in the event that an image is being analyzed, the MGS evaluation model is loaded directly, thereby initiating the MGS evaluation without first going through the mouse detection process. In the event that further analysis is deemed unnecessary, it can also be canceled by selecting the red X.

Upon completion of the analysis, several new options become available. The results can be downloaded in the CSV and XLSX file format. The files contain a variety of columns, including the filename, the file type (designated as "image" or "video" based on the file type), the timestamp (empty for images; for videos, the timestamp indicates the point in the video at which the analyzed image was cropped), the model (the underlying AI model that performs the MGS analysis), the date (date of analysis), the time (time of analysis), orbital, nose, cheeks, ears, and whiskers (the five AUs with values of 0, 1, or 2), the sum (the five AUs added together), the total_mgs (the MGS score), the confidence (the confidence of the AI about the image), and the flag (either 0 or 1, depending on whether there were problems with the analysis). After the analysis of a video, it is also possible to download the final images that were extracted by the mouse detection model in JPG format. Additionally, the results can be viewed directly in the web application (see Figure 4). The selection of this option will result in the presentation of a list containing the individual files that have been analyzed. The file name is displayed as the heading of each box. Furthermore, the type, the timestamp, the five AUs, the sum, and the total_mgs are displayed. To facilitate a rapid assessment of the results, the AU boxes are color-coded according to their values: green for 0, yellow for 1, and red for 2. In the event that a file has been flagged, the entire file-box is colored red. This indicates that the user must manually ascertain the underlying issue. An image is flagged if the AI model does

not evaluate one of the parameters because it cannot recognize it or if something else goes wrong. Additionally, an image is flagged if the mouse detector model is unable to locate a frame that meets predefined quality standards, such as lighting, blurring, mouse detection.

Fig. 4: Web Application AniScale: Viewing results after successful analysis

The objective of the mouse detection model is to automatically suggest the most suitable mouse-face crops from videos and to identify and flag images that do not meet specific quality standards. The present study employed a two-stage approach: The initial step in the process entails the localization of mouse-face regions within the video frames by a detector. In the second step of the process, a face-quality classifier assigns a "face_good" or "not" score to each crop. These scores are then integrated into a multi-factor quality score, which is used to select diverse, high-quality suggestions. The mouse-face localization was achieved

through the implementation of a fine-tuned *Faster Regions with Convolutional Neural Networks (Faster R-CNN)* architecture, incorporating a *Mobile Neural Network Version 3 (MobileNetV3)-Large Feature Pyramid Network (FPN)* backbone. The detector was initialized with *torchvision*'s default *Common Objects in Context (COCO)* weights and trained on a *COCO*-style dataset of annotated mouse-face bounding boxes. The dataset under consideration contains a total of 966 images that have been manually annotated in *Label Studio*. The training process incorporated the *Adaptive Moment Estimation with Decoupled Weight Decay (AdamW)* optimizer (lr=5e-4, weight decay=1e-4), a *Step Learning Rate (StepLR)* scheduler (step_size=8, gamma=0.5), horizontal flip augmentation, and was executed for 20 epochs with a batch size of 4. The selection of optimal checkpoints was determined by the validation loss. To evaluate the quality of the crops, we fine-tuned a *MobileNetV3-Small* classifier that was initialized with *ImageNet1K_ V1* weights in order to predict "face_good" versus "not." A total of 1,668 crops were generated from detector outputs and manually annotated in *Label Studio* for fine-tuning the weights. To make the training images more varied and to create a more robust model, we applied RandomResizedCrop(224), horizontal flips, and color jitter and trained with *AdamW* (lr=1e-3). We also applied *Cosine Annealing Learning Rate (CosineAnnealingLR)* and class-weighted cross-entropy and selected the best model based on macro F1 on a held-out validation set (20 epochs, batch size 64). During inference, we sample frames at one frame per second and detect the quadrants of the observation cage. We compute a quality score for each crop as a weighted sum of detection confidence, normalized sharpness (Laplacian variance), centeredness within the slot, and the probability that the classifier labels it as "face_good." To pick top suggestions, we enforce temporal and visual diversity (minimum frame gap and high-SSIM suppression). Low-probability faces are flagged according to configurable thresholds. Then, the service saves the cropped images, annotated thumbnails, and a meta JSON with diagnostics and flags. *FastAPI* endpoints are used to transfer images to and from the service, to check the current status, and to retrieve the final crops.

4.4 AI Optimization and Dependencies

The performance of the model is contingent upon numerous controllable parameters; among these, the prompt and the sampling temperature are of particular consequence for output quality and reliability. In order to maximize performance, iterative prompt engineering was employed, with task instructions and constraints being progressively refined. In accordance with our working hypothesis, increasing prompt precision and detail exhibited a tendency to enhance alignment with the gold standard. Across models, the final detailed prompt yielded, on average, a 0.24 higher correlation than a simple, general prompt. However, two exceptions were observed. Specifically, GPT 5.1 [26] demonstrated a 0.06 lower correlation with the detailed prompt, while GPT 5.2 [27] exhibited a 0.05 lower correlation. These observations suggest that the model exhibits model-specific

sensitivity or interpretation to prompt form, indicating that prompt optimization benefits may not be uniformly realized across architectures.

Temperature controls how adventurous the model is when choosing the next word. At higher temperatures, the probability is distributed more uniformly across the range of possible outcomes, thereby enhancing randomness and diversity. Conversely, at lower temperatures, choices are concentrated on the most probable options, resulting in outputs that are more deterministic and consistent [12]. In practice, always selecting the most likely continuation can yield repetitive or degenerate text. The employment of sampling methods that retain a flexible "nucleus" of likely tokens has been demonstrated to improve quality relative to fixed-threshold approaches [12]. Consequently, temperature should be tuned together with the sampling strategy to balance coherence and coverage for the task at hand.

It was hypothesized that lower temperatures would yield more stable and accurate outputs, given that higher temperatures promote more exploratory generation. To test this hypothesis, we varied temperature in 0.2 increments while holding all other conditions constant and compared the resulting responses.

Contrary to the predominant hypothesis, the maximum performance was not observed at T = 0.0 (see 5). For GPT 4.1 mini [23], correlation strengthened as temperature elevated, and both GPT 5.1 [26] and GPT 5.2 [27] attained their maximum correlation with the ground truth at T = 0.4.

Concurrently, we observed that several local models yielded unaltered outputs when the temperature was adjusted, suggesting that the user-specified temperature was either disregarded or overridden. Therefore, we include only one local model in Figure 5 as a representative example and omit the others, with the object of enhancing clarity and comprehensibility of the figure.

5 Results

The AniScale framework made it possible to use a wide variety of AI models with the same prompt and test data set. The classification results are listed below and will be discussed in the following chapter.

5.1 Results of Classification

The classification results summarized in Table 2 reveal consistent performance differences between expert human raters, inexperienced humans, and AI-based models across all evaluation metrics.

With respect to ordinal consistency, measured by Spearman's rank correlation, expert raters achieve the highest agreement with the ground truth (ρ up to 0.90), indicating a strong ability to preserve the relative ordering of MGS scores. Inexperienced human raters show moderately lower correlations (ρ 0.70˜0.74), while the best-performing AI model (Grok 4) reaches a comparable level ($\rho = 0.679$). Smaller or locally deployed AI models exhibit substantially lower correlations, in some cases approaching random behavior.

max width=

Model	Spearman	MAE	Mean F1	Mean QWK
Humans				
Experienced 1	0.900	1.106	0.692	0.731
Experienced 2	0.828	1.839	0.477	0.456
Experienced 3	0.832	2.288	0.442	0.426
Inexperienced 1	0.700	1.939	0.529	0.460
Inexperienced 2	0.739	2.515	0.488	0.410
AI-Models				
Gemma 3 12b	-0.135	4.106	0.203	-0.036
Gemma 3 27b	0.100	4.258	0.203	0.041
Minicpm-o-2_6	0.137	4.621	0.203	0.057
Magistral Small 1.2	0.344	3.212	0.340	0.224
InternVL3_5 14b	0.307	3.955	0.319	0.235
Pixtral 12b	0.094	3.803	0.196	0.032
Qwen 2.5 vl 7b	0.148	3.485	0.224	0.051
Grok 4	0.679	2.833	0.407	0.476
GPT-4.1 mini	0.573	4.547	0.198	0.114
GPT-4o	0.516	4.333	0.238	0.128
GPT-5	0.428	3.697	0.285	0.160
GPT-5 mini	0.555	3.167	0.316	0.262
GPT-5.1	0.598	3.000	0.382	0.299
GPT 5.2	0.492	3.121	0.281	0.221
GPT 5.2 pro	0.519	3.649	0.306	0.221
random	0.022	3.000	0.339	0.003

Table 2: Classification performance by the used model.

A similar trend is observed for the mean absolute error (MAE), which reflects the average deviation from the ground truth in absolute units. Expert raters achieve the lowest MAE (1.1), indicating high precision in absolute scoring. In contrast, inexperienced humans and high-performing AI models show increased error levels (MAE 1.9–2.8), while smaller AI models produce significantly larger deviations (MAE >3.5), suggesting limited reliability in precise score estimation.

The F1-score, capturing the balance between precision and recall across discrete AU classifications, further highlights these differences. Experts achieve the highest F1-scores (up to 0.692), while inexperienced humans and advanced AI models reach intermediate values (0.40–0.53). Smaller AI models perform poorly (0.20–0.34), indicating difficulties in correctly classifying individual action units.

Quadratic weighted kappa (QWK), which measures agreement while accounting for ordinal distance, shows a similar pattern. Expert raters achieve substantial agreement (up to $\kappa = 0.731$), whereas inexperienced humans and the best AI models achieve moderate agreement ($\kappa = 0.41\breve{}0.48$). Smaller AI models yield low or near-zero agreement, confirming limited capability in capturing ordinal relationships.

Taken together, these results indicate that expert human raters consistently outperform all other groups across all evaluation metrics. However, large-scale AI models achieve performance levels comparable to inexperienced human raters, particularly in terms of ordinal consistency and categorical agreement. In contrast, smaller AI models exhibit substantially weaker performance across all metrics.

Importantly, the discrepancy between relatively high Spearman correlation and higher MAE values for AI models suggests that these systems capture the general ranking of distress levels but struggle with precise calibration of absolute scores. This behavior indicates that AI models learn global patterns but lack fine-grained sensitivity required for expert-level assessment.

Overall, the results demonstrate a clear performance hierarchy: expert humans > large-scale AI models and inexperienced humans > small-scale AI models. This hierarchy is consistent across all evaluated metrics and highlights both the current limitations and the potential of general-purpose AI systems for MGS assessment.

5.2 Influence of AI-parameter Temperature

LLM-based AI models (transformer) are estimating probabilities for each token. The rate of probability can be adjust by a parameter that is similar to the entropy in physics, caused by the temperature. Therefore, temperature is also an AI-parameter that controls the interplay of creativity and reproducibility. In Figure 5 the classification was repeated by changing the temperature. The graph illustrates that some models do not support temperature, others perform better or worse. The results indicate that temperature is an instrument of fine-tuning that varies between the used model and cannot be generalized.

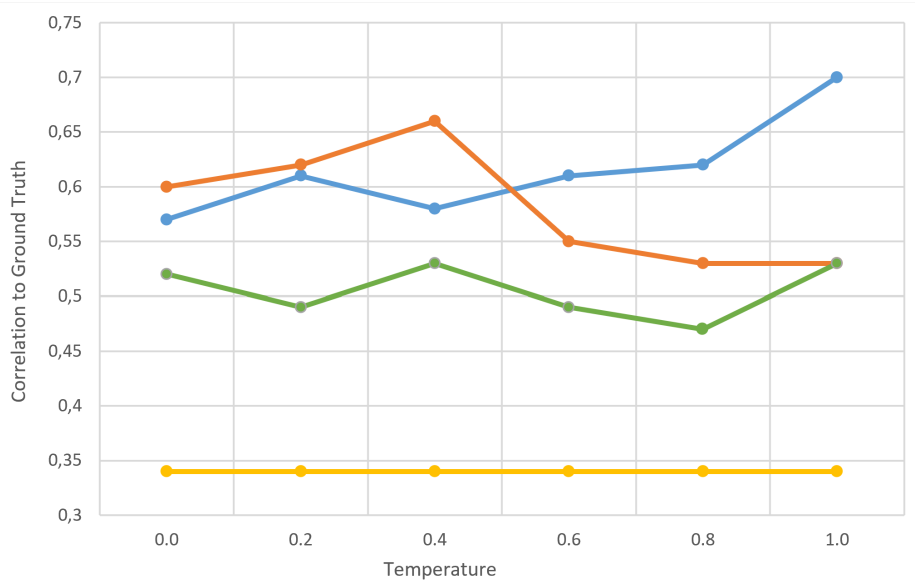

Fig. 5: Temperature dependent Spearman correlation of the AI Models GPT 4.1 mini [23] (blue), GPT 5.1 [26] (orange), GPT 5.2 [27] (green), and Magistral Small 1.2 [1] (yellow) by Spearman.

5.3 ROC-Curve of Burden

The aggregated score of the Mouse Grimace Scale provides the burden of the mouse. The ROC Curve (Receiver Operating Characteristic) is an effective tool to discriminate between two conditions. Figure 6 illustrates the Area Under the Curve (AUC) that measures the discriminatory power.

The graph of the experienced human raters show the best results (e.g. Experient 3, AUC = 0.93), the very large LLMs (e.g. GPT 5.1, AUC = 0.80) are similar to the inexperienced human raters (e.g. Inexperienced 1, AUC = 0.87), and the small local LLMs are slightly better (e.g. Gemma 27b, AUC = 0.54) than a guessing (AUC = 0.5). The ROC indicates that the complexity and size of the LLM models somehow correlate with their performance. However, this trend is not universal, as fine-tuned smaller models can outperform larger, untrained models.

6 Discussion

The presented results clearly show that experts in MGS assessment achieve scores closest to the ground truth. This is expected, as the ground truth dataset itself is based on expert evaluations. However, the inter-rater results also reveal that even experts perform inconsistently on certain action units.

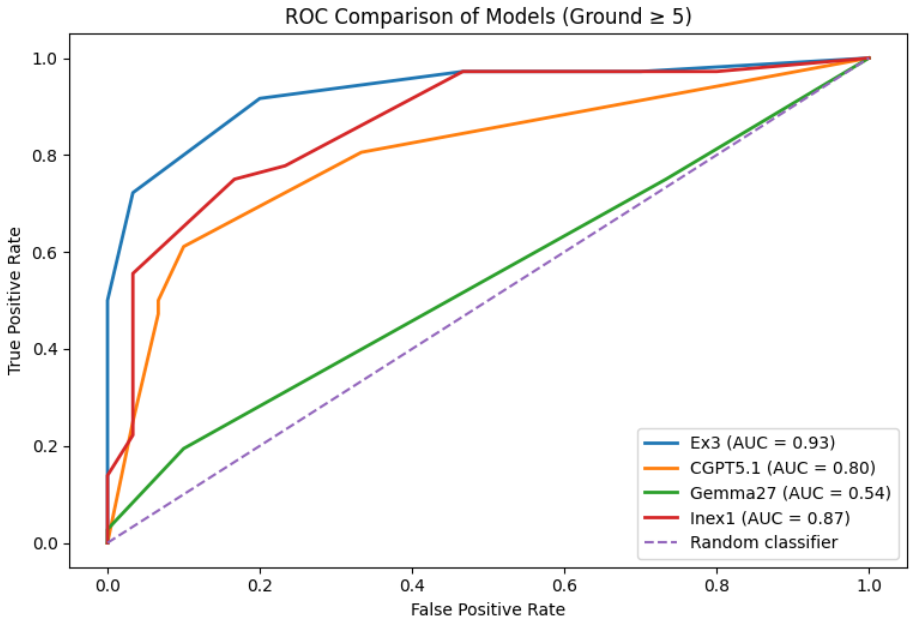

Fig. 6: ROC Curve of an experienced user (Experienced 3, Ex3, blue) in relation to a very large AI-model (ChatGPT-5.1, orange), an inexperienced user (Inex1, red), and a small local LLM (Gemma-27b, Gemma27, green). All raters are better than random (dotted line), the best performer (AUC=0.93) is human.

A key insight of this study is that not all action units contribute equally to reliable assessment. Features such as whisker position and cheek flattening exhibit high variability and low inter-rater agreement, whereas orbital tightening appears to be more robust. This finding challenges the implicit assumption of equal weighting in the traditional MGS formulation. We argue that future MGS systems should incorporate learned or empirically derived weighting schemes, potentially guided by AI-based analysis. Such an approach could transform the MGS from a manually defined scoring system into a data-driven, adaptive metric.

These findings also raise questions about the validity of the MGS itself, as scores may be influenced by additional factors such as age, sex, environmental conditions, time of measurement, or physiological state. Addressing such sources of bias requires reproducible and objective evaluation systems, which AI-based approaches can potentially provide.

Interestingly, AI-based classification (under low-temperature settings) is fully reproducible. This contrasts with human scoring, which is inherently subject to variability. While AI systems currently do not reach expert-level performance, the results indicate that they already approach the level of inexperienced human raters. This suggests that AI-based approaches could support or augment human evaluation. In addition, the findings indicate that the MGS itself may benefit from extensions, such as weighted parameters or alternative representations. Fully end-to-end approaches should also be considered in future work.

During our investigation, we observed that several existing MGS solutions from the literature could not be directly tested or compared within our experimental setup. In some cases, methods only provide binary classification results, while in others, the required input image quality or format is not compatible with our dataset. Furthermore, our results confirm that image and video quality have a significant impact on MGS classification. Some automated systems (e.g., PainFace) rely on high-quality, standardized input data, whereas real-world laboratory data are often of lower quality. Therefore, future MGS systems should incorporate reliability or confidence measures to account for varying input quality.

In our evaluation, all AI models were queried using the same prompt. Since both prompting strategies and temperature settings influence model outputs, further optimization is likely possible. This aspect should be investigated in future studies using fixed models and controlled parameter settings.

Another interesting aspect was the attempt to include artificially generated images of mice representing different states ranging from low to high distress. However, current AI systems often refuse to generate images of distressed animals for ethical reasons. Additionally, real-world datasets of distressed animals are limited for similar reasons. While simulation-based approaches could potentially address this limitation, the lack of balanced data remains a challenge. In our experiments, general-purpose AI models did not fully utilize the entire MGS scale, which likely contributed to higher absolute errors. Nevertheless, the observed correlations indicate that models capture the overall trend (low stress

corresponding to low MGS scores and vice versa). One possible improvement would be normalization or rescaling of the MGS.

The proposed AI-based MGS approach also has several limitations. The analyzed images were automatically extracted from video recordings, and it is likely that pain-related expressions occur intermittently rather than continuously. Therefore, future approaches should consider temporal aggregation across multiple frames. Another important limitation is the small dataset size and the limited number of individual animals. Although the controlled distribution enables systematic bias analysis, the generalizability of the results must be validated on larger and more diverse datasets.

The aim of this work was not to develop a specialized, fine-tuned model optimized for a specific dataset. Instead, our contribution lies in evaluating the performance of general-purpose AI models in comparison to human raters and in demonstrating their potential for automated welfare assessment. Such systems could enable continuous monitoring of animal condition and support timely intervention in laboratory and agricultural settings.

With respect to the predefined hypotheses, the results confirm that trained human experts still outperform both general-purpose AI systems and inexperienced human raters. However, the observed performance trends suggest that AI systems may approach expert-level performance in the future while offering significant advantages in terms of scalability and efficiency.

Finally, the proposed approach is not limited to mice but can be extended to other species. In particular, the combination of AI-based analysis with camera systems offers a scalable solution for improving animal welfare monitoring across different application domains.

7 Conclusion and Outlook

The estimation of the health, pain, stress, and emotional state of laboratory mice is essential for reliable experimental results. These parameters extend physical metrics such as body temperature or weight and are helpful for examining how mammals react to test settings. Because manual analysis of visual appearance is costly, time-consuming, and subjective, we propose an automated, AI-based Mouse Grimace Scale (MGS) assessment. We therefore investigate how accurately human experts in the field of MGS, inexperienced humans, and AI in the form of general-purpose large language models (LLMs) can estimate MGS scores. We present AniScale, a framework for AI-based MGS estimation to support and enhance this assessment, designed a test set of mouse images for objective evaluation, and performed various experiments with both small and high-performance LLMs available as of January 2026. This work demonstrates that general-purpose AI systems are not yet capable of replacing expert human raters in MGS assessment. However, they already achieve performance comparable to inexperienced humans while offering full reproducibility and scalability. More importantly, our findings suggest that AI should not only be viewed as a replacement for human scoring but as a tool to critically evaluate and improve existing welfare metrics.

In particular, the observed variability across action units indicates that the current MGS formulation may benefit from data-driven refinement. Furthermore, this work suggests a transition from manually defined welfare metrics toward data-driven, adaptive assessment frameworks supported by AI. Since MGS is currently performed in specialized observation boxes, future work will focus on continuous MGS assessment in home cages under normal housing conditions. Furthermore, we will extend our AniScale system and integrate trained AI models to achieve higher accuracy and performance. In addition, we will adapt our MGS-based estimation of animal condition to grimace scales for farm animals such as pigs, cows, and poultry in order to enhance their welfare in farming environments.

8 Acknowledgements

We thank University of Rostock, Germany, for providing animal images obtained from approved animal trials (Reg. No. 7221.3-1-035/20, LAAF, Rostock, Germany). We also thank the AI tool FhGenie [37] for linguistic support in improving the expression, spelling, and clarity of this work.
Furthermore, we thank Brigitte Vollmar for their support. This research was funded by KI-TIERWOHL (EXF-25-1031, EXF-25-1034) and by Fraunhofer Society, Germany.

References

1. AI, M.: Magistral small 1.2 (2025), https://lmstudio.ai/models/mistralai/magistral-small-2509, (accessed on 15th December 2025)
2. Andresen, N., Wöllhaf, M., Hohlbaum, K., Lewejohann, L., Hellwich, O., Thöne-Reineke, C., Belik, V.: Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. Plos one **15**(4), e0228059 (2020). https://doi.org/https://doi.org/10.1371/journal.pone.0228059
3. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968). https://doi.org/https://doi.org/10.1037/h0026256
4. Commission, E.: Summary report on the statistics on the use of animals for scientific purposes in the member states of the european union and norway in 2022 (2024), https://ec.europa.eu/transparency/documents-register/detail?ref=SWD(2024)185&lang=en, (accessed on 2nd December 2025)
5. Community, M.: Pixtral 12b (2025), https://huggingface.co/lmstudio-community/pixtral-12b-GGUF, (accessed on 15th December 2025)
6. Faller, K.M., McAndrew, D.J., Schneider, J.E., Lygate, C.A.: Refinement of analgesia following thoracotomy and experimental myocardial infarction using the mouse grimace scale. Experimental physiology **100**(2), 164–172 (2015). https://doi.org/https://doi.org/10.1113/expphysiol.2014.083139
7. Frederickson, R.C., Burgis, V., Edwards, J.D.: Hyperalgesia induced by naloxone follows diurnal rhythm in responsivity to painful stimuli. Science **198**(4318), 756–758 (1977). https://doi.org/https://doi.org/10.1126/science.561998

8. Gupta, S., Yamada, A., Ling, J., Gu, J.G.: Quantitative orbital tightening for pain assessment using machine learning with deeplabcut. Neurobiology of Pain **16**, 100164 (2024). https://doi.org/https://doi.org/10.1016/j.ynpai.2024.100164
9. Hohlbaum, K., Bert, B., Dietze, S., Palme, R., Fink, H., Thöne-Reineke, C.: Severity classification of repeated isoflurane anesthesia in c57bl/6jrj mice—assessing the degree of distress. PloS one **12**(6), e0179588 (2017). https://doi.org/https://doi.org/10.1371/journal.pone.0179588
10. Hohlbaum, K., Bert, B., Dietze, S., Palme, R., Fink, H., Thöne-Reineke, C.: Impact of repeated anesthesia with ketamine and xylazine on the well-being of c57bl/6jrj mice. PloS one **13**(9), e0203559 (2018). https://doi.org/https://doi.org/10.1371/journal.pone.0203559
11. Hohlbaum, K., Corte, G.M., Humpenöder, M., Merle, R., Thöne-Reineke, C.: Reliability of the mouse grimace scale in c57bl/6jrj mice. Animals **10**(9), 1648 (2020). https://doi.org/https://doi.org/10.3390/ani10091648
12. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019). https://doi.org/https://doi.org/10.48550/arXiv.1904.09751
13. Hurley, R.W., Adams, M.C.: Sex, gender, and pain: an overview of a complex field. Anesthesia & Analgesia **107**(1), 309–317 (2008). https://doi.org/https://doi.org/10.1213/01.ane.0b013e31816ba437
14. Jirkof, P., Abdelrahman, A., Bleich, A., Durst, M., Keubler, L., Potschka, H., Struve, B., Talbot, S.R., Vollmar, B., Zechner, D., et al.: A safe bet? inter-laboratory variability in behaviour-based severity assessment. Laboratory animals **54**(1), 73–82 (2020). https://doi.org/https://doi.org/10.1177/0023677219881481
15. for the Protection of Laboratory Animals (Bf3R), G.C.: Use of laboratory animals in 2023 (2025), https://www.bf3r.de/en/offers/test-animal-numbers/test-animal-numbers-2023/, (accessed on 2nd December 2025)
16. Langford, D.J., Bailey, A.L., Chanda, M.L., Clarke, S.E., Drummond, T.E., Echols, S., Glick, S., Ingrao, J., Klassen-Ross, T., LaCroix-Fralish, M.L., et al.: Coding of facial expressions of pain in the laboratory mouse. Nature methods **7**(6), 447–449 (2010). https://doi.org/https://doi.org/10.1038/nmeth.1455
17. Leach, M.C., Klaus, K., Miller, A.L., Scotto di Perrotolo, M., Sotocinal, S.G., Flecknell, P.A.: The assessment of post-vasectomy pain in mice using behaviour and the mouse grimace scale. PloS one **7**(4), e35656 (2012). https://doi.org/https://doi.org/10.1371/journal.pone.0035656
18. McCoy, E.S., Park, S.K., Patel, R.P., Ryan, D.F., Mullen, Z.J., Nesbitt, J.J., Lopez, J.E., Taylor-Blake, B., Vanden, K.A., Krantz, J.L., et al.: Development of painface software to simplify, standardize, and scale up mouse grimace analyses. Pain **165**(8), 1793–1805 (2024). https://doi.org/https://doi.org/10.1097/j.pain.0000000000003187
19. Mellor, D., Stafford, K.: Integrating practical, regulatory and ethical strategies for enhancing farm animal welfare. Australian veterinary journal **79**(11), 762–768 (2001). https://doi.org/https://doi.org/10.1111/j.1751-0813.2001.tb10895.x
20. Miller, A., Kitson, G., Skalkoyannis, B., Leach, M.: The effect of isoflurane anaesthesia and buprenorphine on the mouse grimace scale and behaviour in cba and dba/2 mice. Applied animal behaviour science **172**, 58–62 (2015). https://doi.org/https://doi.org/10.1016/j.applanim.2015.08.038

21. Miller, A.L., Leach, M.C.: The mouse grimace scale: a clinically useful tool? PloS one **10**(9), e0136000 (2015). https://doi.org/https://doi.org/10.1371/journal.pone.0136000
22. OpenAI: GPT-4o (snapshot: 2024-11-20) (2024), https://platform.openai.com/docs/models/gpt-4o?snapshot=gpt-4o-2024-11-20, (accessed on 15th December 2025)
23. OpenAI: GPT-4.1 mini (snapshot: 2025-04-14) (2025), https://platform.openai.com/docs/models/gpt-4.1-mini, (accessed on 15th December 2025)
24. OpenAI: GPT-5 mini (snapshot: 2025-08-07) (2025), https://platform.openai.com/docs/models/gpt-5-mini, (accessed on 15th December 2025)
25. OpenAI: GPT-5 (snapshot: 2025-08-07) (2025), https://platform.openai.com/docs/models/gpt-5, (accessed on 15th December 2025)
26. OpenAI: GPT-5.1 (snapshot: 2025-11-13) (2025), https://platform.openai.com/docs/models/gpt-5.1, (accessed on 15th December 2025)
27. OpenAI: GPT-5.1 (snapshot: 2025-12-11) (2025), https://platform.openai.com/docs/models/gpt-5.2, (accessed on 15th December 2025)
28. OpenAI: GPT-5.1 (snapshot: 2025-12-11) (2025), https://platform.openai.com/docs/models/gpt-5.2-pro, (accessed on 15th December 2025)
29. OpenBMB: Minicpm-o-2_6 (2025), https://huggingface.co/openbmb/MiniCPM-o-2_6-gguf, (accessed on 15th December 2025)
30. OpenGVLab, Bartowski: InternVL3_5-14B (2025), https://huggingface.co/bartowski/OpenGVLab_InternVL3_5-14B-GGUF, (accessed on 15th December 2025)
31. Sotocina, S.G., Sorge, R.E., Zaloum, A., Tuttle, A.H., Martin, L.J., Wieskopf, J.S., Mapplebeck, J.C., Wei, P., Zhan, S., Zhang, S., et al.: The rat grimace scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. Molecular pain **7**, 1744–8069 (2011). https://doi.org/https://doi.org/10.1186/1744-8069-7-55
32. Spearman, C.: The proof and measurement of association between two things. (1961). https://doi.org/https://doi.org/10.2307/1412159
33. Sturman, O., Schmutz, M., Lorimer, T., Zhang, R., Privitera, M., Roessler, F.K., Leonardi, J., Waag, R., Marinescu, A.M., Bekemeier, C., et al.: Grimace: Automated, multimodal cage-side assessment of pain and well-being in mice. bioRxiv pp. 2025–03 (2025). https://doi.org/https://doi.org/10.1101/2025.03.07.642046
34. Team, G.: Gemma 3 (2025), https://goo.gle/Gemma3Report, (accessed on 15th December 2025)
35. Team, Q.: Qwen2.5-vl (2025), https://qwenlm.github.io/blog/qwen2.5-vl/, (accessed on 15th December 2025)
36. Tuttle, A.H., Molinaro, M.J., Jethwa, J.F., Sotocinal, S.G., Prieto, J.C., Styner, M.A., Mogil, J.S., Zylka, M.J.: A deep neural network to assess spontaneous pain from mouse facial expressions. Molecular pain **14**, 1744806918763658 (2018). https://doi.org/https://doi.org/10.1177/1744806918763658
37. Weber, I., Linka, H., Mertens, D., Muryshkin, T., Opgenoorth, H., Langer, S.: Fhgenie: a custom, confidentiality-preserving chat ai for corporate and scientific use. In: 2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C). pp. 26–31. IEEE (2024). https://doi.org/10.1109/icsa-c63560.2024.00011
38. Whittaker, A.L., Liu, Y., Barker, T.H.: Methods used and application of the mouse grimace scale in biomedical research 10 years on: a scoping review. Animals **11**(3), 673 (2021). https://doi.org/https://doi.org/10.3390/ani11030673

39. xAI: Grok 4 (2025), `https://grok.com/`, (accessed via webui troughout December 2025)