

LLMs stabilize Deauville scoring through longitudinal reasoning in lymphoma

Wei-Fang Zhang

tsy1997@126.com

Department of Nuclear Medicine, Peking University Third Hospital

Meixin Zhao

Peking University

Mingyang Yu

Shanghaitech University

Yuan Li

Department of Nuclear Medicine, Peking University People's Hospital

Han Li

Department of Nuclear Medicine, Beijing Chaoyang Hospital, Capital Medical University, 8th Gongtinanlu Road, Chaoyang District, Beijing 100020, China.

Xiaoyan Hou

Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191

Annan Zhang

Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191

Shijia Weng

Department of Nuclear Medicine, Peking University People's Hospital, Beijing 100044

Meng Wang

Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191

Na Guo

Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191

Anhui Zhu

Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191

Yiran Shu

United Imaging Intelligence Co., Ltd., Shanghai, China.

Jingyu Liu

United Imaging Intelligence Co., Ltd., Shanghai, China.

Ying Shao

Shanghai United Imaging Intelligence Co. Ltd.

Yanbo Chen

Shanghai United Imaging Intelligence Co. Ltd.

Min-Fu Yang

Beijing Chaoyang Hospital, Capital Medical University

Yaozong Gao

Shanghai United Imaging Intelligence Co. Ltd. <https://orcid.org/0000-0002-7547-5209>

Dinggang Shen

ShanghaiTech University <https://orcid.org/0000-0002-7934-5698>

Article

Keywords:

Posted Date: June 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9934993/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. Some authors are affiliated with Shanghai United Imaging Intelligence Co., Ltd. The remaining authors declare no competing interests.

LLMs stabilize Deauville scoring through longitudinal reasoning in lymphoma

Meixin Zhao^{1†}, Mingyang Yu^{2†}, Yuan Li^{3†}, Han Li⁴, Xiaoyan Hou¹,
Annan Zhang¹, Shijia Weng¹, Meng Wang¹, Na Guo¹, Anhui Zhu¹,
Jingyu Liu⁵, Yiran Shu⁵, Ying Shao⁵, Yanbo Chen⁵, Minfu Yang⁴,
Yaozong Gao^{5*}, Dinggang Shen^{2,5,6*}, Weifang Zhang^{1*}

As one of the most common cancers worldwide, lymphoma requires reliable tools to assess treatment response. The Deauville score (DS) is a widely used PET/CT-based standard for this purpose and plays a central role in clinical decision-making and research. In many regions, DS is embedded in clinical workflows and directly influences treatment adjustment. However, DS assessment remains reader-dependent, as it requires complex longitudinal interpretation of prior and current PET/CT findings, leading to limited interobserver agreement and challenges for large-scale application. Here, we introduce a large language model (LLM)-based framework that leverages longitudinal PET/CT reports to

perform structured, stepwise reasoning for automated Deauville scoring. In a multicentre dataset of 6,009 lymphoma cases from three hospitals, the model achieved 85.8% accuracy for six-class classification across internal and external cohorts, outperforming human readers and state-of-the-art baselines, with the greatest advantage observed in cases with low inter-reader agreement. These findings suggest that LLM-based approaches can provide a practical strategy for addressing complex imaging-derived clinical assessment tasks through structured semantic reasoning, enabling more standardized and scalable lymphoma response assessment.

1 Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191, China. 2 School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China. 3 Department of Nuclear Medicine, Peking University People’s Hospital, Beijing 100044, China. 4 Department of Nuclear Medicine, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China. 5 Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China. 6 Shanghai Clinical Research and Trial Centre, Shanghai 201210, China. Correspondence to: Weifang Zhang, Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191, China (tsy1997@126.com); Yaozong Gao (yaozong.gao@uii-ai.com); and Dinggang Shen (Dinggang.Shen@gmail.com). † These authors contributed equally.

1. Introduction

The Deauville score (DS) [16] was developed for imaging-based response assessment on interim FDG PET in lymphoma and has since become a widely used criterion for treatment response assessment on FDG PET/CT, supporting therapeutic decision-making in clinical practice [6, 17]. This five-point visual scale evaluates treatment response on PET/CT by comparing lesion radiotracer uptake with reference tissues (mediastinal blood pool and liver), focusing on residual lymphoma-related disease sites. In this system, scores of 1–3 are generally considered consistent with complete metabolic response, whereas scores of 4–5 indicate residual metabolically active disease. Category X denotes newly identified uptake patterns that are likely unrelated to lymphoma and thus cannot be directly accommodated within the standard scoring scale [18]. Although the criteria are conceptually simple, consistent Deauville scoring in real-world practice remains difficult. The detailed scoring definitions used in this study are provided in Supplementary Note 1.

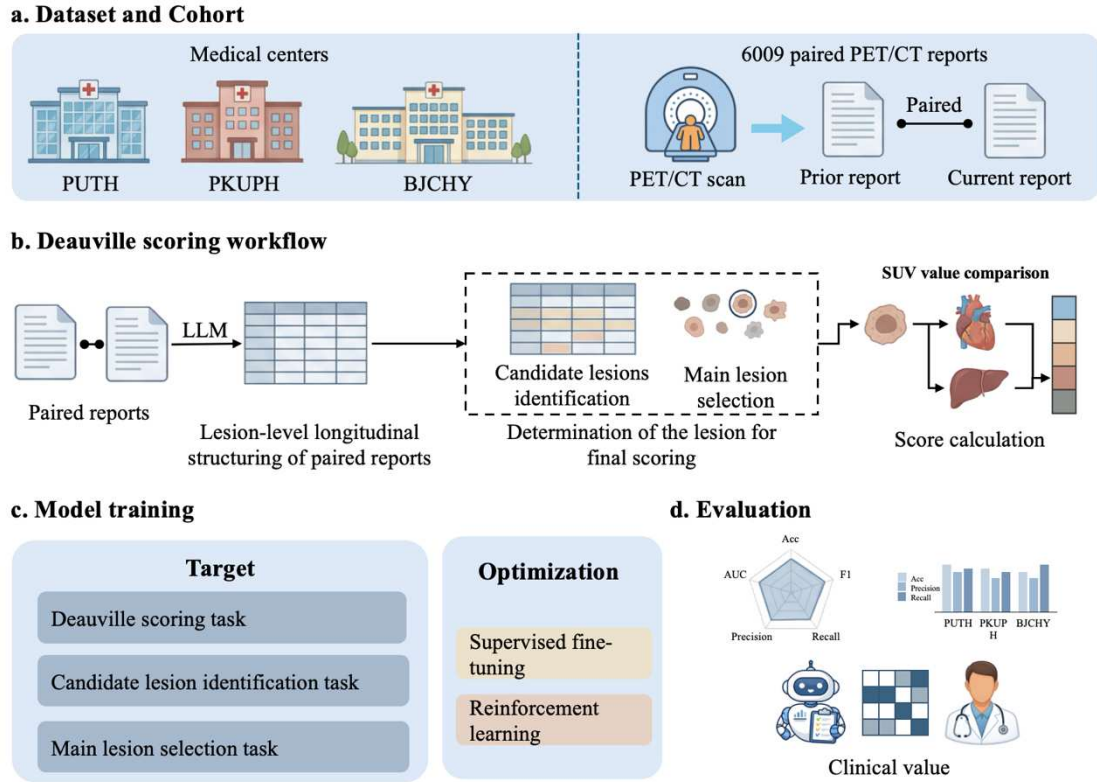


Figure 1: Overview of dataset construction, Deauville scoring workflow, model training, and evaluation. **a**, Dataset and cohort. A multicentre dataset comprising 6,009 paired PET/CT reports was collected from three medical centres. Each sample contained a prior report and a current report. **b**, Deauville scoring workflow. Paired reports were longitudinally structured at the lesion level using an LLM to support candidate lesion identification and main lesion selection, after which the Deauville score was determined by comparison with reference uptake. **c**, Model training. The model was trained with three tasks: Deauville scoring, candidate lesion identification, and main lesion selection, using supervised fine-tuning and reinforcement learning. **d**, Evaluation. Performance was evaluated using quantitative metrics, centre-wise comparisons, and clinical value analysis.

Importantly, the main challenge is not numerical uptake comparison alone. In routine follow-up assessment, readers must first determine whether FDG-avid findings are likely lymphoma-related rather than physiological or treatment-related inflammatory uptake, then decide which residual or newly detected lesions should be included for scoring, and finally select the lesion that should determine the final score in the context of prior examinations. This combined process of uptake attribution, lesion inclusion, and representative-lesion selection is especially difficult in multifocal disease and is a major source of interobserver disagreement [1, 5, 21]. Dual reading can partially mitigate this variability, but it increases workload and does not fully standardize the reasoning process.

Most prior studies on automated Deauville assessment have focused on image-based analysis [14]. These approaches typically depend on lesion segmentation, lesion registration across serial scans, lesion classification, and end-to-end score prediction from PET/CT volumes [4, 8–10, 22, 23]. However, such methods face several practical limitations in longitudinal lymphoma imaging: lesions may disappear, merge, shift, or newly emerge over time; physiological and inflammatory FDG uptake is common; and most published systems prioritize binary classification over multi-class scoring, despite the latter being more clinically interpretable and more informative for subsequent clinical decision making [25, 26]. As a result, existing methods have had limited ability to explicitly standardize the reasoning steps that drive disagreement in real-world Deauville scoring.

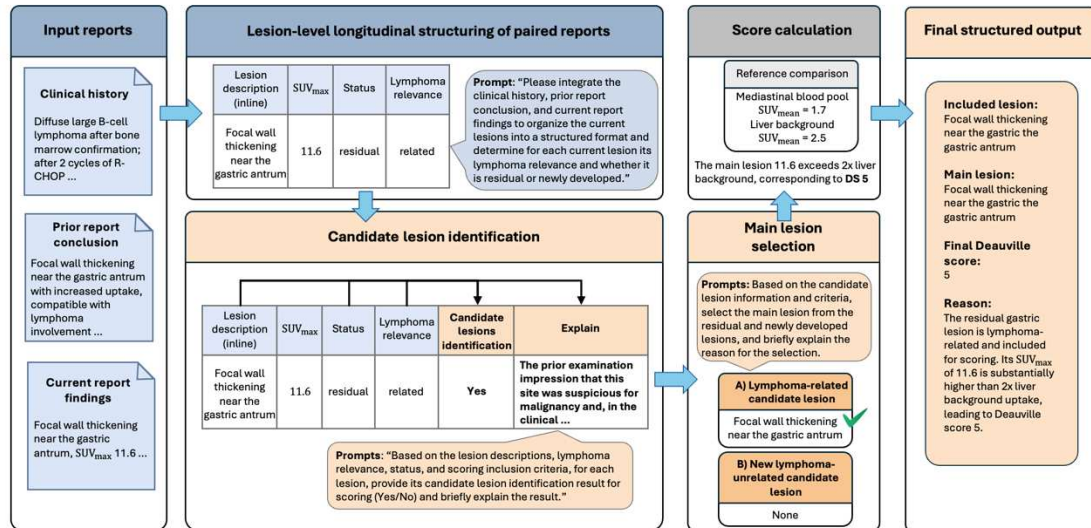


Figure 2: Deauville scoring workflow. For each case, paired prior and current reports are first organized into lesion-level structured descriptions, including lesion description (inline), uptake value (e.g., SUV_{max}), temporal status (residual or newly developed), and lymphoma relevance (yes or no). Based on predefined scoring inclusion criteria, each lesion is evaluated for candidate lesion identification, indicating whether it should be considered for scoring. Among the eligible candidate lesions, the main lesion for final scoring is selected by comparing the highest-SUV_{max} residual lymphoma-related lesion with any eligible newly developed lesion. The Deauville score is then determined by comparing the uptake of the selected main lesion with reference regions (mediastinum and liver).

To address this limitation, we reformulated Deauville scoring as a longitudinal reasoning task over paired follow-up PET/CT reports. This design is clinically motivated. In daily practice, follow-up interpretation is rarely performed from a single examination in isolation. Instead, clinicians compare the current study with previous

reports and revisit raw images when needed. PET/CT reports are therefore not merely textual surrogates of the scans, but condensed, expert-validated summaries that already encode clinically relevant lesion-level information. When converted into structured representations, they provide a workflow-aligned substrate for modelling uptake attribution, lesion selection, and rule-based score assignment.

Nevertheless, reliable report-based Deauville scoring remains challenging. Pretrained LLMs lack task-specific knowledge for multifocal lymphoma assessment, may make errors in numerical comparison with reference tissues, and can deviate from the ordered workflow required in clinical scoring. We therefore developed a structured, stepwise framework that decomposes the task into four sequential steps, as illustrated in Fig. 2: (1) lesion-level longitudinal structuring of paired follow-up reports, which organizes the two reports into aligned lesion-level representations; (2) candidate lesion identification, which screens all reported lesions and retains only those potentially eligible for Deauville scoring; (3) main lesion selection, which determines the lesion that should represent the patient for final scoring; and (4) final score reasoning, which assigns the Deauville score to the selected lesion under the Deauville criteria and outputs a structured result including both the score and its explanation. Detailed prompting protocol, reasoning steps, full prompt templates, and representative case examples spanning the original reports, structured inputs, and model-generated reasoning outputs are provided in Supplementary Notes 2–6.

Using longitudinal PET/CT report pairs from three hospitals, we evaluated the framework in 2-class, 4-class and 6-class settings. The method achieved agreement

comparable to expert interobserver agreement and consistently outperformed existing baselines. More importantly, its clinical value was most evident in the cases that are most difficult and disagreement-prone, supporting the role of LLM-based, workflow-aligned reasoning as a practical strategy for stabilizing Deauville scoring in real-world lymphoma care. Representative examples of longitudinal PET/CT reports, structured inputs, and model-generated reasoning outputs are provided in the Supplementary Information (Supplementary Note 4-6).

2 Results

2.1 Study Design

The overall study design is summarized in Fig. 1. We first constructed a multicentre cohort of 6,009 longitudinal 18F-FDG PET/CT report pairs from three hospitals, including Centre PUTH (Peking University Third Hospital; n=3,908), Centre PKUPH (Peking University People's Hospital; n=1,813), and Centre BJCYH (Beijing Chao-Yang Hospital; n=288). Each pair linked at least two consecutive whole-body PET/CT examinations from the same patient and was associated with a documented Deauville score. The three centres exhibited clear dataset-level differences (Fig. 3a, b), including differences in lesion burden and score distribution, thereby forming a realistic multicentre domain-shift setting for model development and evaluation.

For model development and evaluation, we adopted a centre-aware split. Cases from Centres PUTH and PKUPH were used for training and internal testing, whereas Centre

BJCYH was reserved exclusively for external evaluation. Specifically, 600 cases from Centre PUTH and 300 cases from Centre PKUPH were randomly sampled as the internal test set (n=900), and the remaining 4,821 cases were used for training. The 288 cases from Centre BJCYH served as an independent external test cohort. This design enabled evaluation of both in-domain performance and cross-centre generalizability. Representative case examples, including the original PET/CT reports, structured inputs, and model-generated reasoning outputs, are provided in the Supplementary Information (Supplementary Notes 4–6).

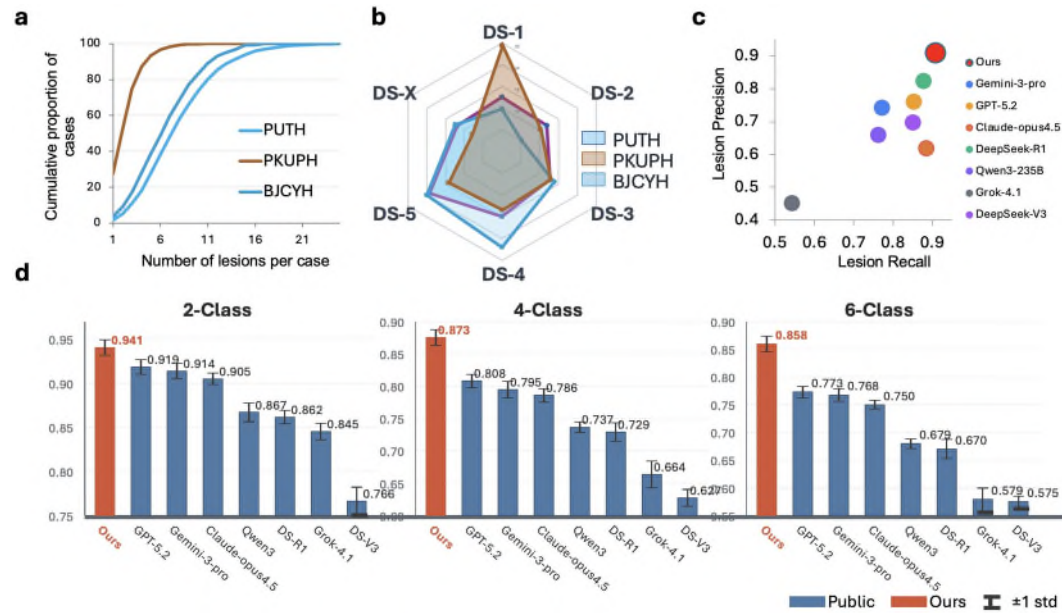


Figure 3: Cross-centre data heterogeneity and model performance in Deauville scoring.

a, Empirical cumulative distribution functions of the number of recorded lesions per case in the three centres. PKUPH is shifted toward lower lesion counts, whereas PUTH and BJCYH show broader distributions with a larger proportion of cases containing multiple lesions. **b**, Radar plot of the relative composition of Deauville score categories across centres, showing centre-specific class distributions and confirming dataset-level

heterogeneity in score prevalence. **c**, Comparison of recall and precision for candidate lesion identification indicated that the model’s score predictions were grounded in clinically relevant lesion evidence rather than in score estimation alone. **d**, Comparison of overall accuracy under 2-class, 4-class, and 6-class Deauville scoring settings. Our model shows the highest performance across all settings.

Within this framework, we evaluated the proposed method from four aspects. First, we compared our method with representative public models under the same test setting to assess overall Deauville scoring performance and lesion-level evidence quality. Second, we examined robustness across both centres and Deauville score categories to determine whether model performance remained stable under institutional heterogeneity and class-dependent variation. Third, we performed ablation analysis on the pooled cohort to assess the contribution of candidate lesion identification, main lesion selection, supervised multi-task training, and reinforcement learning. Finally, we compared the scoring behaviour of our model with that of human readers and further evaluated model performance in difficult cases with low inter-observer agreement.

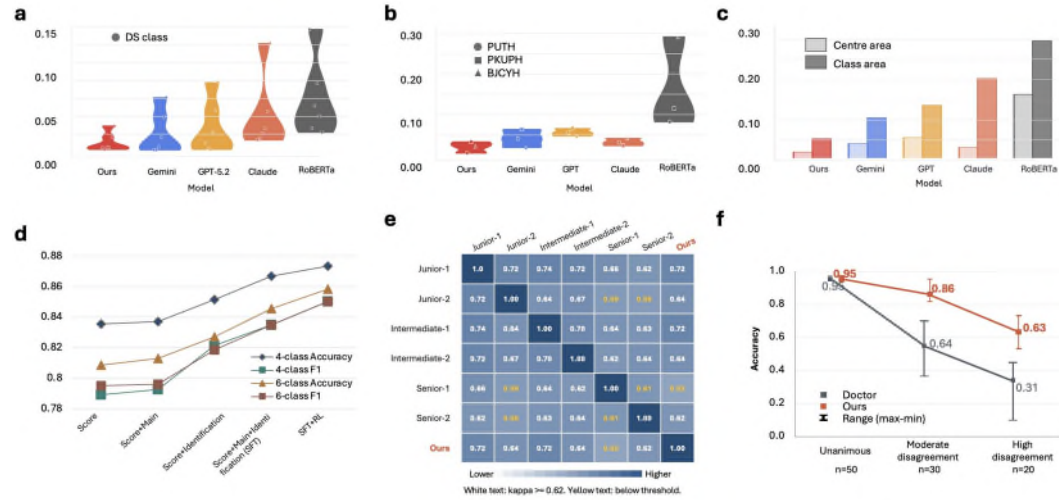


Figure 4: Evaluation of robustness, ablation effects, and comparison with human readers. **a**, Violin plots of class-wise robustness, measured as the normalized absolute deviation of class-wise F1 from macro-F1 across the six Deauville score classes. **b**, Violin plots of cross-centre robustness, measured as the normalized absolute deviation of centre-specific accuracy from overall accuracy across PUTH, PKUPH, and BJCYH. **c**, Summary comparison of class-level and centre-level robustness, shown as the mean deviation from a and b, respectively, where lower values indicate greater robustness. **d**, Ablation study on the pooled cohort, showing progressive gains in 4-class and 6-class accuracy and F1. **e**, Agreement between physicians and our method, showing that model-reader agreement falls within the human-human range. **f**, Accuracy stratified by physician disagreement level, showing that our method is more accurate in difficult cases with greater inter-reader disagreement.

2.2 Comparison with Public Large Language Models

We first compared our method with representative public large language models under task-specific prompting, including explicit Deauville criteria and expert interpretive guidance. As shown in Fig. 3d, our method consistently outperformed all evaluated public models, reaching an overall 0.941 accuracy in binary-class scoring, 0.873 in four-class scoring and 0.858 in six-class, indicating a consistent performance advantage over all evaluated public models. However, comparison based on final score accuracy alone cannot fully reflect model capability, because evaluation using only the final score does not reveal whether the model truly identified the clinically relevant lesions for scoring or merely fit the statistical pattern of score labels.

To further understand this performance gap, we compared lesion-level evidence quality across different models using lesion recall and lesion precision for candidate lesion identification from all lesions in the current reports. These metrics measured whether a model could correctly identify the lesions eligible for Deauville scoring, rather than relying only on the final score prediction. As shown in Fig. 3c, our method achieved both high lesion recall (0.9084) and high lesion precision (0.9089), outperforming all public baselines and placing it in the upper-right region of the comparison space. This suggests that the superior performance of our method arose not simply from predicting the final label, but from more accurate identification of clinically relevant lesion evidence for scoring.

2.3 Robustness across centres and score categories

To evaluate robustness, we compared our method with representative baselines from two complementary perspectives: robustness to cross-Centre heterogeneity and robustness across Deauville score categories. All models were evaluated on the same multi-centre test split. For centre-level robustness, we quantified the normalized absolute deviation of centre-specific accuracy from each model’s mean accuracy across PUTH, PKUPH, and BJCYH. For class-level robustness, we quantified the normalized absolute deviation of class-wise F1 from each model’s own macro-F1 across the six Deauville score classes. Lower deviation indicates that model performance remains closer to its overall level and is therefore more robust to centre shift or category-specific degradation.

As shown in Fig. 3a, our method exhibited the most compact class-wise robustness distribution, indicating stronger robustness across Deauville score categories and less sensitivity to class-specific performance collapse. Public LLM baselines showed broader distributions, suggesting weaker robustness across score classes, whereas the RoBERTa-based [27] classifier displayed the largest spread overall. Similarly, Fig. 3b shows that our method had the tightest cross-centre robustness distribution, supporting stronger robustness under institutional variation. In contrast, public baselines showed weaker cross-centre robustness, and RoBERTa exhibited the greatest instability across centres.

To summarize these two robustness experiments, we further computed the mean deviation for both centre-level and class-level distributions (Fig. 3c). Our method achieved the smallest deviation in both settings, demonstrating that its advantage was

not limited to higher average performance, but also extended to more robust behavior across heterogeneous centres and uneven score categories. Together, these results indicate that our framework is robust not only to multi-centre distribution shift, but also to class-dependent degradation in Deauville scoring.

2.4 Ablation Study

We next performed an ablation study to examine the effectiveness of our training strategy (Fig. 4d). The baseline model was trained only on the Deauville scoring task (Score). We then progressively introduced two additional supervised tasks: candidate lesion identification (Identification) and main lesion selection (Main). In this formulation, supervised fine-tuning (SFT) jointly optimized all three tasks, corresponding to Score+Main+Identification. Finally, reinforcement learning was applied only to the Deauville scoring task, using the final score as the reward signal, yielding the full model denoted as SFT+RL.

As shown in Fig. 4d, performance improved progressively from Score to Score+Main, Score+Identification, SFT, and finally SFT+RL, across both 4-class and 6-class settings. This trend indicates that the gain of our framework was not driven by a single component, but by the effectiveness of the overall multi-task training strategy. In particular, adding candidate lesion identification and main lesion selection both contributed clear improvements over the baseline scoring-only model, while joint SFT over the three tasks further strengthened performance. Reinforcement learning on the

final score provided an additional improvement on top of SFT, leading to the strongest overall results.

These findings are clinically meaningful. Deauville scoring depends not only on assigning the final score itself, but also on correctly identifying candidate lesions and selecting the lesion that should determine the final judgement. By explicitly introducing candidate lesion identification and main lesion selection into training, the model became better aligned with the actual clinical decision process. The additional gain from reinforcement learning suggests that optimizing the final scoring objective further improves consistency between the model’s intermediate decision process and its final Deauville score output.

2.5 Performance in challenging cases with low inter-observer agreement

Because limited inter-observer agreement is a major challenge in Deauville scoring, we further compared the scoring behavior of our model with that of human readers and examined model performance under different levels of human agreement. In this experiment, six doctors independently scored 100 cases. We first quantified pairwise agreement among all readers using Cohen’s kappa and summarized the results in a pairwise agreement matrix (Fig. 4e), where each cell represents the agreement between two readers and the comparisons between our model and individual doctors are shown alongside the human-human comparisons.

As shown in Fig. 4e, human-human agreement was only moderate overall, confirming limited interobserver agreement in Deauville scoring even among experienced readers. Importantly, the agreement between our model and individual doctors consistently fell within the range of human agreement, indicating that the scoring pattern of our model was comparable to that of human readers rather than behaving as an outlier. In other words, the model produced human-like scoring behavior in terms of its consistency with individual readers.

We next asked whether the model's advantage became more evident in clinically difficult cases with low inter-observer agreement. For each case, we defined the agreement level according to the largest number of doctors assigning the same Deauville score (top-n) among the six readers. Cases were categorized as unanimous when all six doctors assigned the same score (top-n = 6, n = 50), as moderate-disagreement when the largest voting group contained four or five doctors (top-n = 4 or 5, n = 30), and as high-disagreement when no score received support from more than three doctors (top-n $\leq$ 3, n = 20). We then compared the accuracy of our model with human performance in each subset (Fig. 4f).

When human agreement was high, our model performed similarly to human readers, with both our method and human majority voting achieving an accuracy of 0.95 in the unanimous subset. However, the advantage of our model became increasingly apparent as inter-observer agreement decreased. In the moderate-disagreement subset, our model achieved an accuracy of 0.86, compared with 0.64 for majority voting and a doctor-level range of 0.37–0.70. In the high-disagreement subset, where clinical interpretation

was most challenging, our model still achieved an accuracy of 0.65, whereas majority voting reached only 0.31 and individual doctor accuracy ranged from 0.10 to 0.45. These findings suggest that the value of our method lies not only in producing human-like scoring behaviour, but also in providing more reliable decisions in cases where inter-observer agreement is weakest.

3. Discussion

This study demonstrates that an LLM-based structured, stepwise reasoning framework can reduce interpretive uncertainty in Deauville scoring by standardizing the sequence of decisions that clinicians actually make during follow-up assessment. Rather than treating Deauville scoring as a direct label prediction problem, the framework explicitly models lesion analysis, lesion selection, and final score reasoning. This design yielded not only strong overall accuracy but also better performance in cases with high inter-reader disagreement, suggesting that the main clinical value of the approach lies in the very scenarios where conventional interpretation is most variable and resource-intensive.

The Deauville scale is a semi-quantitative system based on visual comparison and is not intended for definitive interpretation from a single examination alone. In real-world practice, a reader must decide whether observed FDG uptake likely reflects residual lymphoma or instead physiological, inflammatory, or treatment-related activity; determine which findings should be considered eligible for scoring; and identify the lesion that best represents overall disease status. These decisions are inherently

longitudinal, because the meaning of a current abnormality often depends on its relationship to prior disease sites and prior treatment response. This helps explain why Deauville scoring, despite apparently simple criteria, remains subject to substantial reader variability [5, 7, 17, 21]. Reflecting this workflow, our framework uses clinical history, prior report findings and impressions, and current report findings as input, while deliberately excluding current impressions to avoid direct label leakage. This preserves the causal sequence of clinical interpretation and allows the score to be inferred from upstream evidence rather than copied from the contemporaneous report. In this sense, the model is better understood as a structured second-reader or standardization tool than as a black-box label generator.

From a methodological perspective, our findings contribute to the broader discussion on the reproducibility of FDG-PET-based response assessment. Previous international benchmarking studies have shown that even quantitative PET metrics, such as total metabolic tumor volume (TMTV), can be affected by interpretive choices, including how diffuse splenic or bone marrow uptake is handled [1, 3, 18, 19, 24]. These observations highlight that reproducibility in PET response assessment depends not only on measurement standardization, but also on the consistency of upstream interpretive decisions. In this context, our framework provides a complementary approach by explicitly constraining lesion attribution, lesion inclusion, and main-lesion selection through longitudinal, rule-based reasoning. This may help reduce systematic variability in Deauville scoring, consistent with prior multicentre reports of only moderate inter-observer agreement [11, 19, 20].

Compared with automated approaches that depend on strict lesion matching or end-to-end optimization from image input to final score, the present framework is more compatible with longitudinal lymphoma assessment. One-to-one lesion correspondence is often unstable, as lesions may resolve, merge, shift in position, or newly emerge over time, and these challenges are amplified across centres by variation in reporting style and local interpretive habits [10, 14]. By operating on structured lesion representations and explicit reasoning steps, our approach improves auditability, permits error localization to specific stages, and better aligns with how follow-up PET/CT is interpreted in routine nuclear medicine practice. Although developed here for Deauville scoring, the underlying principle may be extendable to other categories within the Lugano classification [6] and to PET-based response assessment in settings such as immunotherapy [2, 12, 13, 15].

The observation that model benefit increased as human disagreement increased is particularly important from a clinical deployment perspective. Straightforward cases rarely drive workload or management uncertainty; by contrast, equivocal cases often trigger dual reading, senior review, additional imaging, or delayed treatment decisions. A system that mainly improves performance in this subset could therefore serve as a triage aid, quality-control layer, or decision-support tool in multidisciplinary lymphoma care.

Several limitations should be acknowledged. First, the present study validates primarily the reasoning stage of the overall pipeline and assumes reliable availability of high-

quality structured PET/CT report elements; automated generation of these structured descriptions from imaging data remains an essential component for full workflow integration. Second, because the system is report-based, its performance depends on the completeness and quality of the original report and it cannot replace direct image review when report descriptions are insufficient or potentially inaccurate. Third, we evaluated agreement with expert-derived labels rather than long-term clinical outcomes, so the prognostic value of model-derived scores remains unknown and requires prospective validation. Fourth, although the results in high-disagreement cases are encouraging, a prospective reader-assistance study is still needed to determine whether the system can reduce reporting time, improve inter-reader agreement, or change downstream management in routine practice. In summary, longitudinal semantic reasoning over paired PET/CT reports appears to be a clinically meaningful and scalable strategy for stabilizing Deauville scoring, and it provides a practical foundation for deployable PET/CT response-assessment tools that assist, rather than replace, expert lymphoma interpretation.

4 Methods

4.1 Ethics Statement

This retrospective multicentre study was approved by the institutional review board of Peking University Third Hospital (IRB00006761-M20260170), with authorization for data use obtained from the participating hospitals. All PET/CT reports were fully de-identified prior to analysis. Given the retrospective nature of the study and the use of

anonymized textual data, the requirement for informed consent was waived in accordance with institutional and national guidelines.

4.2 Dataset and annotation

We conducted a retrospective multicentre study using longitudinal 18F-FDG PET/CT reports from lymphoma patients collected at three hospitals: Peking University Third Hospital (PUTH), Peking University People’s Hospital (PKUPH), and Beijing Chao-Yang Hospital (BJCHY). As described above, after data cleaning and harmonization, 6,009 report pairs were retained for analysis. To reduce inter-centre inconsistency, particularly in cases involving category X, lesions were rechecked under a unified scoring standard and incorrect labels were corrected where necessary. Cases from PUTH and PKUPH were used for model development and internal evaluation, whereas cases from BJCHY were reserved exclusively for external testing.

To support structured longitudinal decision-making, the original reports were first converted into lesion-level structured representations. The Qwen3-235B-A22B large language model was used to process the clinical history, the current imaging findings, and the conclusion of the prior examination, while the conclusion of the current examination was deliberately excluded to prevent leakage of the reported Deauville score. As illustrated in Fig. 2, the model was prompted to integrate the clinical history, prior report conclusion, and current report findings, organize the current lesions into a structured format, and determine for each lesion its lymphoma relevance and whether it was residual or newly developed. The resulting structured table recorded lesion

description, SUVmax, temporal status, and lymphoma relevance. Hepatic and mediastinal blood-pool uptake values were also extracted as reference signals for subsequent score calculation. This structured representation, together with the clinical history, was then used as the input to the Deauville scoring model.

The output of the scoring model was designed to include not only a final Deauville score but also explicit intermediate decision steps. Because these intermediate annotations could not be manually established at scale, we generated initial drafts using a hindsight-guided reverse-reasoning prompting framework and then refined them through iterative physician–model review until a stable supervision standard was established.

The final annotation framework was organized into three clinically motivated tasks:

- 1) Candidate lesion identification: Based on lesion description, lymphoma relevance, temporal status (Residual / Newly detected), and scoring inclusion criteria, each lesion was assigned a candidate lesion identification result (Yes/No), which indicated whether the lesion should be included for Deauville scoring, together with a brief explanation. This step identified all lesions potentially eligible for subsequent Deauville scoring and enabled explicit supervision of lesion inclusion decisions.
- 2) Main lesion selection: Based on the candidate lesion information and the Deauville scoring criteria, the model selected the main lesion from the residual and newly developed lesions and provided a brief explanation for the selection. This step reflected the clinical requirement that the final score should be determined by the lesion most representative of overall disease status.

3) Score calculation: The selected main lesion was compared against the mediastinal blood-pool and liver background reference uptake according to the Deauville criteria. For example, when the main lesion SUVmax exceeded twice the liver background uptake, the case was assigned DS 5. This step produced the final case-level Deauville score together with an explicit explanation of the decision.

The final structured output for each case therefore contained four components: the included lesion set used for scoring, the selected main lesion, the final Deauville score, and a textual explanation describing why the selected lesion led to that score under the reference comparison. This output format preserved the full decision path from lesion inclusion to final score assignment, thereby enabling supervision of both intermediate tasks and the final scoring outcome.

This annotation strategy was designed to align supervision with the real-world workflow of lymphoma response assessment and to support an LLM-based structured decision framework, rather than a direct black-box classifier.

4.3 Experimental Setup

All experiments were conducted on a Linux-based server equipped with dual Intel Xeon CPUs and eight NVIDIA H20 GPUs, each with 144 GB of memory. Training was implemented using the open-source LLaMA-Factory framework. The supervised training stage (SFT) was initialized from Qwen3-4B and performed using LoRA fine-tuning. LoRA adapters were applied to the attention and feed-forward projection layers

(q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj). The SFT model was trained with a maximum sequence length of 5,120 tokens, a per-device batch size of 1, gradient accumulation of 16 steps, a learning rate of 6×10^{-4} , cosine learning-rate scheduling, and 6 epochs of training. A warm-up ratio of 0.05 was used, with 100 warm-up steps, and training was performed in bfloat16 precision. Validation was performed in a stepwise manner using a held-out split of 0.5% of the training data.

Reinforcement learning was then performed using DAPO, initialized from the SFT model. The DAPO stage used a GRPO-style advantage estimator. The maximum prompt length and response length were set to 4,096 and 8,192 tokens, respectively. Training used a prompt batch size of 256, a mini-batch size of 32, and 16 sampled responses per prompt. The actor model was optimized with a learning rate of 5×10^{-6} , weight decay of 0.1, gradient clipping of 1.0, and token-mean loss aggregation. Training was run for 10 epochs. During rollout, sequence-parallel training and tensor-parallel generation were used to support long-context optimization on multi-GPU hardware. The reward was computed from the final scoring outcome, so that reinforcement learning directly optimized the consistency between reasoning process and final Deauville score assignment.

4.4 Supervised Fine-tuning

We initialized our model from Qwen3-4B and first performed supervised fine-tuning (SFT) on structured instruction–response pairs derived from the annotated report representations. Each target response contained both the explicit stepwise reasoning

process and the final Deauville score, so that supervision was applied not only to the output label but also to the intermediate reasoning process. In this way, the supervised stage was intended to teach the model domain-specific scoring knowledge, expert interpretive heuristics, and the structured reasoning trajectory required for Deauville assessment. Model parameters were optimized by minimizing the negative log-likelihood of the target response:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^N \sum_{t=1}^{|y^{(i)}|} \log \pi_{\theta}(y_t^{(i)} | x^{(i)}, y_{<t}^{(i)}), \quad (1)$$

where $x^{(i)}$ denotes the structured input for case i , $y^{(i)}$ denotes the corresponding target response, and π_{θ} is the autoregressive policy parameterized by θ . To enable parameter-efficient adaptation, we adopted Low-Rank Adaptation (LoRA) for a subset of transformer weight matrices. For each adapted weight matrix, the pretrained parameter W_0 was reparametrized as

$$W = W_0 + \Delta W, \quad \Delta W = BA, \quad (2)$$

where $A \in \mathcal{R}^{r \times d}$ and $B \in \mathcal{R}^{k \times r}$ are the trainable low-rank matrices and r is the LoRA rank. The corresponding forward computation becomes

$$h = W_0 x + \frac{\alpha}{r} B A x, \quad (3)$$

where α is a scaling factor. This supervised stage guides the model to follow a clinically aligned stepwise reasoning process, including lesion analysis, main-lesion selection and final score determination, but it remains a token-level objective and therefore does not directly optimize the correctness of the final Deauville score.

4.5 Reinforcement Learning with DAPO

Starting from the SFT model, we further optimized the policy using Decoupled Clipping and Dynamic Sampling Policy Optimization (DAPO). The purpose of this stage was to move beyond token-level imitation and directly reinforce reasoning trajectories that lead to the correct final score. For each prompt, multiple candidate responses were sampled from the current policy and evaluated according to the correctness of the final Deauville score. Policy optimization was performed using the DAPO objective

$$J_{\text{DAPO}}(\theta) = \mathbb{E}[\min(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}})\hat{A}_{i,t})], \quad (4)$$

where

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(y_{i,t}|x_i, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x_i, y_{i,<t})} \quad (5)$$

is the importance sampling ratio between the updated policy and the reference policy, and $\hat{A}_{i,t}$ denotes the group-normalized advantage estimated from multiple sampled candidates for the same prompt. Because Deauville scoring yields a discrete and directly verifiable final label, we used a rule-based scalar reward defined on the final predicted score:

$$\mathcal{R}(\hat{y}_i, y_i) = \begin{cases} 1, & \text{if } \hat{y}_i = y_i, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where $\hat{y}_i$ is the predicted Deauville score extracted from the model output and y_i is the ground-truth label. This avoids the need for a learned reward model and directly optimizes outcome correctness. Compared with standard PPO/GRPO-style optimization, DAPO is better suited to long-form clinical reasoning because it uses

decoupled clipping to preserve exploration and reduce entropy collapse, dynamic sampling to avoid ineffective zero-gradient groups, token-level policy-gradient aggregation to improve credit assignment over long reasoning traces, and overlong reward shaping to stabilize excessively long responses. In our setting, these properties are particularly important because the model must produce not only a correct score but also a coherent stepwise reasoning process that remains consistent with the final Deauville decision.

Code availability

The code and trained model weights are available at [Hugging Face repository](#).

Data availability

The PET/CT report data used in this study are not publicly available due to patient privacy and institutional restrictions but may be available from the corresponding authors upon reasonable request and with appropriate institutional approvals.

Acknowledgements:

This work was supported by the National Key Research and Development Program of China (2024YFA1014102), the Peking University Clinical Scientist Training Program (BMU2026PYJH004; the Fundamental Research Funds for the Central Universities), the Beijing Natural Science Foundation (7262234), the Key Clinical Project of Peking

University Third Hospital (BYSYZD2024005 and BYSYZD2023016), the Peking University Third Hospital–United Imaging Research Institution Intelligent Imaging Joint Research and Development Centre Foundation, the National Natural Science Foundation of China (82441023, U23A20295, 62131015, and 82394432), the STI2030-Major Projects of the China Ministry of Science and Technology (2022ZD0209000 and 2022ZD0213100), the Shanghai Municipal Central Guided Local Science and Technology Development Fund (YDZX20233100001001), the Key R&D Program of Guangdong Province, China (2023B0303040001), and the HPC Platform of ShanghaiTech University. The authors sincerely thank the Department of Nuclear Medicine, Beijing Chaoyang Hospital, Capital Medical University, for providing validation data.

Author contributions

M.Z. (Meixin Zhao) and Y.G. (Yaozong Gao) designed the study. M.Z. and collaborators from participating hospitals collected and annotated the data. M.Z., M.Y. (Mingyang Yu) and Y.G. developed the model and performed the experiments and analysis. M.Z., M.Y., and Y.L. (Yuan Li) wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests

Additional information

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to W.Z., Y.G. or D.S.

References

- [1] Hugo JA Adams and Thomas C Kwee. Proportion of false-positive follow-up FDG-PET scans in lymphoma: Systematic review and meta-analysis. *Critical Reviews in Oncology/Hematology*, 141:73–81, 2019.
- [2] Akram Al-Ibraheem, Ahmed Saad Abdlkadir, Malik E Juweid, Kamal Al-Rabi, Mohammad Ma’koseh, Hikmat Abdel-Razeq, and Asem Mansour. FDG-PET/CT in the monitoring of lymphoma immunotherapy response: current status and future prospects. *Cancers*, 15(4):1063, 2023.
- [3] Ronald Boellaard, Irène Buvat, Christophe Nioche, Luca Ceriani, Anne-Ségolène Cottreau, Luca Guerra, Rodney J Hicks, Salim Kanoun, Carsten Kobe, Annika Loft, et al. International benchmark for total metabolic tumor volume measurement in baseline 18F-FDG PET/CT of lymphoma patients: a milestone toward clinical implementation. *Journal of Nuclear Medicine*, 65 (9):1343–1348, 2024.
- [4] Tyler Bradshaw, Amy Weisman, Scott Perlman, and Steve Cho. Automatic image classification using labels from radiology text reports: predicting Deauville scores, 2020.

- [5] Coreline N Burggraaff, Alexander C Cornelisse, Otto S Hoekstra, Pieterella J Lugtenburg, Bart De Keizer, Anne IJ Arens, Filiz Celik, Julia E Huijbregts, Henrica CW De Vet, and Josée M Zijlstra. Interobserver agreement of interim and end-of-treatment 18F-FDG PET/CT in diffuse large B-cell lymphoma: impact on clinical practice and trials. *Journal of Nuclear Medicine*, 59(12): 1831–1836, 2018.
- [6] Sara Duarte, Adriana Roque, Tiago Saraiva, Carolina Afonso, Bárbara Almeida Marques, Carla Barros Lima, Dulcelena Neves, Ana Catarina Lai, Gracinda Costa, Augusta Cipriano, et al. Interim FDG18-PET SUVmax variation adds prognostic value to Deauville 5-point scale in the identification of patients with ultra-high-risk diffuse large B cell lymphoma. *Clinical Lymphoma Myeloma and Leukemia*, 23(2):e107–e116, 2023.
- [7] Thomas W Georgi, Lars Kurch, Dirk Hasenclever, Victoria S Warbey, Lucy Pike, John Radford, Osama Sabri, Regine Kluge, and Sally F Barrington. Interobserver variability in interim PET assessment in Hodgkin lymphoma—reasons and solutions. *PLoS One*, 18(3):e0283694, 2023.
- [8] Ida Häggström, Doris Leithner, Jennifer Alvé, Gabriele Campanella, Murad Abusamra, Honglei Zhang, Shalini Chhabra, Lucian Beer, Alexander Haug, Gilles Salles, et al. Deep learning for [18F] fluorodeoxyglucose-PET-CT classification in patients with lymphoma: a dual-centre retrospective analysis. *The Lancet Digital Health*, 6(2):e114–e125, 2024.

- [9] Zachary Huemann, Changhee Lee, Junjie Hu, Steve Y Cho, and Tyler J Bradshaw. Domain adapted large language models for classifying nuclear medicine reports. *Radiology: Artificial Intelligence*, 5(6):e220281, 2023.
- [10] Daniel T Huff, Victor Santoro-Fernandes, Song Chen, Meijie Chen, Carl Kashuk, Amy J Weisman, Robert Jeraj, and Timothy G Perk. Performance of an automated registration-based method for longitudinal lesion matching and comparison to inter-reader variability. *Physics in Medicine & Biology*, 68(17):175031, 2023.
- [11] Emmanuel Itti, Michel Meignan, Alina Berriolo-Riedinger, Alberto Biggi, Amanda F Cashen, Pierre Véra, Hervé Tilly, Barry A Siegel, Andrea Gallamini, René-Olivier Casasnovas, et al. An international confirmatory study of the prognostic value of early PET/CT in diffuse large B-cell lymphoma: comparison between Deauville criteria and $\Delta\text{SUV}_{\text{max}}$. *European journal of nuclear medicine and molecular imaging*, 40(9):1312–1320, 2013.
- [12] William Karlsen, Lin Akily, Monika Mierzejewska, Jacek Teodorczyk, Artur Bandura, Renata Zaucha, and Wojciech Cytawa. Is ^{18}F -FDG-PET/CT an Optimal Imaging Modality for Detecting Immune-Related Adverse Events after Immune-Checkpoint Inhibitor Therapy? Pros and Cons. *Cancers*, 16(11):1990, 2024.
- [13] Zahra Kiamanesh, Narjess Ayati, Ramin Sadeghi, Eliza Hawkes, Sze Ting Lee, and Andrew M Scott. The value of FDG PET/CT imaging in outcome prediction and response assessment of lymphoma patients treated with immunotherapy: a meta-analysis and systematic review. *European journal of nuclear medicine and molecular imaging*, 49(13):4661–4676, 2022.

- [14] Carsten Kobe, Markus Dietlein, and Dirk Hellwig. PET/CT for lymphoma post-therapy response assessment in Hodgkin lymphoma and diffuse large B-cell lymphoma. In *Seminars in nuclear medicine*, volume 48, pages 28–36. Elsevier, 2018.
- [15] Guillaume Manson, Amaeshi Chukwunonye Lemchukwu, Fatima-Zohra Mokrane, Egesta Lopci, Nicolas Aide, Laetitia Vercellino, Roch Houot, and Laurent Dercle. Interpretation of 2-[18F] FDG PET/CT in Hodgkin lymphoma patients treated with immune checkpoint inhibitors. *European Radiology*, 32(9):6536–6544, 2022.
- [16] Michel Meignan, Andrea Gallamini, Michel Meignan, Andrea Gallamini, and Corinne Haioun. Report on the first international workshop on interim-PET scan in lymphoma. *Leukemia & lymphoma*, 50(8):1257–1260, 2009.
- [17] Susan K Parsons, Kristina S Yu, Nicholas Liu, Supriya Kumar, Michelle A Fanale, Katie Holmes, Carlos Flores, Andy Surinach, Darcy R Flora, and Andrew M Evens. Observations of oncologists on treatment selection with interim positron emission tomography–adapted approaches in classic Hodgkin lymphoma: The real-world connect study. *JCO Oncology Practice*, 19(6):e867–e876, 2023.
- [18] Fabien Ricard, Bruce Cheson, Sally Barrington, Judith Trotman, Annette Schmid, Guenther Brueggenwerth, Gilles Salles, Larry Schwartz, Greg Goldmacher, Rudresh Jarecha, et al. Application of the Lugano classification for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the prolog consensus initiative (part 1—clinical). *Journal of Nuclear Medicine*, 64(1):102–108, 2023.

[19] May Sadik, Jesús López-Urdaneta, Johannes Ulén, Olof Enqvist, Per-Ola Andersson, Rajender Kumar, and Elin Trägårdh. Artificial intelligence increases the agreement among physicians classifying focal skeleton/bone marrow uptake in Hodgkin's lymphoma patients staged with [18F] FDG PET/CT—a retrospective study. *Nuclear Medicine and Molecular Imaging*, 57(2): 110–116, 2023.

[20] Pallawi Torka, Levi D Pederson, Michael V Knopp, David Poon, Jun Zhang, Brad S Kahl, Howard R Higley, Gary Kelloff, Jonathan W Friedberg, Lawrence H Schwartz, et al. Is local review of positron emission tomography scans sufficient in diffuse large B-cell lymphoma clinical trials? A CALGB 50303 analysis. *Cancer medicine*, 12(7):8211–8217, 2023.

[21] Lu Wang, Shixiong Zhang, and Jun Xin. Predicting diffuse large B-cell lymphoma outcomes with lesion-to-liver maximum standardized uptake value for interim-treatment and end-of-treatment positron emission tomography-computed tomography. *Quantitative Imaging in Medicine and Surgery*, 13(10):6789, 2023.

[22] Yuntao Yu, Pierre Decazes, Isabelle Gardin, Pierre Vera, and Su Ruan. 3D lymphoma segmentation in PET/CT images based on fully connected CRFs. In *International Workshop on Computational Methods for Molecular Imaging*, pages 3–12. Springer, 2017.

[23] Yuntao Yu, Pierre Decazes, Jérôme Lapuyade-Lahorgue, Isabelle Gardin, Pierre Vera, and Su Ruan. Semi-automatic lymphoma detection and segmentation using fully conditional random fields. *Computerized Medical Imaging and Graphics*, 70:1–7, 2018.

- [24] Gerben JC Zwezerijnen, Jakoba J Eertink, Coreline N Burggraaff, Sanne E Wiegers, Ekhlash AIN Shaban, Simone Pieplenbosch, Daniela E Oprea-Lager, Pieterella J Lugtenburg, Otto S Hoekstra, Henrica CW de Vet, et al. Interobserver agreement on automated metabolic tumor volume measurements of Deauville score 4 and 5 lesions at interim 18F-FDG PET in diffuse large B-cell lymphoma. *Journal of Nuclear Medicine*, 62(11):1531–1536, 2021.
- [25] Galtier, L Vercellino, L Chartier, et al. Positron emission tomography-imaging assessment for guiding strategy in patients with relapsed/refractory large B-cell lymphoma receiving CAR T cells. *Haematologica*, 108(1):171–180, 2023.
- [26] L Kurch, J Landman-Parker, TW Georgi, et al. Quantitative Deauville Scoring to Uncover Prognostic Information from 18F-FDG PET-Based Response Assessment: Data from the EuroNet-PHL-C1 Trial. *Journal of Nuclear Medicine*, 66(9):1331–1337, 2025.
- [27] Yinhan Liu, Myle Ott, Naman Goyal, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplement.docx](#)