United European Gastroenterology Journal

WILEY | ueg journal

ORIGINAL ARTICLE OPEN ACCESS

Performance of GPT-5 in the Interpretation of IBD Histopathology Reports

Marcello Maida[1,2] | Alessandro Vitello[1,2] | Fabio Salvatore Macaluso[3] | Marco Daperno[4] | Giammarco Mocci[5] | Antonio Rispo[6] | Giulio Calabrese[6] | Nicola L. Decarli[7] | Lucrezia Laschi[8] | Caterina Fattorini[9] | Giorgia Locci[10] | Rachele Del Sordo[11] | Dario Ligresti[12] | Matteo Tacelli[13] | Manuele Furnari[14,15] | Sandro Sferrazza[16] | Giovanni Marasco[17,18] | Antonio Facciorusso[19] | Ambrogio Orlando[3] | Vincenzo Villanacci[20]

[1]Department of Medicine and Surgery, University of Enna "Kore", Enna, Italy | [2]Gastroenterology Unit, Umberto I Hospital, Enna, Italy | [3]Inflammatory Bowel Disease Unit, "Villa Sofia-Cervello" Hospital, Palermo, Italy | [4]Gastroenterology Unit, AO Ordine Mauriziano, Turin, Italy | [5]Division of Gastroenterology, "Brotzu" Hospital, Cagliari, Italy | [6]Gastroenterology, Department of Clinical Medicine and Surgery, School of Medicine "Federico II" of Naples, Naples, Italy | [7]Department of Pathology, Misericordia Hospital, Grosseto, Italy | [8]Pathology Section, Oncology Department, San Giovanni di Dio Hospital, Florence, Italy | [9]Pathology Unit, Azienda Sanitaria Toscana Nord Ovest, Pisa, Italy | [10]Unit of Anatomic Pathology, ARNAS G. Brotzu, Cagliari, Italy | [11]Department of Medicine and Surgery, Section of Anatomic Pathology and Histology, Medical School, University of Perugia, Perugia, Italy | [12]Endoscopy Service, Department of Diagnostic and Therapeutic Services, IRCCS - ISMETT, Palermo, Italy | [13]Division of Biliopancreatic Endoscopy and Endoscopic Ultrasonography Unit, IRCCS San Raffaele Hospital, Milan, Italy | [14]Gastroenterology Unit, Department of Internal Medicine, University of Genoa, Genoa, Italy | [15]Gastroenterology Unit, Department of Medical Specialties, IRCCS Ospedale Policlinico San Martino, Genoa, Italy | [16]Gastroenterology and Endoscopy Unit, ARNAS Civico-Di Cristina-Benfratelli, Palermo, Italy | [17]IRCCS Azienda Ospedaliero Universitaria Di Bologna, Bologna, Italy | [18]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy | [19]Experimental Medicine, Università del Salento, Lecce, Italy | [20]Institute of Pathology ASST-Spedali Civili, University of Brescia, Brescia, Italy

Correspondence: Giovanni Marasco (giovanni.marasco4@unibo.it)

Received: 17 October 2025 | **Revised:** 1 December 2025 | **Accepted:** 9 December 2025

Keywords: crohn disease | GPT-5 | inflammatory bowel diseases | natural language processing | pathology | ulcerative colitis

ABSTRACT

Background: Histopathological interpretation is crucial for diagnosing inflammatory bowel disease (IBD), distinguishing between Crohn's Disease (CD), Ulcerative Colitis (UC), IBD-Unclassified (IBD-U), and Non-IBD colitis (NIBDC). However, interobserver variability and limited expertise can reduce diagnostic accuracy. Large Language Models (LLMs) such as GPT-5 may offer clinical support in interpreting histology reports.

Methods: We analyzed 100 real-life histological reports from ileo-colonoscopies, equally representing CD, UC, IBD-U, and NIBDC, collected across five Italian healthcare centers, including both IBD-specialized and non-specialized hospitals. A reference standard was established by an expert pathologist. Independent classifications were generated by GPT-5, five gastrointestinal pathologists, five IBD-expert gastroenterologists (GIs), and five non-expert GIs. Diagnostic performance (accuracy, recall, precision, F1-score), agreement with the reference standard (Cohen's κ), and inter-rater reliability (Fleiss' κ) were assessed.

Results: GPT-5 achieved the highest agreement with the reference standard with the highest accuracy (76.0%), compared to pathologists (68.6%), IBD-experts (69.2%), and non-experts (63.2%). Agreement with the reference standard was substantial for GPT-5 ($\kappa = 0.671$) and moderate for human groups ($\kappa = 0.508$–0.588). GPT-5 showed perfect recall for CD and UC, high recall for NIBDC (96.0%), but poor performance for IBD-U (recall 8.0%, F1-score 14.3%). Fleiss' κ indicated moderate agreement among pathologists and IBD-experts, and fair agreement among non-experts.

Conclusion: GPT-5 demonstrated reliable performance in interpreting IBD histological reports, exhibiting high accuracy and strong agreement with the reference standard. While unreliable for IBD-U, GPT-5 may serve as a supportive tool in histopathological interpretation of IBD, particularly in centers with limited access to expert pathologists or IBD-specialists.

1 | Introduction

Inflammatory Bowel Disease (IBD) encompasses Crohn's Disease (CD) and Ulcerative Colitis (UC), two chronic inflammatory disorders of the gastrointestinal tract characterized by an aberrant immune response [1].

As for European guidelines, there is no single definitive test for diagnosing CD or UC. Instead, the diagnosis requires an integrated approach combining clinical evaluation, biochemical tests, stool analysis, endoscopic examination, cross-sectional imaging and histological assessment [2].

The initial approach to accurately diagnosing IBD typically involves endoscopy together with microscopic evaluation of an extensive and maximally comprehensive biopsy sampling. This histopathological assessment plays a critical role not only in confirming the clinically hypothesized diagnosis but also in differentiating between CD, UC, histopathological doubtful cases classifiable as IBD-Unclassified (IBD-U), and Non-IBD colitis (NIBDC) [2–6].

However, histopathological interpretation can be challenging for clinicians, especially when the histology report is inconclusive or when physicians lack adequate expertise in this field.

To address this challenge, multilingual Large Language Models (LLMs), such as ChatGPT (Chat Generative Pretrained Transformer, OpenAI), may have the potential to assist in the interpretation of colorectal biopsy histology by providing consistent and timely support in identifying key diagnostic features and reducing interobserver variability.

LLMs have already shown effectiveness in supporting physicians across diverse areas of gastroenterology [7], including IBD [8–10]. However, to date, no study has specifically investigated their potential to recognize the histopathological patterns of IBD.

Based on these premises, this exploratory study aims to evaluate the potential of ChatGPT in assisting both IBD-specialists and non-specialist clinicians with the interpretation of histopathological reports, specifically in distinguishing IBD from Non-IBD conditions and in classifying IBD subtypes (CD, UC, and IBD-U).

2 | Methods

We analyzed 100 randomly selected real-life histological reports from ileo-colonoscopies obtained at five hospitals including Italian tertiary centers for IBD and non-specialized gastroenterology departments. The dataset included twenty-five randomly selected cases for each diagnostic category: UC, CD, IBD-U, and NIBDC [i.e., Infective colitis, Ischemic colitis, Microscopic colitis (Lymphocytic and Collagenous), Segmental colitis associated with diverticulosis (SCAD), Autoimmune colitis, Eosinophilic colitis].

Eligible cases were ileo-colonoscopies with well-oriented biopsies from at least three distinct ileo-colonic segments. Oriented biopsies were defined as specimens placed on a pre-cut cellulose acetate filter in the correct anatomical orientation immediately after sampling.

Exclusion criteria were pediatric patients, sampling limited to a single segment (e.g., isolated ileitis or proctitis), and inadequately oriented biopsies.

The diagnoses selected for the study were rendered by pathologists specializing in the gastrointestinal field. All reports were anonymized before analysis with removal of patient identifiers, clinical data, and the original local diagnosis. Only the number of biopsy samples and essential endoscopic details were retained, specifically the mucosal findings and the anatomical distribution of inflammatory lesions.

An expert pathologist (VV), blinded to the original local diagnosis, reviewed and labeled each report, serving as the reference standard.

The anonymized reports were then submitted to ChatGPT (GPT-5 model, OpenAI) using predefined prompts, and responses were recorded (full dataset, prompts, and outputs available in the Supporting Information S1).

For comparison, the same reports were independently evaluated by three groups of human raters: five gastrointestinal expert pathologists (ND, LL, CF, GL, RDS), five senior gastroenterologists (GIs) with IBD expertise (AO, FSM, GMo, MD, AR), and five GIs without specific IBD expertise (DL, GMa, MT, MF, SS).

IBD-expert GIs were defined as senior gastroenterologists over the age of 40 with more than 10 years of experience in treating IBD and practicing in IBD referral centers.

Data collection was performed via a dedicated Google Forms module.

The primary outcome was the accuracy of GPT-5 in distinguishing CD, UC, IBD-U, and NIBDC compared with the reference standard. The secondary outcome was the agreement between GPT-5 and the three groups of human raters.

This study was reported in accordance with the STARD guidelines (STARD checklist is provided in the Supporting Information S1).

Key Summary

- Summarise the established knowledge on this subject
  - Histopathological assessment is central to diagnosing inflammatory bowel disease (IBD), but interpretation is challenging and interobserver variability is common.
- What are the significant and/or new findings of this study
  - This study shows that GPT-5 achieved the highest agreement with the reference standard with higher overall accuracy in classifying IBD histological reports.
  - In detail, GPT-5 showed excellent performance in classifying Crohn's disease, ulcerative colitis, and non-IBD colitis, but poor performance for IBD-Unclassified (IBD-U).
  - GPT-5 may support clinicians by reducing diagnostic variability and providing timely assistance, especially in settings lacking IBD expertise, while reinforcing the need for multidisciplinary review in IBD-U.

2.1 | Statistical Analysis

Diagnostic performance metrics were derived from confusion matrices generated for each rater group against the reference standard diagnosis. Accuracy, recall, precision, and F1-score were calculated separately for each diagnostic category (UC, CD, IBD-U, and NIBDC). For pathologists, IBD-expert GIs, and non-expert GIs, all individual evaluations from the five members of each group were considered independently and pooled into a single category, yielding 500 observations per group (5×100 cases). GPT-5, by contrast, generated one prediction per report, for a total of 100 observations.

Agreement with the reference standard was quantified using Cohen's κ, which accounts for chance agreement, to directly compare each rater group and GPT-5 with the reference standard. Inter-rater reliability among human raters was assessed using Fleiss' κ to quantify categorical agreement beyond chance, interpreted according to Landis and Koch's criteria (< 0 poor, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, 0.81–1.00 almost perfect) [11].

All statistical analyses were performed using SPSS v. 30.0 for Macintosh (SPSS Inc., Chicago, USA) and R v.4.2.1 (R Foundation for Statistical Computing, Vienna, Austria).

3 | Results

Overall diagnostic performance across all rater groups was evaluated by comparing each classification with the reference standard. Key findings are summarized below, followed by detailed confusion matrices (Figure 1) and performance metrics (Table 1).

Diagnostic accuracy was 68.6% for pathologists, 69.2% for IBD-expert GIs, 63.2% for non-expert GIs, and 76.0% for GPT-5 (Figure 2).

When stratified by disease category, recall varied substantially. For IBD-U, recall was low across all groups, with values of 33.6%, 30.4%, and 32.8% for pathologists, IBD-experts, and non-experts, respectively, and only 8.0% for GPT-5. In contrast, recall for CD was high, ranging from 68.0% for non-experts GIs to 86.4% for IBD-experts GIs, with GPT-5 achieving 100%. Similarly, recall for UC was 84.8%, 81.6%, and 76.0% for pathologists, experts, and non-experts, respectively, again with GPT-5 at 100%. For NIBDC, recall was 76.0%, 78.4%, and 76.0% in the three human groups, and 96.0% for GPT-5 (Table 1).

Precision also showed variability across groups. For IBD-U, precision was 56.0%, 52.1%, and 42.3% for pathologists, experts, and non-experts, respectively, compared to 66.7% for GPT-5. For CD, precision ranged from 76.6% to 83.3% among human raters, while GPT-5 achieved 67.6%. Precision for UC was 60.6%, 72.3%, and 57.9% in the human groups, compared to 80.6% for GPT-5. Finally, for NIBDC, precision was 73.1%, 67.6%, and 69.3% for the three human groups, and 82.8% for GPT-5 (Table 1).

The F1-score, combining precision and recall, confirmed these findings. For IBD-U, F1 was 42.6%, 38.5%, and 36.9% for pathologists, experts, and non-experts, respectively, and only 14.3% for GPT-5. For CD, F1 was 81.6%, 81.3%, and 75.0% in the human groups, compared to 80.6% for GPT-5. For UC, F1 was 70.0%, 76.7%, and 65.9% in the human groups, with GPT-5 achieving the highest value at 89.3%. For NIBDC, F1 was 74.5%, 72.8%, and 72.5% in the human groups, and 89.0% for GPT-5 (Table 1).

Agreement with the reference standard, expressed as Cohen's κ, was 0.580 for pathologists (moderate agreement), 0.588 for IBD-expert GIs (moderate agreement), 0.508 for non-expert GIs (moderate agreement), and 0.671 for GPT-5 (substantial agreement) (Figure 3).

Inter-rater agreement within human subgroups, assessed with Fleiss' κ, showed moderate agreement among pathologists ($\kappa = 0.531$, 95% CI 0.449–0.610) and IBD-experts ($\kappa = 0.475$, 95% CI 0.386–0.558), while non-experts achieved only fair agreement ($\kappa = 0.375$, 95% CI 0.293–0.454).

4 | Discussion

The application of AI to histopathological assessment in IBD appears particularly promising, supported by an expanding body of literature [12–14] and reinforced by the high awareness and generally favorable perceptions of AI systems among GIs [15, 16].

Our findings align with emerging literature indicating that LLMs may assist pathologists in interpreting complex histopathological information while facing important contextual and methodological limitations [17].

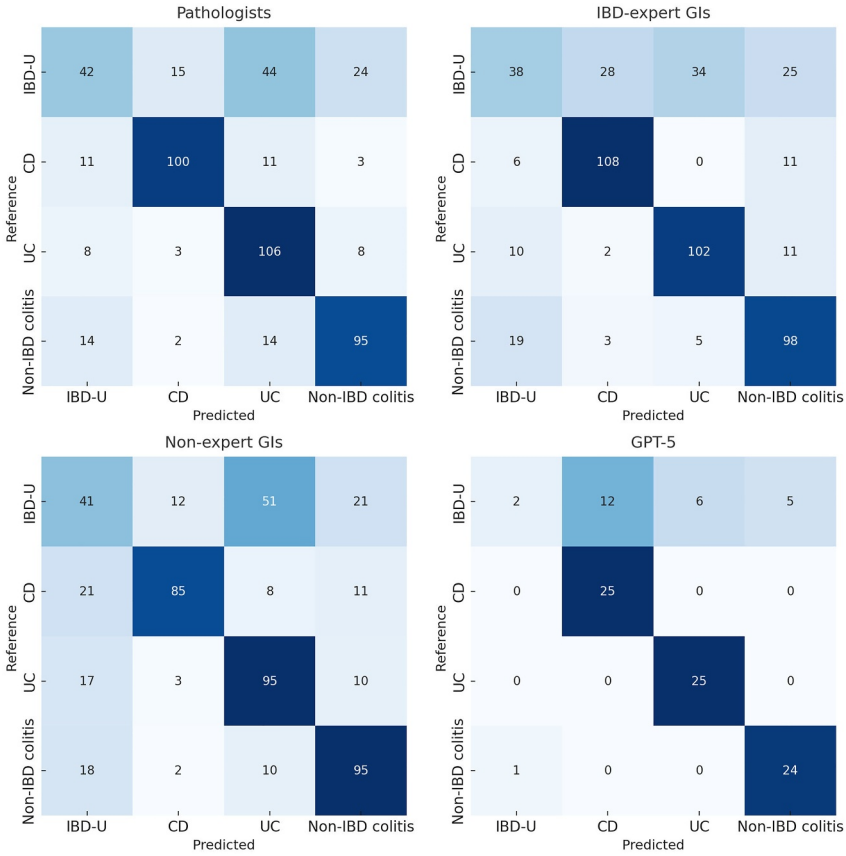

FIGURE 1 | Confusion matrices of diagnostic classifications by different rater groups (*rows* represent the reference—true—diagnosis; *columns* represent the predicted diagnosis; values along the diagonal correspond to correct classifications; off-diagonal values represent misclassifications).

TABLE 1 | Comparative performance metrics of pathologists, IBD-expert GIs, non-expert GIs, and GPT-5 compared with the reference standard.

Metric	Pathologists	IBD-expert GIs	IBD non-expert GIs	GPT-5
Accuracy (%)	68.6 (64.4–72.5)	69.2 (65.0–73.1)	63.2 (58.9–67.3)	76.0 (66.8–83.3)
Recall (%)				
-IBD-U	33.6 (25.9–42.3)	30.4 (23.0–38.9)	32.8 (25.2–41.4)	8.0 (2.2–25.0)
-CD	80.0 (72.1–86.1)	86.4 (79.3–91.3)	68.0 (59.4–75.5)	100.0 (86.7–100.0)
-UC	84.8 (77.5–90.0)	81.6 (73.9–87.4)	76.0 (67.8–82.6)	100.0 (86.7–100.0)
-NIBDC	76.0 (67.8–82.6)	78.4 (70.4–84.7)	76.0 (67.8–82.6)	96.0 (80.5–99.3)
Precision (%)				
-IBD-U	56.0 (44.7–66.7)	52.1 (40.8–63.1)	42.3 (32.9–52.2)	66.7 (20.8–93.9)
-CD	83.3 (75.7–88.9)	76.6 (69.0–82.8)	83.3 (74.9–89.3)	67.6 (51.5–80.4)
-UC	60.6 (53.2–67.5)	72.3 (64.4–79.1)	57.9 (50.3–65.2)	80.6 (63.7–90.8)
-NIBDC	73.1 (64.9–80.0)	67.6 (59.6–74.7)	69.3 (61.2–76.4)	82.8 (65.5–92.4)
F1-score (%)				
-IBD-U	42.6 (33.1–50.0)	38.5 (28.6–46.5)	36.9 (28.9–44.7)	14.3 (0.0–33.3)
-CD	81.6 (75.7–86.6)	81.3 (76.0–85.9)	75.0 (68.0–81.3)	80.6 (67.9–90.5)
-UC	70.0 (64.5–76.4)	76.7 (70.9–82.1)	65.9 (59.5–71.6)	89.3 (80.0–96.7)
-NIBDC	74.5 (68.5–80.0)	72.8 (66.4–78.0)	72.5 (66.1–78.7)	89.0 (77.8–96.4)

However, existing research has largely overlooked the potential of LLMs to assist physicians in interpreting histopathological reports for patients with IBD.

In this exploratory study, we evaluated the ability of GPT-5, a state-of-the-art LLM, to classify histopathological reports of IBD in comparison to expert gastrointestinal pathologists, GIs with

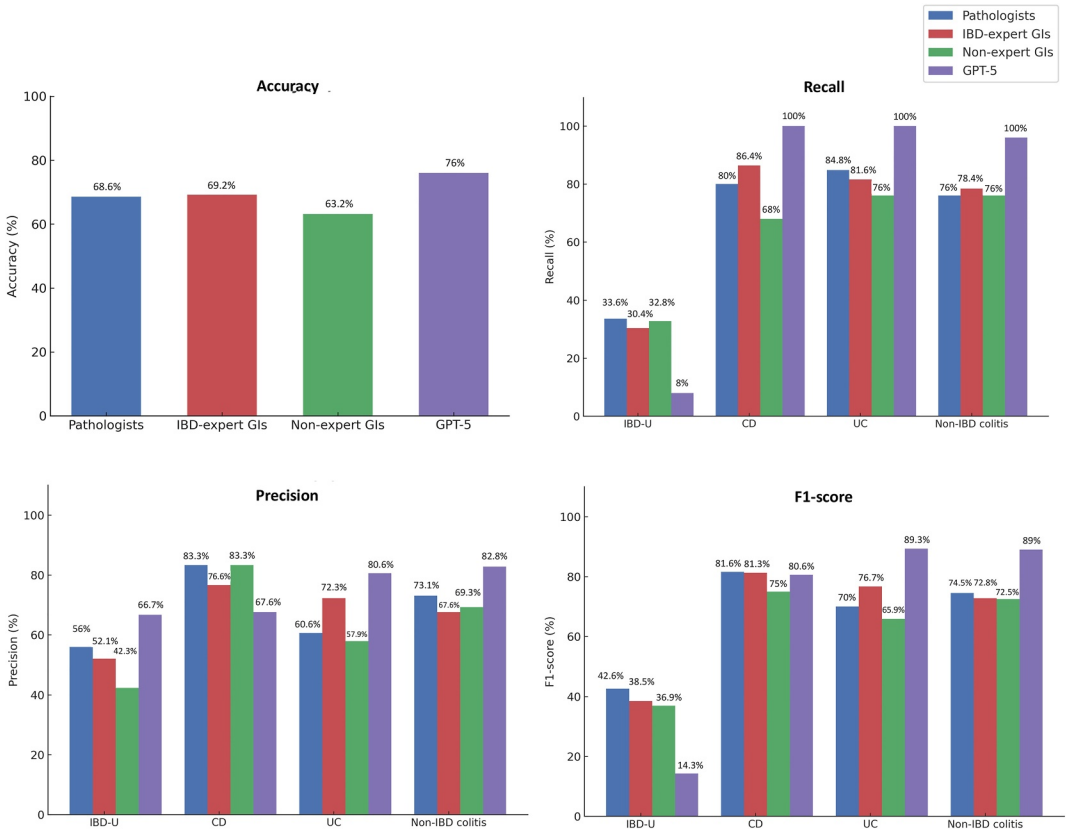

FIGURE 2 | Diagnostic performance (accuracy, recall, precision, and F1-score) of different rater groups.

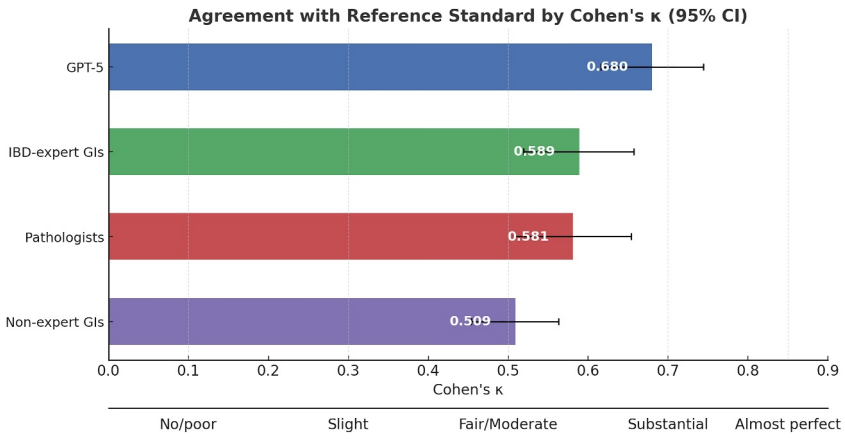

FIGURE 3 | Agreement with the reference standard for pathologists, IBD-expert GIs, non-expert GIs, and GPT-5 by Cohen's κ.

expertise in IBD, and non-expert GIs. GPT-5 achieved higher overall accuracy and substantial agreement with the reference standard, performing excellently for UC and CD, acceptably for NIBDC, but poorly for IBD-U.

Notably, both GPT-5 and human raters struggled with IBD-U, highlighting the inherent diagnostic challenges associated with this category.

The poor performance for IBD-U reflects the intrinsic uncertainty of this diagnosis, which often serves as a "diagnostic placeholder" when the features of UC and CD overlap. In real-world practice, this difficulty is often exacerbated by the limited expertise of general pathologists and suboptimal orientation or inadequate biopsy sample, which can hinder an accurate evaluation of mucosal architecture and inflammatory distribution.

Our results align with several consensus documents that acknowledge the limited reproducibility of IBD-U diagnoses [2–6].

In particular, GPT-5 showed the poorest performance for IBD-U. This may be explained by its low prevalence (and consequently limited representation of this category in model training), the highly heterogeneous and non-standardized terminology used within IBD-U reports, and the frequent use of

nuanced or attenuated histological descriptors in this category that human readers can contextualize, unlike the LLM.

We believe that, in most cases, the IBD-U designation reflects an inconclusive label rather than a distinct IBD subtype, particularly when the disease is in its early or fulminant stages, presents in atypical forms, or when the pathologist does not have access to complete patient information (e.g., crucial indicators related to the patient's clinical history, such as prior medication use, among others).

Therefore, in accordance with the latest European guidelines [2] and supported by the findings of our study, we recommend that, in cases of IBD-U, well-oriented biopsy specimens be reassessed by an expert gastrointestinal pathologist. Additionally, repeat endoscopy with biopsies, wireless video capsule endoscopy (VCE), imaging, or a combination of these modalities should be considered to aid in disease reclassification.

This study contributes to the existing literature by quantitatively benchmarking GPT-5 against multiple rater groups, providing one of the first comparative evaluations of a cutting-edge LLM in histopathology. A key advantage of using GPT-5, the most recently released model, is its significant improvements in contextual understanding, factual accuracy, and reasoning ability compared with earlier GPT versions. Unlike many other LLMs, GPT-5 incorporates advanced multimodal training and a broader medical knowledge base, which may account for its superior performance in most categories.

From a clinical perspective, GPT-5 could serve as a valuable decision-support tool for clearly identifiable UC, CD, or NIBDC cases. It has the potential to reduce diagnostic variability and provide assistance in centers with limited access to expert pathologists or IBD-specialists. However, for IBD-U, both AI and human raters prove to be unreliable, emphasizing the importance of multidisciplinary case discussions and the cautious integration of AI.

Despite the encouraging results, it is crucial to emphasize that GPT-5 should be used with caution and regarded strictly as a clinician-support tool rather than an autonomous diagnostic system. In fact, the model can only provide probabilistic confidence outputs and cannot operate independently as a diagnostic decision model. Any practical application would require expert verification and the careful integration of AI-generated outputs with clinical and endoscopic information, ensuring that such classifications support rather than replace human diagnostic judgment.

Moreover, the good performance of the LLM should not be interpreted as diagnostic superiority over human experts, as all performance metrics necessarily reflect agreement with the reference pathologist rather than estimates of absolute diagnostic correctness.

This study has several strengths, including the use of real-world histology reports based on well-oriented biopsies collected from multiple centers with varying levels of clinical and diagnostic specialization in IBD, a balanced dataset across different diagnoses, and the inclusion of multiple rater groups with rigorous statistical analyses.

Limitations include the relatively small sample size, which also limited the feasibility of meaningful subgroup analyses, its retrospective design, and the restriction to endo-histological information. In this regard, clinical data were intentionally excluded to avoid influencing raters, but this decision may have impacted diagnostic accuracy since clinical correlation is often crucial for resolving indeterminate cases.

A further limitation is the use of a single expert pathologist as the reference standard. Although this approach ensures internal consistency and avoids multi-reader adjudication, it inherently lacks the robustness of a multi-expert consensus, potentially limiting the reproducibility of the diagnostic classification.

\Finally, the strict inclusion criteria of this study, requiring well-oriented biopsies from at least three ileo-colonic segments, although adopted to maximize internal validity, may limit external generalizability, since sampling quality in routine clinical practice can vary across centers and patient scenarios.

Future directions should encompass larger prospective studies with a greater representation of IBD-U cases and the evaluation of LLMs across languages and healthcare contexts, given that cultural and linguistic factors may influence AI outputs. Furthermore, AI systems should be trained on multimodal data that integrate histology, endoscopy, and clinical metadata to more accurately reflect real-world diagnostic reasoning. Comparative studies involving text-based LLMs and image-based AI pathology models are also warranted, along with the development of explainable AI outputs to enhance transparency in ambiguous scenarios.

Finally, future studies should also evaluate whether the use of LLMs has a measurable impact on patient-centered outcomes, which were not assessed in the present study.

5 | Conclusion

GPT-5 demonstrated strong overall performance in interpreting histological reports from both IBD and non-IBD patients, highlighting its potential as a supportive tool in clinical practice. However, its performance was consistently poor in IBD-U, mirroring the challenges also faced by human raters. These findings emphasize that GPT-5 should complement, rather than replace, clinical decision-making by assisting clinicians in the interpretation of histopatological reports, particularly in diagnostically ambiguous cases.

Greater efforts should focus on improving diagnostic precision through multidisciplinary collaboration and enhanced communication between pathologists and clinicians.

Author Contributions

M.M., A.V. and V.V.: conception, design of the study, drafting of the manuscript. M.M., A.V., V.V., F.S.M., A.O.: critical revision of the manuscript for relevant intellectual content. M.M.: statistical analysis. M.D., G.M., A.R., G.C., N.L.D., L.L., C.F., G.L., R.D.S., D.L., M.T., M.F., S.S., A.F.: comments. All the authors contributed to the manuscript

editing and had full control over the preparation, final reading, and approval of the manuscript.

Acknowledgements

Open access publishing facilitated by Universita degli Studi di Bologna, as part of the Wiley - CRUI-CARE agreement.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Available upon reasonable request.

References

1. T. Chhibba, B. Gros, J. A. King, et al., "Environmental Risk Factors of Inflammatory Bowel Disease: Toward a Strategy of Preventative Health," *Journal of Crohn's and Colitis* 19, no. 4 (2025): jjaf042, https://doi.org/10.1093/ecco-jcc/jjaf042.

2. T. Kucharzik, S. Taylor, M. Allocca, et al., "Ecco-Esgar-Esp-Ibus Guideline on Diagnostics and Monitoring of Patients With Inflammatory Bowel Disease: Part 1," *Journal of Crohn's and Colitis* 19, no. 7 (2025): jjaf106, https://doi.org/10.1093/ecco-jcc/jjaf106.

3. F. Magro, C. Langner, A. Driessen, et al., "European Consensus on the Histopathology of Inflammatory Bowel Disease," *Journal of Crohn's and Colitis* 7, no. 10 (2013): 827–851, https://doi.org/10.1016/j.crohns.2013.06.001.

4. C. Langner, F. Magro, A. Driessen, et al., "The Histopathological Approach to Inflammatory Bowel Disease: A Practice Guide," *Virchows Archiv* 464, no. 5 (2014): 511–527, https://doi.org/10.1007/s00428-014-1543-4.

5. V. Villanacci, L. Reggiani-Bonetti, T. Salviato, et al., "Histopathology of Ibd Colitis: A Practical Approach From the Pathologists of the Italian Group for the Study of the Gastrointestinal Tract (Gipad)," *Pathologica* 113, no. 1 (2021): 39–53, https://doi.org/10.32074/1591-951x-235.

6. F. Magro, G. Doherty, L. Peyrin-Biroulet, et al., "Ecco Position Paper: Harmonization of the Approach to Ulcerative Colitis Histopathology," *Journal of Crohn's and Colitis* 14 (2020): 1503–1511, https://doi.org/10.1093/ecco-jcc/jjaa110.

7. M. Maida, C. Celsa, L. H. S. Lau, et al., "The Application of Large Language Models in Gastroenterology: A Review of the Literature," *Cancers* 16, no. 19 (2024): 3328, https://doi.org/10.3390/cancers16193328.

8. N. Labarile, A. Vitello, E. Sinagra, et al., "Artificial Intelligence in Advancing Inflammatory Bowel Disease Management: Setting New Standards," *Cancers* 17, no. 14 (2025): 2337, https://doi.org/10.3390/cancers17142337.

9. O. Ozturk, M. Ergul, Y. Cagir, et al., "Assessing chatgpt-v4 for Guideline-Concordant Inflammatory Bowel Disease: Accuracy, Completeness, and Temporal Drift," *Journal of Clinical Medicine* 14, no. 13 (2025): 4599, https://doi.org/10.3390/jcm14134599.

10. M. Sciberras, Y. Farrugia, H. Gordon, et al., "Accuracy of Information Given by Chatgpt for Patients With Inflammatory Bowel Disease in Relation to Ecco Guidelines," *Journal of Crohn's and Colitis* 18, no. 8 (2024): 1215–1221, https://doi.org/10.1093/ecco-jcc/jjae040.

11. J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics* 33, no. 1 (1977): 159–174, https://doi.org/10.2307/2529310.

12. C. Furlanello, N. Bussola, N. Merzi, et al., "The Development of Artificial Intelligence in the Histological Diagnosis of Inflammatory Bowel Disease (ibd-ai)," *Digestive and Liver Disease* 57, no. 1 (2025): 184–189, https://doi.org/10.1016/j.dld.2024.05.033.

13. M. Iacucci, G. Santacroce, P. Meseguer, et al., "Endo-Histo Foundational Fusion Model: A Novel Artificial Intelligence for Assessing Histologic Remission and Response to Therapy in Ulcerative Colitis Clinical Trial," *Journal of Crohn's and Colitis* 19, no. 7 (2025): jjaf108, https://doi.org/10.1093/ecco-jcc/jjaf108.

14. G. Santacroce, I. Zammarchi, O. M. Nardone, et al., "Rediscovering Histology - The Application of Artificial Intelligence in Inflammatory Bowel Disease Histologic Assessment," *Therapeutic Advances in Gastroenterology* 18 (2025): 17562848251325525, https://doi.org/10.1177/17562848251325525.

15. M. Maida, S. Sferrazza, G. Calabrese, et al., "Perceptions of Artificial Intelligence Among Gastroenterologists in Italy: A National Survey," *Cancers* 17, no. 8 (2025): 1353, https://doi.org/10.3390/cancers17081353.

16. C. L. Leggett, S. Parasa, A. Repici, T. M. Berzin, S. A. Gross, and P. Sharma, "Physician Perceptions on the Current and Future Impact of Artificial Intelligence to the Field of Gastroenterology," *Gastrointestinal Endoscopy* 99, no. 4 (2024): 483–489.e2, https://doi.org/10.1016/j.gie.2023.11.053.

17. E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine With a Focus on Digital Pathology - A Recent Scoping Review," *Diagnostic Pathology* 19, no. 1 (February 2024): 43: PMID: 38414074; PMCID: PMC10898121, https://doi.org/10.1186/s13000-024-01464-7.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.
Supporting Information S1: ueg270161-sup-0001-suppl-data.docx.