Deadline: September 26 (Thursday) | 11:00 AM Pacific time zone | 2:00 PM EASTERN time zone | 8:00PM Amsterdam time

Title (40 words)

Towards a Multicentric Open Digital Pathology Assistant Benchmark: Initial Results from the DALPHIN Study

Body (2500 characters in total)

Background

Artificial Intelligence in the form of chatbots is an emerging reality. Equipped with vision-language capabilities, models like GPT-4o or Gemini 1.5 Pro can process both images and text, representing a possibility of multi-modal virtual assistants in healthcare. While there is an urgent need and growing expectation to adopt digital assistants to support clinical diagnostics, clinicians and researchers must question chatbots' capability to answer diagnostic questions. To address this necessity, we created DALPHIN, a multicentric open benchmark for virtual assistants applied to diagnostic problems in digital pathology. Here we present the initial results of the DALPHIN study.

Design

For this pilot study, we collected 56 whole-slide images from 4 medical centers in 4 countries, covering 11 organs and 32 diseases (including neoplasms and inflammatory diseases). Pathologists identified regions of interest (ROIs) and formulated "entry-level" and "advanced" questions to answer by analyzing one or more ROIs from the same slide. As entry-level, we asked the organ present in the image (Q1), whether a neoplasm was present (Q2), and if yes, whether the neoplasm was benign or malignant (Q3). As advanced questions, we asked about the diagnosis of the lesion (Q4), organ-, tissue- or lesion-specific questions with multiple-choice answers (Q5), and free response questions (Q6). In total, DALPHIN comprises 128 unique ROIs and 292 ROI-question pairs. We considered two generalist (GPT-4o, Gemini 1.5 Pro) and one pathology-specific (PathChat 1.0) models, fed them with ROIs and questions optimized as prompts, and analyzed the produced text output. Having the expected answers, we measured accuracy for all questions, with Q4 and Q6 assessed by a pathologist.

Results

PathChat outperformed generalist models on both entry-level and advanced questions (Table 1). For entry-level, generalist models achieved an average accuracy >55.2%, showing a degree of pathology knowledge, while performance dropped to <26% in advanced questions. PathChat correctly answered most entry-level and multiple-choice questions, but achieved lower performance on Q4 and Q6, highlighting the complexity of distinguishing lesions (Q4) and of generating free-text answers (Q6).

	Entry-level				Advanced			
	Q1	**Q2**	**Q3**	**Average**	**Q4**	**Q5**	**Q6**	**Average**
GPT-4o	32.14%	75.00%	58.54%	55.23%	14.29%	51.79%	11.11%	25.70%
Gemini 1.5 Pro	33.93%	73.21%	73.17%	60.10%	7.14%	50.89%	18.52%	25.50%
PathChat 1.0	67.87%	96.43%	90.24%	84.84%	39.29%	74.11%	40.74%	51.38%

Table 1. Quantitative results for the different types of questions and models used in DALPHIN.

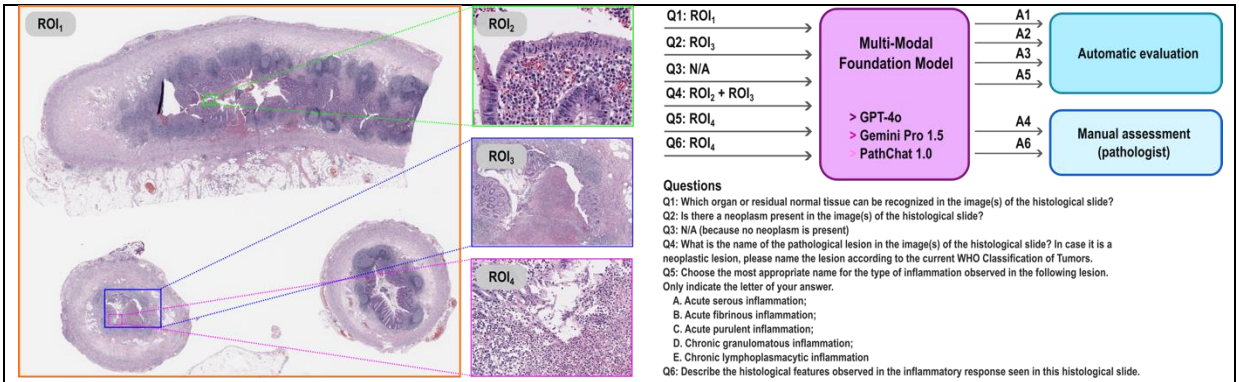

Figure 1. Schematic overview of the DALPHIN data processing. ROI: Region Of Interest. Q1-Q6: questions; A1-A6: answers. Automatic evaluation and manual assessment by pathologist were done based on the answers produced by each foundation model compared to the expected answers provided by each clinical center.

Conclusion

The pathology-specific model outperformed generalist models, and while its performance on some advanced questions is suboptimal, this shows some promise for supporting pathologists in their diagnostics while focusing on advanced aspects. We aim at expanding the DALPHIN benchmark and make it publicly accessible.

Carlijn Lems^1* <carlijn.lems@radboudumc.nl>
Natalie Klubickova^1,2* <Natalie.Klubickova@radboudumc.nl>
Biagio Brattoli^3 <biagio@lunit.io>
Taebum Lee^3 <taebum.lee@lunit.io>
Seokhwi Kim^4 <seokhwikim@ajou.ac.kr>
Veronica Vilaplana Besler^5 <veronica.vilaplana@upc.edu>
Pedro Fernandez^6 <plfernandez.germanstrias@gencat.cat>
Laura Pons^6 <lponsmar.germanstrias@gencat.cat>
Arvydas Laurinavicius^7 <Arvydas.Laurinavicius@vpc.lt>
Julius Drachneris^7 <Julius.Drachneris@vpc.lt>
Diana Montezuma Felizardo^8 <diana.felizardo@impdiagnostics.com>
Domingos Oliveira^8 <domingos.oliveira@impdiagnostics.com>
Shoko Vos^1 <Shoko.Vos@radboudumc.nl>
Maschenka Balkenhol^1 <Maschenka.Balkenhol@radboudumc.nl>
Jolique van Ipenburg^1 <Jolique.vanIpenburg@radboudumc.nl>
Anne-Marie Vos^1 <Anne-Marie.M.Vos@radboudumc.nl>
Milda Poceviciute^1# <milda.poceviciute@radboudumc.nl>
Nadieh Khalili^1# <nadieh.khalili@radboudumc.nl>
Francesco Ciompi^1# <francesco.ciompi@radboudumc.nl>

1. Radboud University Medical Center, Nijmegen (Netherlands)
2. Biopticka Laboratory Ltd., Pilsen (Czech Republic)
3. Lunit, Seoul (South Korea)
4. Ajou University School of Medicine (South Korea)
5. Universitat Politecnica de Catalunya, Barcelona (Spain)
6. Hospital Universitari Germans Trias I Pujol, Badalona (Spain)
7. Vilnius University and National Centre of Pathology, Vilnius (Lithuania)
8. Imp Diagnostics, Porto (Portugal)

- * equal contribution
- # joint supervision