Proceedings of the 6th International Conference on Biological Engineering and Medical Science
DOI: 10.54254/2753-8818/2026.36040

Closing the Loop: An Integrated Vision-Calibration-Actuation System with LLM-Assisted Diagnostic Interface for Autonomous Topical Therapy of Urticaria

Zhilin Sui

The Hun School of Princeton, Princeton, USA
suizhilin0715@outlook.com

Abstract. Urticaria is common, recurrent, and often managed outside the clinic, where patients must repeatedly localize lesions and apply topical agents themselves. Current AI dermatology systems mainly stop at diagnosis, while dermatologic robots remain procedure-specific and rarely combine reasoning with physical delivery. This paper reports a feasibility study of an integrated topical-therapy pipeline for urticaria. The system combines (i) ResNet-50 U-Net segmentation trained on 98 manually annotated lesion photographs, (ii) 9-point affine hand-eye calibration followed by deterministic boustrophedon coverage planning for a 6-DOF arm, and (iii) a LoRA-adapted Llama-3.2-3B diagnostic interface trained on a 2,752-pair urticaria instruction corpus with per-patient memory for multi-turn allergen-elimination dialogue. The LLM runs in parallel with the actuation chain and does not control the robot's geometric plan. On a benchtop phantom setup, the released artifact achieves held-out Dice 0.810 ± 0.083 on the validation split, sub-millimeter calibration residuals in a representative session, and complete planned coverage of segmented lesion regions. The contribution is not a new segmentation, calibration, planning, or LLM algorithm in isolation, but a reproducible integration of these modules toward autonomous robotic dermatologic care.

Keywords: medical robotics, large language model, image segmentation, dermatology, hand-eye calibration

1. Introduction

Urticaria, characterized by recurrent pruritic wheals and angio-edema, is one of the most prevalent inflammatory skin conditions worldwide, with an estimated lifetime prevalence approaching 20% and rising age-standardized disability metrics in the most recent global burden analyses [1] (Figure 1).

DOI: 10.54254/2753-8818/2026.36040

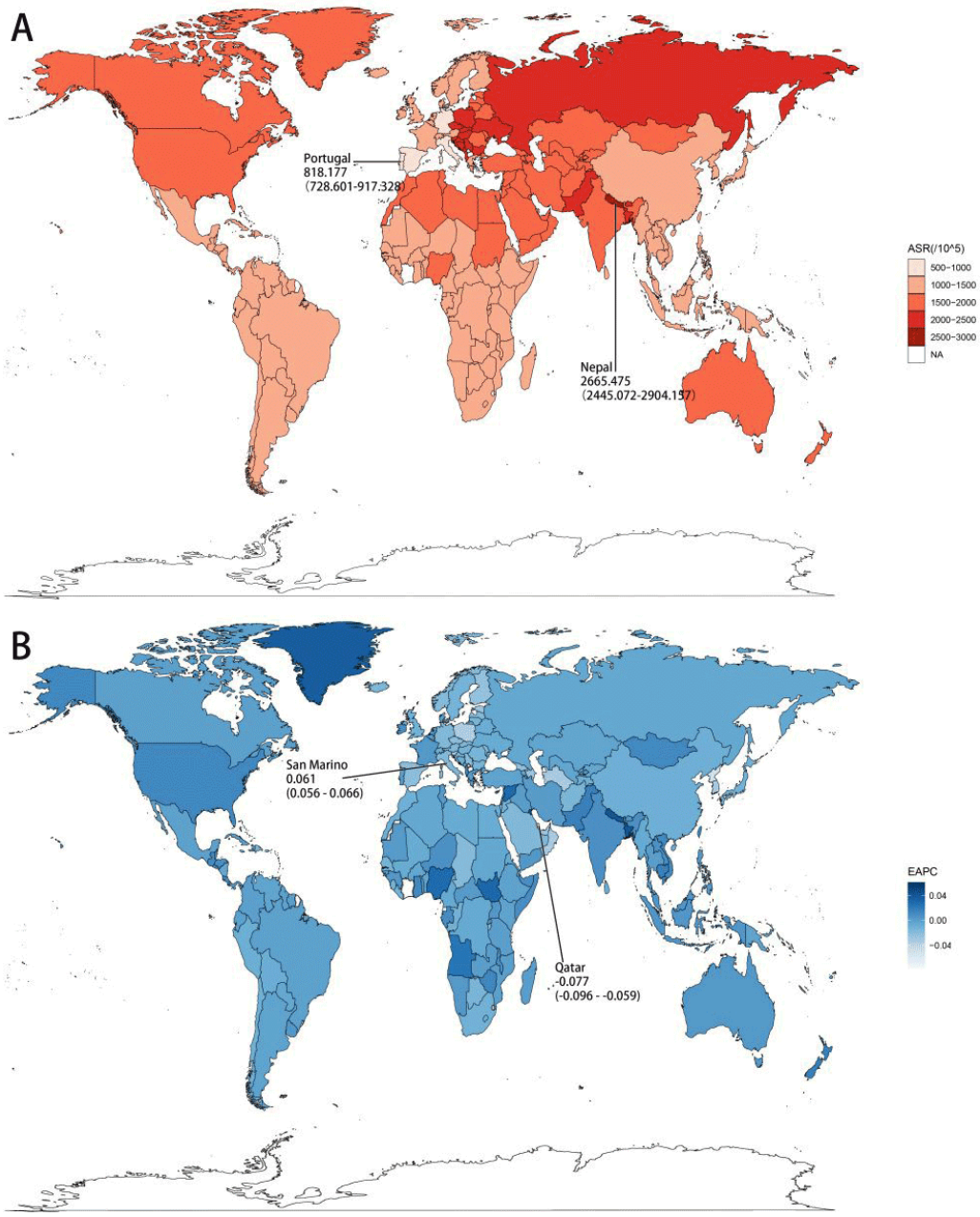

Figure 1. Global incidence and prevalence footprint of urticaria

Current international guidelines recommend a step-wise regimen anchored on H1-antihistamines and supplemented with topical agents (corticosteroids, emollients, antipruritic creams) for symptomatic relief [2]. In practice, management is often self-administered: lesions are episodic, distributed, and difficult to cover uniformly. This creates an opening for assistive systems that can localize affected skin and deliver topical therapy reproducibly.

Dermatologic AI and medical robotics have advanced along separate tracks. Vision models can classify or segment skin findings [3, 4], and multimodal medical LLMs can produce clinically shaped text from images and history [5, 6]. Robotic dermatology has also shown procedure-specific feasibility, most prominently in hair restoration and cosmetic light delivery [7, 8]. Yet these systems

DOI: 10.54254/2753-8818/2026.36040

leave a perception—reasoning—actuation gap: diagnostic systems do not treat, and dermatologic robots rarely include patient-facing AI reasoning.

This paper asks a system-integration question: can a single benchtop platform combine an LLM-assisted urticaria interface, pixel-level lesion localization, hand-eye-calibrated motion, and topical delivery by a 6-DOF arm? The goal is not to claim novelty in any single module, but to show that the full stack can be assembled, documented, and evaluated as a reproducible feasibility artifact.

The proposed system has three modules. The vision module trains a ResNet-50 U-Net [9] on 98 annotated urticaria photographs. The actuation module aligns camera pixels to robot coordinates through 9-point affine hand-eye calibration [10] and sweeps the resulting mask with a deterministic boustrophedon planner [11]. The diagnostic module adapts Llama-3.2-3B with LoRA on a 2,752-pair urticaria instruction corpus and keeps a per-patient JSON memory for multi-turn allergen-elimination reasoning. The LLM is a parallel diagnostic interface; it does not modulate the robot's geometric plan.

The contributions of this paper are as follows:

• C1. This paper presents an integrated urticaria-care pipeline that combines lesion segmentation, calibrated 6-DOF actuation, and an LLM-assisted diagnostic interface.

• C2. The work implements and evaluate a reproducible vision-to-actuation chain, reporting held-out segmentation, calibration, and coverage behavior on the released artifact.

• C3. This work adapt a small LLM to urticaria dialogue with LoRA and patient memory, while explicitly separating diagnostic reasoning from robot motion control.

The remainder of the paper reviews related work, describes the system, reports experiments, and discusses limitations and safety.

2. Related work

The proposed platform combines three mature research lines: dermatologic AI, robotic skin therapy, and medical LLM interfaces. This section summarizes only the work needed to position the integration claim.

2.1. AI-assisted dermatologic diagnosis

Convolutional models have reached strong performance on photographic skin-lesion classification and differential diagnosis [3, 4]. SkinGPT-4 extends this direction by pairing visual features with a language model for free-text dermatologic responses [5]. These systems are valuable, but their endpoint is still a label, report, or recommendation. The therapy step remains manual. This work uses vision output as an input to downstream motion planning, closing a practical part of the loop that diagnostic-only systems leave open.

2.2. Robotic systems in dermatology and skin therapy

Medical robotics has enabled minimally invasive surgery and supervised autonomy [12, 13]. Dermatology has narrower precedents: ARTAS automates follicular-unit extraction for hair restoration [7], and Mu et al. report a vision-guided 6-DOF arm for cosmetic photo-rejuvenation [8]. These systems show that robotic skin interaction is feasible, but they are built for specific procedures and do not include an LLM diagnostic interface or lesion-mask-driven topical delivery.

DOI: 10.54254/2753-8818/2026.36040

2.3. LLMs in clinical and dermatologic decision support

Med-PaLM and other medical LLMs demonstrate that large models can encode clinical knowledge and answer medical questions [6, 14]. LoRA makes domain adaptation possible without full-parameter fine-tuning [15]. Here the LLM is used as a patient-facing diagnostic and education interface with longitudinal memory. It is intentionally not placed in the robot control loop; segmentation and actuation remain deterministic.

2.4. Position with respect to prior work

Table 1. Comparison of related work across five dimensions: vision-based segmentation (Vis), robotic actuation (Rob), large language model (LLM)-assisted decision (LLM), dermatologic application (Derm), and integrated end-to-end pipeline (E2E)

Work	Vis	Rob	LLM	Derm	E2E
SkinGPT-4	✓	—	✓	✓	—
Esteva et al.	✓	—	—	✓	—
Photo-rejuvenation	✓	✓	—	✓	—
ARTAS	✓	✓	—	✓	—
da Vinci	✓	✓	—	—	—
Med-PaLM	—	—	✓	—	—
This work	✓	✓	✓	✓	✓

Table 1 summarizes the positioning. Prior systems cover subsets of the desired capabilities; this work contributes their task-grounded composition for topical urticaria therapy.

3. Methodology

The system is organized as a vision-to-actuation chain plus a parallel LLM diagnostic interface. The implemented modules are lesion segmentation, hand-eye calibration, coverage path planning, hardware delivery, and urticaria-focused dialogue support.

3.1. System overview

Figure 2 shows the pipeline. A HIKROBOT structured-light camera captures an RGB image of the affected region. A ResNet-50-backbone U-Net predicts a binary lesion mask at 512\times512 resolution. A deterministic boustrophedon planner converts the mask into image-plane waypoints, and a 9-point affine hand-eye calibration maps those waypoints into the coordinate frame of a 6-DOF robotic arm. A custom multi-nozzle end-effector then applies topical agent under contact-pressure gating. In parallel, the same case context is passed to an LLM interface that produces a structured urticaria report and multi-turn allergen-elimination dialogue.

The diagnostic LLM is not a robot controller. It does not change the geometric coverage path, dose, or dwell time in the current artifact; those steps are deterministic and separately safety-gated.

DOI: 10.54254/2753-8818/2026.36040

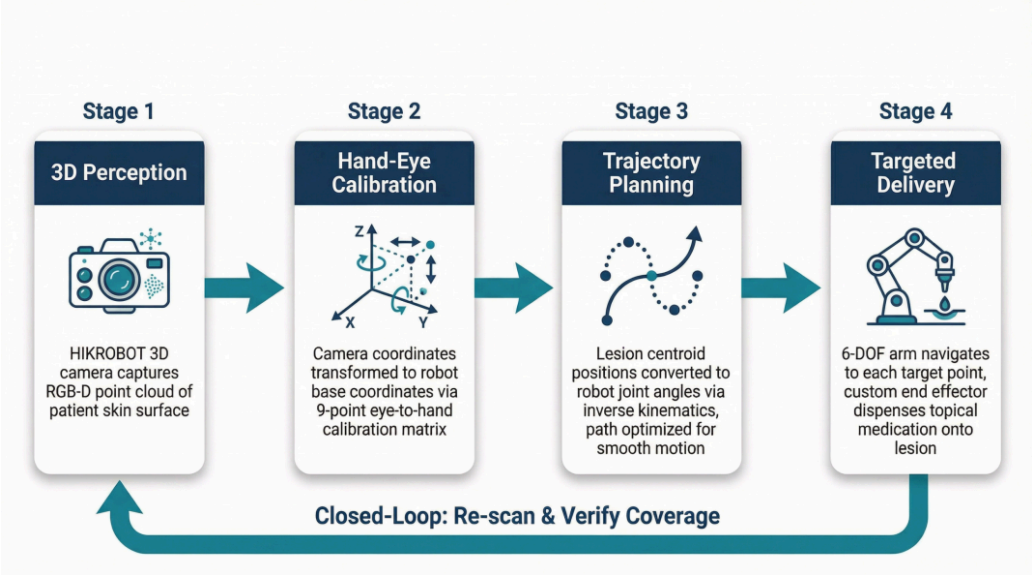

Figure 2. System pipeline from RGB-D capture to lesion segmentation, calibration, path planning, topical delivery, and parallel LLM-assisted reporting

3.2. Vision-based lesion segmentation

The segmentation module uses the U-Net design [9] with a ResNet-50 encoder. The encoder extracts a four-level feature pyramid; the decoder upsamples, concatenates skip connections, and outputs two logits (background vs.\ lesion) through a final 1\times1 convolution. The prediction is thresholded into a binary mask that is consumed by the planner.

Training uses 98 LabelMe-annotated urticaria photographs. The model is optimized with Dice loss, Adam at learning rate 10^{-4}, and a two-stage schedule: 50 frozen-backbone epochs followed by 100 full-network epochs. Augmentation includes scale jitter, horizontal flip, HSV jitter, and padded resizing to 512\times512. This is a small-data medical segmentation baseline rather than a new network architecture.

3.3. Hand-eye calibration

The benchtop task is approximated as a local planar workspace. Pixel coordinates (x,y) are mapped to robot task-space coordinates (X,Y) with an affine transform:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

Nine manually touched points on a planar calibration grid provide correspondences $\{(x_i, y_i) \to (X_i, Y_i)\}_{i=1}^{9}$. The six free parameters are estimated by least squares:

$$\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{A}\boldsymbol{\theta} - \mathbf{b}\|_2^2 \tag{2}$$

Calibration quality is reported as reprojection error:

DOI: 10.54254/2753-8818/2026.36040

$$\mathrm{RMSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(X_i - \widehat{X}_i\right)^2 + \left(Y_i - \widehat{Y}_i\right)^2} \tag{3}$$

The affine reduction is appropriate for the current near-planar phantom experiments; curved anatomy would require depth-aware SE(3) calibration such as classical hand-eye methods [10].

3.4. Coverage planning and hardware

Given a binary mask, the planner sweeps horizontal scan lines through the positive region using a configurable step size (default 20 px). Rows are traversed left-to-right and right-to-left in alternation, the standard boustrophedon pattern [11]. Gaps larger than the step size split a row into separate runs, so the path does not intentionally cross background between disconnected lesion components. The affine transform then converts the image path into arm waypoints.

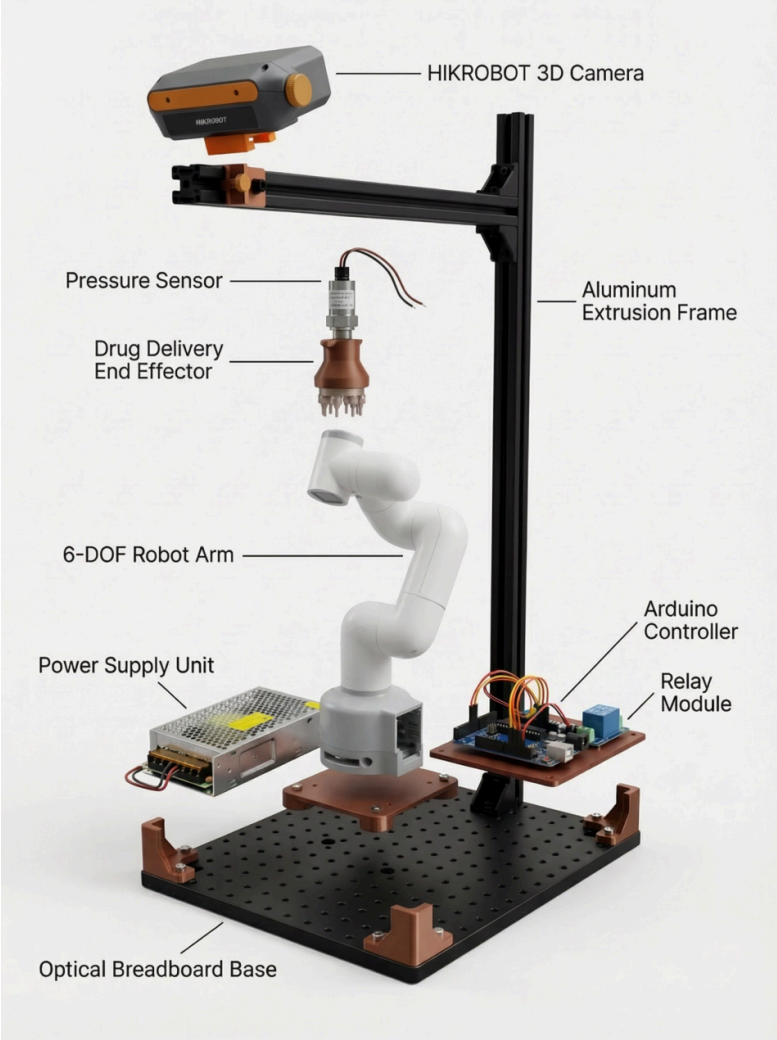

Figure 3. Benchtop platform: RGB-D camera, 6-DOF arm, custom multi-nozzle end-effector, pressure sensor, Arduino/relay control, and power electronics

The hardware platform in Figure 3 uses a commercial 6-DOF arm, a HIKROBOT RGB-D camera, a 3D-printed multi-nozzle end-effector, miniature peristaltic pumps, and an Arduino-controlled relay circuit. Safety is handled by three conservative checks: pressure-gated delivery,

DOI: 10.54254/2753-8818/2026.36040

software emergency stop on over-pressure, and a calibration pre-flight check that aborts if RMSE exceeds 1.5 mm.

3.5. LLM diagnostic interface

The diagnostic module adapts Llama-3.2-3B [16] with Low-Rank Adaptation (LoRA) [15]. Rank-8 adapters with α=16 are inserted into the q_proj and v_proj attention projections; the base model remains frozen, and roughly 4 M parameters are trained. Supervised fine-tuning uses a 2,752-pair augmented urticaria instruction corpus, maximum length 512, AdamW at learning rate 2\times10^{-4}, and 200 training steps.

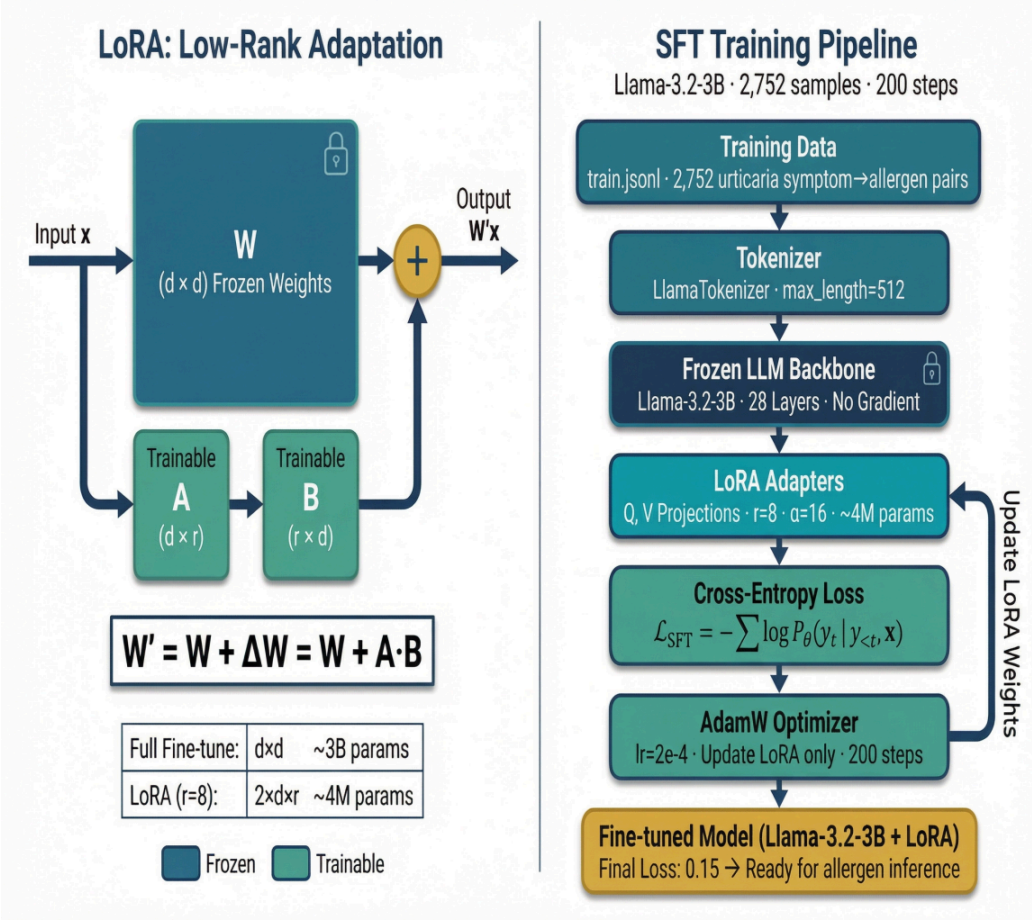

Figure 4. LoRA adaptation and supervised fine-tuning pipeline for the urticaria diagnostic interface

For longitudinal use, each patient has a JSON memory profile storing reported triggers, lesion sites, treatment outcomes, and resolution timelines. At inference time, a 7-field questionnaire and the patient memory are concatenated into the prompt, allowing the model to compare new symptoms with earlier visits and narrow possible triggers over multiple turns. Figure 8 illustrates the intended interaction pattern. Emergency red flags such as dyspnea or laryngeal edema trigger referral language rather than home-care advice.

DOI: 10.54254/2753-8818/2026.36040

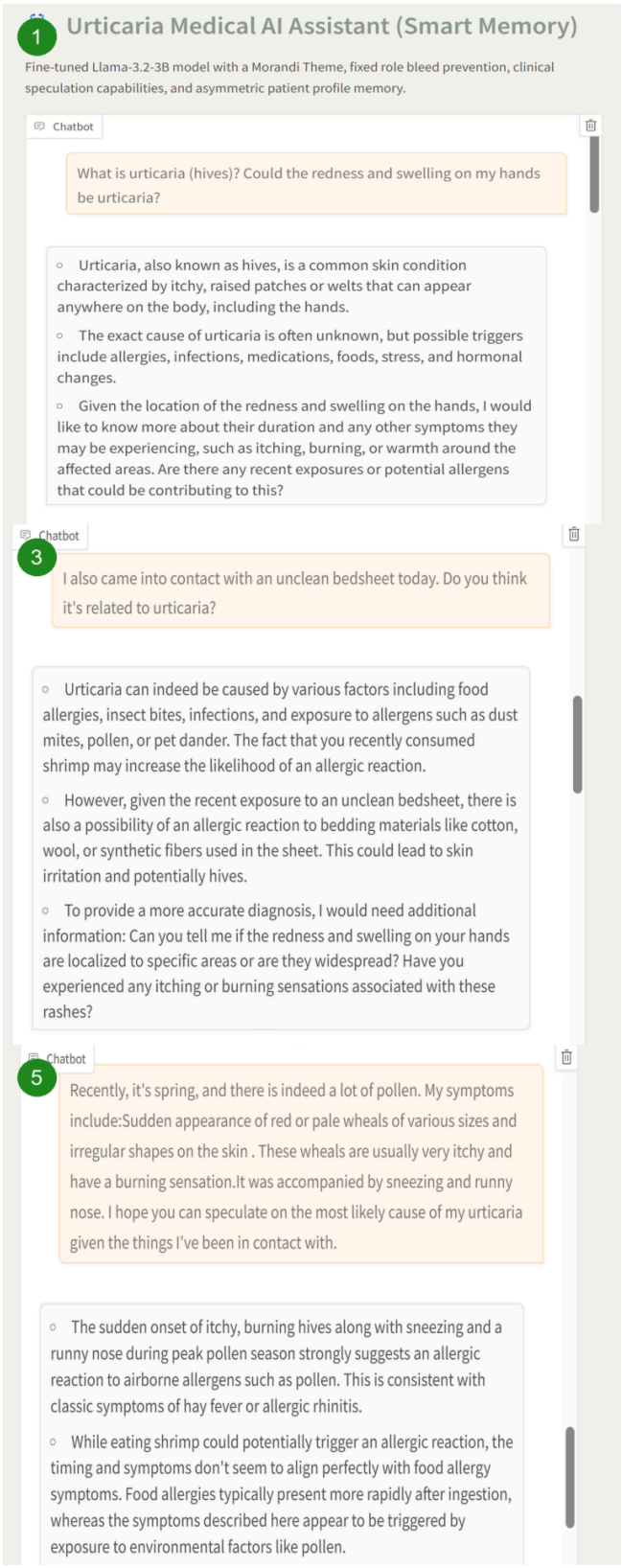

Figure 5. Representative multi-turn allergen-elimination dialogue, with the model updating candidate triggers as patient context accumulates

DOI: 10.54254/2753-8818/2026.36040

The LLM output is intended for triage and patient education. It is not used to command the arm in this version of the system.

4. Experiments

The experiments characterize the released feasibility artifact rather than benchmark every component against alternatives. All robotic tests use a medical-grade silicone arm phantom rather than live patients.

4.1. Experimental setup

The platform includes a myCobot 280-Pi 6-DOF manipulator, a Hikrobot MV-EB435i RGB-D camera, an Arduino Uno R3 relay controller, miniature peristaltic pumps, and a pressure-sensing multi-nozzle end-effector. The camera and arm are fixed to a shared optical breadboard to reduce calibration drift.

Two datasets are used. The lesion segmentation set contains 98 LabelMe-annotated urticaria photographs from the publicly available DermNet image library [17]. The LLM corpus contains 2,752 augmented urticaria instruction—output pairs spanning general, physical, pediatric, lifestyle, allergen-testing, and special-cause topics; a 396-pair sample is released with the code. Segmentation is implemented in PyTorch, calibration in NumPy, planning in deterministic Python, and LoRA fine-tuning in transformers with peft.

4.2. Vision module

The segmentation data are split into 68 training, 20 validation, and 10 test samples. On validation, the ResNet-50 U-Net achieves Dice 0.810 ± 0.083 and IoU 0.689 ± 0.118. On test, it achieves Dice 0.744 ± 0.114 and IoU 0.605 ± 0.140. Pooling the 30 held-out samples gives Dice 0.788 ± 0.099 and IoU 0.661 ± 0.132. Figure 4 shows representative predictions. The model handles compact wheals and moderate lighting variation, while blur and strong specular highlights remain common failure modes.

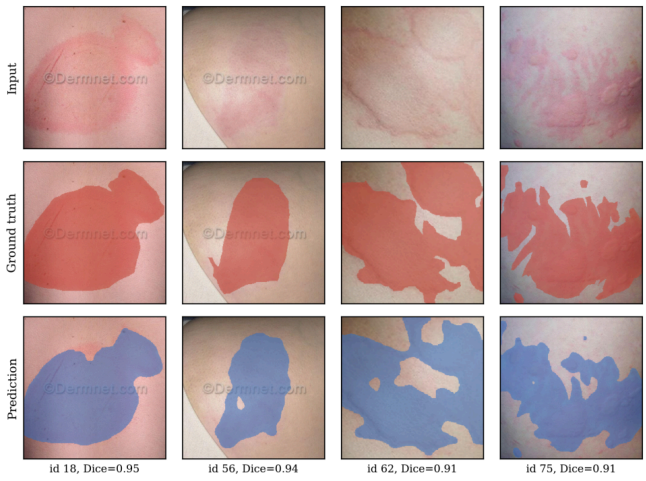

Figure 6. Held-out U-Net predictions: RGB input, ground-truth mask, and predicted mask with Dice scores

DOI: 10.54254/2753-8818/2026.36040

No head-to-head comparison with DeepLabv3+, Mask R-CNN, or Segment Anything is reported because the 98-sample dataset is too small for a stable architecture study. A larger multi-fold benchmark is left for future work.

4.3. Calibration and coverage

The 9-point affine calibration is run at setup. In a representative session with approximately 0.5 px corner-detection noise and a workspace scale near 0.5 mm/px, the released code reports a mean reprojection residual of 0.21 mm and a maximum residual of 0.31 mm, well below the 1.5 mm safety abort threshold.

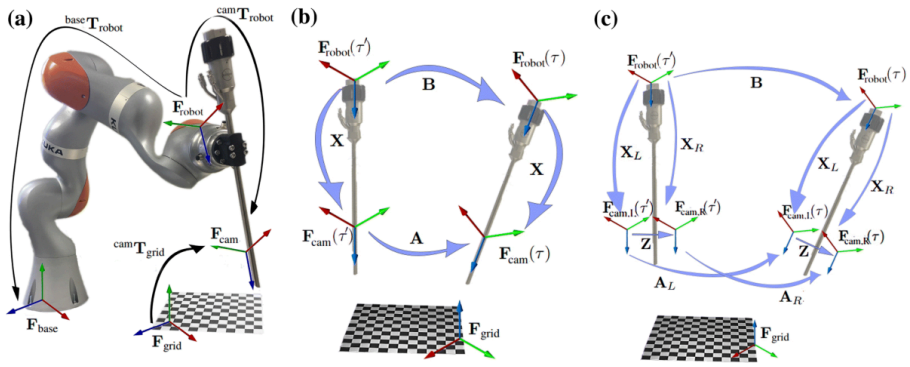

Figure 7. Calibration geometry and residual-check workflow for the planar affine hand-eye reduction

Figure 6 shows a representative boustrophedon coverage path on a segmented lesion mask. With the default 20 px step, every positive mask pixel lies within one step of a waypoint; the planner prioritizes complete coverage and deterministic behavior over path length optimality.

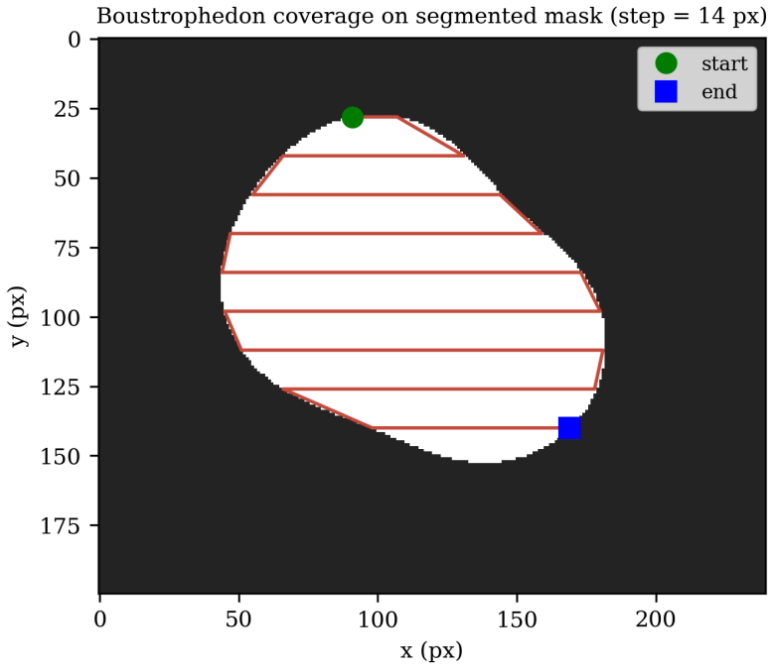

Figure 8. Boustrophedon coverage path over a predicted lesion mask

DOI: 10.54254/2753-8818/2026.36040

4.4. LLM module and end-to-end demonstration

The LoRA adapter (r=8, α=16, q_proj/v_proj) is trained for 200 steps on the 2,752-pair urticaria corpus and converges to cross-entropy loss 0.15. The memory-augmented dialogue interface can carry forward reported triggers across turns, as shown in Figure 8. No diagnostic accuracy score is claimed: that would require dermatologist-adjudicated ground truth on a larger case set.

Figure 9 summarizes the complete benchtop workflow: image capture, segmentation, calibrated path overlay, and topical delivery through the pressure-gated end-effector. A formal repeatability study measuring coverage percentage, application error, and session-level robustness is the next experimental step.

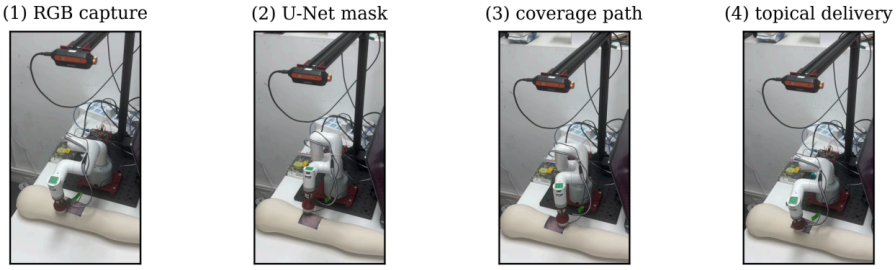

Figure 9. End-to-end demonstration: capture, segmentation, path overlay, and topical delivery

5. Discussion

The value of the system lies in integration rather than isolated module novelty. Compared with diagnostic dermatology models such as SkinGPT-4 [3], the proposed pipeline adds a physical topical-delivery path. Compared with procedure-specific dermatologic robots such as ARTAS [4], it operates on patient-specific lesion masks and includes an LLM diagnostic interface. Compared with standalone medical LLMs, it keeps language reasoning separate from deterministic robot planning. This separation is deliberate: the current artifact is a feasibility platform, not a validated closed-loop medical device.

5.1. Limitations

Dataset scale. The segmentation model is trained on only 68 images and evaluated on 30 held-out samples. The reported Dice scores should therefore be interpreted as same-distribution feasibility numbers, not as clinical generalization estimates. Skin-tone, age-group, lighting, and anatomical-site robustness remain unmeasured.

Clinical validation. The experiments are performed on a benchtop phantom, not live patients. The work has not undergone IRB review and makes no claim about dose-response behavior, lesion resolution, or superiority to manual application.

LLM validation. The LoRA-adapted LLM is not evaluated against dermatologist-adjudicated case labels. Its output should be treated as triage and patient-education support, not as a diagnosis. Hallucination and unsafe over-confidence remain risks for any deployed medical LLM.

Geometry assumptions. The affine calibration assumes a local planar working surface. High-curvature anatomy will require depth use, SE(3) calibration, and surface-aware planning.

Control coupling. The LLM does not modulate the robot's path, dose, dwell time, or re-treatment schedule. Such closed-loop coupling is future work and would require stricter safety validation.

DOI: 10.54254/2753-8818/2026.36040

5.2. Safety and reproducibility

The platform is framed as assistive topical delivery, not autonomous prescription. Red-flag symptoms such as dyspnea, laryngeal edema, or systemic anaphylaxis trigger referral language in the diagnostic prompt. At the hardware level, pressure sensing gates delivery, over-pressure causes emergency stop, and calibration RMSE above 1.5 mm blocks actuation until recalibration.

The released artifact includes the U-Net training code and weights, calibration script, boustrophedon planner, end-effector firmware, LLM prompting/fine-tuning framework, and urticaria knowledge resources.

6. Conclusion

This paper presented an integrated feasibility platform for autonomous topical urticaria therapy. The system combines ResNet-50 U-Net lesion segmentation, 9-point affine hand-eye calibration, boustrophedon coverage planning, pressure-gated 6-DOF topical delivery, and a LoRA-adapted Llama-3.2-3B diagnostic interface with patient memory. The released artifact achieves validation Dice 0.810 ± 0.083 on the 20-sample validation split, sub-millimeter calibration residuals in a representative session, and complete planned coverage of segmented lesion masks on a benchtop phantom.

The work should be read as a reproducible system-integration study, not as a clinical device validation. Future work will focus on IRB-approved patient studies, larger and more diverse lesion datasets, depth-aware SE(3) calibration for curved anatomy, quantitative repeatability testing, and true closed-loop coupling between diagnostic context and safe actuation parameters.

References

[1] Zhang, L., Wang, X., Yang, H., & others. (2023). Burden of and trends in urticaria globally, regionally, and nationally from 1990 to 2019: Systematic analysis. *JMIR Public Health and Surveillance, 9*, e50114.

[2] Zuberbier, T., Abdul Latiff, A. H., Abuzakouk, M., Aquilina, S., Asero, R., Baker, D., Ballmer-Weber, B., Bangert, C., Ben-Shoshan, M., Bernstein, J. A., & others. (2022). The international EAACI/GA^{2}LEN/EuroGuiDerm/APAAACI guideline for the definition, classification, diagnosis, and management of urticaria. *Allergy, 77*(3), 734–766.

[3] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

[4] Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., & others. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature Medicine, 26*(6), 900–908.

[5] Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., & Gao, X. (2024). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications, 15*(1), 5004.

[6] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., & others. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180.

[7] Avram, M. R., & Watkins, S. A. (2014). The ARTAS robotic hair restoration system. *Dermatologic Surgery, 40*(12), 1319–1327.

[8] Mu, G., Zhang, Y., Li, H., Wang, J., & Liu, H. (2022). Robotics meets cosmetic dermatology: Development of a vision-guided system for laser photo-rejuvenation. *IEEE/ASME Transactions on Mechatronics, 27*(5), 3528–3539.

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI* (pp. 234–241).

[10] Tsai, R. Y., & Lenz, R. K. (1989). A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. *IEEE Transactions on Robotics and Automation, 5*(3), 345–358.

DOI: 10.54254/2753-8818/2026.36040

[11] Choset, H. (2000). Coverage of known spaces: The boustrophedon cellular decomposition. *Autonomous Robots, 9*(3), 247–253.
[12] Rashid, T. G., Said, S., Kotwal, S., Stewart, L., & Ahmad, I. (2023). A review of the role of robotics in surgery: To DaVinci and beyond! *Cureus, 15*(9), e44973.
[13] Shademan, A., Decker, R. S., Opfermann, J. D., Leonard, S., Krieger, A., & Kim, P. C. W. (2016). Supervised autonomous robotic soft tissue surgery. *Science Translational Medicine, 8*(337), 337ra64.
[14] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus, 15*(6), e40895.
[15] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.
[16] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., & others (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
[17] Trust, D. N. Z. (2024). DermNet dermatology image library. Public dermatology image and information resource. \urlhttps://dermnetnz.org/