https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

Brief Communication: Structured Virtual Expert Panels for Interdisciplinary Ideation in Natural Hazard Science

Jui-Ming Chang[1,2], Nativ Ron[3], Qi Zhou[4], Shang Jyh Yiin[5]

[1]Department of Civil Engineering, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
[2]Disaster Prevention and Water Environment Research Center, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan
[3]Géosciences Rennes, University of Rennes, Rennes 6118, France
[4]GFZ Helmholtz Centre for Geosciences, Potsdam, Germany
[5]Data Decision.ai, Taipei 100, Taiwan

Correspondence to: Jui-Ming Chang (geomingical@gmail.com; aming@nycu.edu.tw)

Abstract. We present a virtual expert panel workflow for early-stage interdisciplinary ideation in natural hazard research. Using the Six Thinking Hats framework, a moderator questions multiple virtual experts to elicit points of consensus, contested assumptions, knowledge gaps, and follow-up clarifications. We demonstrate this workflow using a debris-flow monitoring case at the Illgraben, Switzerland. The panel highlights concept drift: year-to-year environmental change shifts the link between seismic signals and event labels and reduces machine-learning generalization. The output motivates ideas on deployment-realistic evaluation, including time-ordered data splits for training and testing and forward-chaining validation. The workflow provides a traceable record for hypothesis formulation.

1 Introduction

Scientific progress in Earth sciences relies on creativity, particularly the ability to integrate diverse lines of evidence into new hypotheses (Fisher, 1997). Recent studies argue that interdisciplinary teams must actively cultivate creativity to address complex Earth system challenges (Fazey et al., 2020). Environmental seismology exemplifies this integrative approach, bridging geophysics, hydrology, natural hazards and geomorphology to monitor Earth surface processes (e.g., Nativ et al., 2025; Chang et al., 2025b). The hazard monitoring field has recently adopted machine learning approaches to improve event detection and signal interpretation from continuous seismic records, especially when manual inspection becomes a bottleneck for large, continuous datasets (e.g., Chmiel et al., 2021). With machine learning increasingly used to replace manual screening at scale, early input from multiple specialists helps align research questions with evaluation design before model development begins. Without such discussion, key deployment assumptions can remain implicit.

Traditional expert panels, such as review boards, workshop committees, or ad-hoc advisory groups, remain invaluable for scientific evaluation, but may remain difficult to convene when researchers face tight time constraints (Rockwell, 2009). Early interdisciplinary exchange can therefore be limited, even when broad perspectives are important for framing research questions

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

and deciding how evidence will be evaluated. Recent advances in large language models now make it possible to construct virtual expert panels. Related ideation workflows have begun to show practical value in fields such as chemistry and biology (Zhang et al., 2025). In geoscience, comparable uses for early-stage interdisciplinary ideation and evaluation planning are still emerging, and practical examples remain limited. Here we propose a computational framework for early interdisciplinary dialogue using a virtual expert panel. This multi-agent system runs a moderator-led exchange among virtual experts, where a dedicated moderator agent posts round-specific questions and compiles the responses (Chen et al., 2025). Our workflow produces a transparent record of assumptions, perceived research gaps, and potential research directions that domain experts can review and refine before formal consultations. This study asks what inspectable outputs such a panel produces in practice, and whether the workflow can help formulate testable research ideas and deployment-relevant evaluation choices under temporal change.

2 Method

The framework is implemented as a multi-agent AI system (Code Availability). Users define the panel composition by specifying the number of virtual experts and assigning each expert a disciplinary role and perspective to approximate an interdisciplinary team. We illustrate the workflow using a field study case in the Illgraben catchment, Switzerland, a well-instrumented site frequently used for debris flow-related research and method development (Badoux et al., 2009). Using a widely-studied site provides a concrete context for interpreting the panel outputs and linking them to operational debris-flow monitoring questions.

We manually combine four virtual experts in our demonstration session: an engineering geologist, an environmental seismologist, a machine learning specialist, and a scientific editor (Supplement S1). We refer to each of these experts as a "panelist". A separate moderator agent, here implemented as a scientific editor, facilitates the discussion by issuing round-specific prompts and synthesizing the panel responses. Section 2.1 describes the input materials and initialization. Section 2.2 describes the panel discussion protocol. Section 2.3 describes the moderator synthesis and final outputs (Fig. 1).

2.1 Input Stage

In the input stage, all panelist agents receive background scientific materials and research questions that serves as the discussion's starting point. This step gives every agent common knowledge basis before any discussion begins. For the Illgraben case, the background material is a peer-reviewed debris-flow paper (Zhou et al., 2025), providing a common empirical and methodological reference for all panelists. As a demonstration, we specify a focused discussion goal as research questions: "Which machine learning strategies can improve debris-flow detection at the Illgraben using seismic measurements exclusively? *How should evaluation be designed when models trained on a few wet seasons must be generalized to future years as environmental conditions change?"* These questions set the topic and limit the discussion to detection strategy and evaluation under temporal change.

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

2.2 Panel Discussion Stage

The panel discussion follows a facilitation protocol based on the Six Thinking Hats (De Bono, 1985), using one hat per round to define a single, shared viewpoint for all panelists. A moderator agent (in this case, a journal Editor) leads each round by posting a hat-specific question to the full panel (Fig. 1). The moderator applies the six hats in a fixed order. The sequence includes White Hat for facts, evidence, and missing information, Red Hat for intuitive reactions and concerns, Purple Hat for unexplored directions, used in this study in place of the traditional Black Hat (De Bono, 1995), Yellow Hat for feasibility and value, Green Hat for creative alternatives and assumption challenges, and Blue Hat for process control, prioritization, and predefined evaluation criteria and decision thresholds. In each round, the moderator first sets the hat role for the discussion and posts the corresponding question. All panelists follow the same hat role when answering, so the round reflects a single shared perspective. The sequence is recorded as a transcript that the moderator later synthesizes.

2.3 Output and Follow-up Stage

After each round, the framework's main outputs are generated by the moderator agent. Based on the full set of round-by-round responses, the moderator then produces a synthesis summarizing recurring points of agreement, debated points, and potential research directions. In addition, the moderator performs a gap analysis that explicitly identifies what was not discussed, but is likely to matter for rigor, evaluation in later testing. For transparency purposes, the initial, independent opinions of each expert and the full sequence of hat-constrained rounds are also provided in the Supplement (Supplement S2), allowing the user to trace the evolution of the dialogue. A follow-up loop then allows the user to pose targeted clarification questions, and a final report consolidates the synthesis and follow-up insights into a concise set of research directions.

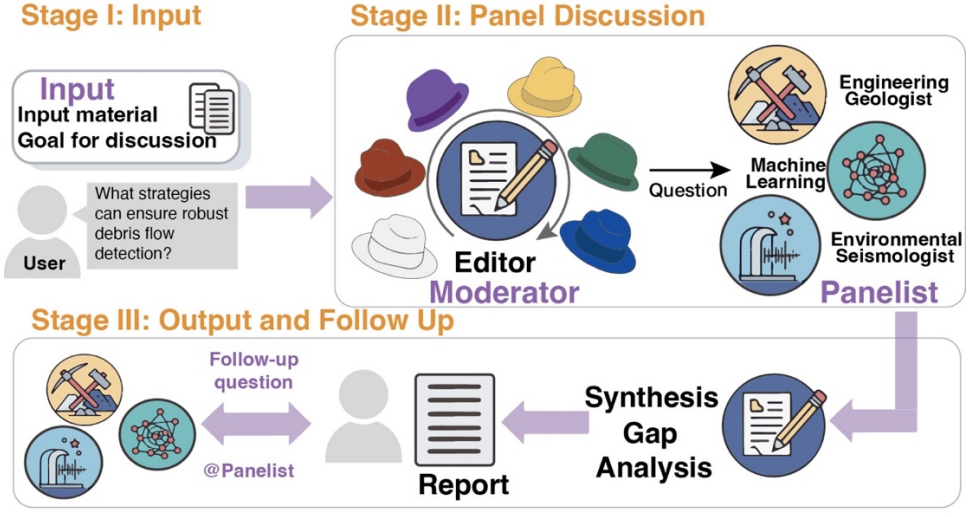

Figure 1. Structured workflow of the large language model (LLM)-based virtual expert panel. The framework operates in three stages. Stage I defines the user-provided input materials and discussion goals. Stage II conducts a moderated panel discussion

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

in which virtual experts rotate through six complementary analytical roles to examine the problem from factual, critical, constructive, and strategic perspectives. Stage III produces a moderator synthesis that summarizes consensus, unresolved divergences, and missing considerations, followed by user-driven follow-up questions and a consolidated report of ideation outcomes. Follow-up questions can be addressed to the full panel or to specific panelists by explicitly directing the query to one or more agents (e.g., "@Panelist").

3 Result

Our example case study produced a fixed set of outputs defined by the workflow. The moderator's report summarizes three shared points (topics raised consistently across panelists), three debated points (topics where panelists expressed contrasting views), and three missing topics (important issues not raised during the initial rounds). The moderator also recorded a targeted follow-up clarification on evaluation design under temporal change, and a final prioritized list of research directions. The full round-by-round record is provided in the Supplement S2. Table 1 summarizes how each output is treated as an inspectable workflow record and how scientist can use it as inputs for later testing, rather than as a standalone conclusion.

3.1 Panel outputs

The moderator synthesis reported three consensus points. First, all experts identified temporal generalization across years as the core risk that requires explicit testing. Second, they agreed that seismic-only detection approach is feasible in principle, but is sensitive to year-to-year environmental change that shifts seismic signal characteristics over time and can alter the relationship between seismic observations and debris-flow occurrence, thereby compromising the strength of debris-flow detection. Third, they agreed that Leave-One-Year-Out evaluation, one year is held out for testing and the remaining years are used for training, is a necessary baseline for cross-year generalization results.

The moderator synthesis also identified three debates. The first debate was whether seismic signals alone are sufficient for debris-flow detection, or whether additional sensors are needed for reliable deployment. The second debate was whether the choice of seismic features really matters for detection, and whether stronger links to physical mechanisms are needed for reliable models. The third debate focused on which strategy is best suited to sustaining performance over multiple years in a changing environment. Proposed solutions include anomaly detection with periodic threshold updates, learning representations intended to transfer across years, and approaches that incorporate physical source constraints. The supporting statements are provided in Supplement S2.

3.2 Panel outputs

The gap analysis identified three topics that were not addressed in the initial panel rounds. These are uncertainty and calibration when the same model is applied to following years with shifted seismic features (concept drift), drift monitoring and model

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

updating when labels are sparse (few confirmed debris-flow events), and modern time-series architectures with self-supervised baselines that are now common expectations in the field.

A follow-up discussion clarified how to evaluate models when performance may change over time. The panelists stated that all model development must use only data that occur before the test period. This includes training data, validation data for model selection, and threshold tuning. The panelists hence recommended forward-chaining evaluation, where for a given test year, models are trained on earlier years and tested on the immediately following year to track performance decay over time. By contrast, Leave-One-Year-Out can train on years that occur after the test year. This violates a prospective evaluation setting for operational use, because it allows information from later regimes to influence model development and can overstate performance on earlier years.

The final report consolidated the session into three prioritized directions. (1) Monitor performance over time: test the model on newer data using the same setup as deployment, to track performance decline. (2) a direction focused on drift-aware seismic-only detection with calibrated uncertainty and decision thresholds under shifting conditions. (3) proposing feature representations designed to transfer across years, using self-supervised learning and domain adaptation, and evaluated under chronological testing.

Table 1. Scientific functions of workflow outputs from the virtual expert panel

Stage	Output artifact	Product	Scientific function	Role of human experts
Stage I	Independent statements	Discipline-specific readings of the same input	Make initial assumptions explicit	Check domain accuracy and framing
Stage II	Hat-constrained rounds	Short, role-specific viewpoints across key dimensions	Ensure structured coverage	Flag missing physics or constraints
Stage III	Moderator synthesis	Explicit consensus and disagreement summary	Stabilize the problem landscape	Judge what is scientifically meaningful
Stage III	Gap analysis	Unaddressed topics and likely rigor issues	Identify scientist-facing concerns	Select gaps for follow-up or testing
Stage III	Follow-up clarification	Targeted resolution of contested points	Probe evaluation choices	Challenge outputs and refine questions
Stage III	Final report	Prioritized research directions	Testable study designs	Accept, reject, or redesign directions

4 Discussion

4.1 From panel outputs to testable hypotheses

The central scientific contribution of the virtual panel is at the stage of hypothesis formulation. The workflow does not assume that any single suggestion is correct. Instead, it generates a structured set of hypotheses and methodological opportunities that can be taken forward towards empirical validation. To place the panel outputs in context, we point to recent studies that pursue

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

similar methodological directions in debris-flow though machine learning technology for Illgraben. Recent studies have begun testing anomaly-detection approaches that treat debris flows as departures from background seismic activity (e.g., Kamper et al., 2025). Others have used unsupervised deep-learning to characterize and cluster debris flows and ambient noise (Huang, Z. et al., 2025). Work using modern transformer-based architectures (Liu et al., 2025; Song et al., 2025) for debris-flow detection provides another example of how ideas from the panel output are being explored in practice. The above cited studies were published after Zhou et al. (2025), which forms the input material used in the panel session. These studies do not establish the correctness of the panel suggestions. Instead, they indicate that several outputs are consistent with active research directions and can be tested empirically.

4.2 Concept drift as an organizing problem for operational validity

The virtual panel results point to a higher-level concern beyond model choice. The follow-up clarification makes evaluation design central under temporal environmental change. In this context, concept drift refers to time-dependent environmental change that alters the relationship between seismic observations and debris-flow labels, thereby weakening temporal generalization even when within-year performance appears stable. The panelists emphasized strict chronological splitting and forward-chaining as deployment-valid approaches. Experts also cautioned that cross-year protocols can be misleading when model development is influenced by data from later years.

This framing motivates concept drift as an organizing problem for debris-flow monitoring in environmental seismology. The issue is not whether drift exists, but whether evaluation is done chronologically, so reported performance matches deployment on future, unseen conditions. Guided by the panel outputs and follow-up clarifications, we adopted concept drift as the core message of a subsequent study using the same Illgraben dataset (Chang et al., 2025). The paper implemented a strictly chronological design: training on 2017-2018 data, validation/tuning on 2019 data, and testing on 2020 data. The analysis examined how key elements of an adaptive pipeline, including drift-aware feature selection, ensemble design, and post-processing, influence performance in the held-out test year.

4.3 Limits of AI outputs and scientific oversight

A virtual panel can widen the space of ideas, but it also inherits the limits of large language models. Hallucination remains a known risk, in which plausible statements are not grounded in evidence (Huang et al., 2025). Additionally, outputs can vary across model choices and prompts, and fluency is not evidence of correctness. For these reasons, the workflow requires domain experts to check claims against the input study and supporting materials and to challenge assumptions through targeted follow-up questions. In this framing, the virtual panel supports ideation, and researchers retain responsibility for verification and study design. Table 1 makes this division of labor explicit. The workflow produces inspectable artifacts, including the moderator synthesis, identified gaps, and follow-up clarifications, each paired with a defined role for human checking and revision.

Variability across large language model runs can limit reproducibility, because different prompts, sampling settings, or model versions can produce different answers. The same variability can also support exploration at the ideation stage, because

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

multiple runs can reveal alternative framings and competing assumptions. Domain experts can then compare the alternatives, discard unsupported claims, and select hypotheses for empirical testing. Recent discussions in geoscience and scientific workflows describe large language models as aids for hypothesis generation that require strong disciplinary oversight, rather than as autonomous decision makers (Hadid et al., 2024).

5 Conclusion

The Illgraben case demonstration shows how a virtual panel can externalize early interdisciplinary reasoning into an inspectable record that supports hypothesis formulation. The discussion converged on temporal generalization as a key constraint for seismic-based debris-flow detection, and follow-up clarified that evaluation and threshold tuning should respect temporal order under changing environmental conditions. The resulting research ideas remain hypotheses that require empirical testing, but several suggested directions align with recent methodological developments in debris-flow and environmental seismology. The workflow, therefore, defines a clear division of labour: the system generates documented intermediate outputs, and domain experts verify claims, refine questions, and design deployment-relevant evaluations. The same workflow can be transferred to other scientific problems by changing the input materials and the set of virtual experts. Further evaluation is needed to test performance and usability across sites, scientific questions, and disciplinary settings.

Code and data availability

The code used to implement the virtual panel workflow is publicly available at https://github.com/geomingical/AI_Brainstorm_sixhats No new observational data were generated for this study.

Author contributions

S.J. contributed to early conceptual discussions related to AI-supported virtual panels. J.M. developed the Six Thinking Hats–based workflow, implemented the code, and prepared the initial manuscript draft. J.M., R., and Q. jointly developed the narrative structure and interpretation. All authors reviewed and approved the final manuscript.

Acknowledgements

The virtual expert panel workflow was implemented using large language models to simulate role-based scientific discussion. The virtual expert personas were generated with Gemini 3 Pro Preview, and panel interactions were conducted using GPT-5.2 and Gemini 3 Pro Preview. These tools were used solely to support structured ideation and documentation. All interpretations, methodological decisions, and manuscript preparation were carried out by the authors.

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

Competing interests

The authors declare that they have no competing interests.

Financial support

J.-M. was supported by a postdoctoral fellowship from the Foundation for the Advancement of Outstanding Scholarship, Taiwan.

References

Badoux, A., Graf, C., Rhyner, J., Kuntner, R., and McArdell, B. W.: A debris-flow alarm system for the alpine illgraben catchment: design and performance. Natural hazards, 49, 517–539, 2009.

Chang, J. M., Zhou, Q., Tang, H., Turowski, J. M., and Ko, K.: Adaptive Machine Learning Framework for Debris Flow Monitoring in Nonstationary Environments in Illgraben, Switzerland, Engineering Geology (under reviewed). Available at SSRN: https://ssrn.com/abstract=5736972, 2025(a).

Chang, J.M., Yang, C.M*., Chao, W.A., Ku, C.S., Huang, M.W., Hsieh, T.C., and Hung, C.Y. Unraveling Landslide Mechanisms with Seismic Signal Analysis for Enhanced Pre-Survey Understanding. Nature Hazards and Earth System Sciences, 25, 451–466. https://doi.org/10.5194/nhess-25-451-2025, 2025(b).

Chen, N., Tong, Y., Wu, J., Duong, M. D., Wang, Q., Zou, Q., Hooi, B., and He, B.: Beyond Brainstorming: What Drives High-Quality Scientific Ideas? Lessons from Multi-Agent Collaboration. arXiv preprint arXiv:2508.04575, 2025.

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B. W., and Hibert, C.: Machine learning improves debris flow warning. Geophysical Research Letters, 48(3), e2020GL090874, 2021.

De Bono, E.: Six thinking hats. Penguin Books, 1985.

De Bono, E.: Serious creativity. The Journal for Quality and Participation, 18(5), 1995.

Fazey, I., Schäpke, N., Caniglia, G., Hodgson, A., Kendrick, I., Lyon, C., et al.:Transforming knowledge systems for life on Earth: Visions of future systems and how to get there. Energy research & social science, 70, 101724. https://doi.org/10.1016/j.erss.2020.101724, 2020.

Fisher, S. G.: Creativity, idea generation, and the functional morphology of streams. Journal of the North American Benthological Society, 16(2), 305-318, 1997.

Hadid, A., Chakraborty, T., and Busby, D.: When geoscience meets generative AI and large language models: Foundations, trends, and future challenges. Expert Systems, 41(10), e13654. https://doi.org/10.1111/exsy.13654, 2024.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1-55. https://doi.org/10.1145/3703155, 2025.

https://doi.org/10.5194/egusphere-2026-884
Preprint. Discussion started: 20 May 2026

Huang, Z., Yang, Z., Pang, B., Wu, Z., and Feng, L.: Characterizing and clustering debris flow and environmental noise seismic signals using unsupervised deep learning. Geophysical Journal International, 243(2), ggaf353, 2025.

Kamper, F., Walter, F., Paitz, P., Meyer, M., Volpi, M., and Salzmann, M.: Exploring seismic mass-movement data with anomaly detection and dynamic time warping. EGUsphere. https://doi.org/10.5194/egusphere-2025-3864, 2025.

Liu, Y., Li, S., Turowski, J. M., Tang, H., Ouyang, C., Guo, X., Xu, Q., Zhang, B., Yang, J., An, B. and Zhou, Q.: Debris Flow Early Warning Using the Patch Fourier Transformer. Authorea Preprints. DOI: 10.22541/essoar.176409475.52042815/v1, 2025.

Nativ, R., Turowski, J., Chang, J.M., Hovius, N., Yang, C.J., Chen, W.S., Chang, W.Y., and Laronne, J.: Stationary Boulders Increase River Seismic Frequency via Turbulence. Geophysical Research Letters 52, e2024GL113784, 2025.

Rockwell, S.: The FDP faculty burden survey. Research management review, 16(2), 29, 2009.

Song, Y., He, S., Chen, H., Zhang, Z., Liu, W., and Lyu, A.: Deep Learning Supports Cross-Regional Debris Flow Warning: A Case Study of Two Regions. Authorea Preprints. DOI: 10.22541/essoar.174440748.83873138/v1, 2025.

Wang, H., Du, X., Yu, W., Chen, Q., Zhu, K., Chu, Z., Yan, L., and Guan, Y. (2025): Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. Neurocomputing, 618, 129063, 2025.

Zhang, Y., Khan, S. A., Mahmud, A., Yang, H., Lavin, A., Levin, M., Frey, J., Dunnmon, J., Evans, J., Bundy, A., Dzeroski, S., Tegner, J., & Zenil, H. (2025) Advancing the Scientific Method with Large Language Models: From Hypothesis to Discovery, npj Artificial Intelligence, 2025 (preprint: arXiv:2505.16477). https://arxiv.org/abs/2505.16477

Zhou, Q., Tang, H., Hibert, C., Chmiel, M., Walter, F., Dietze, M., and Turowski, J. M.: Enhancing debris flow warning via machine learning feature reduction and model selection. Journal of Geophysical Research: Earth Surface, 130(4), e2024JF008094, 2025.