ISSN 2594-2352 · revistaaihc.mx

Nehnemi: Explainable AI Support for Rural Students' Mental Health

Nehnemi: Apoyo de IA para la salud mental de estudiantes rurales

Haziel Álvarez González[1,*], Norma Elva Chávez[2], Saiph Savage[3]

[1] *ORCID, Facultad de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México, México*
[2] *ORCID, Facultad de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México, México*
[3] *ORCID, Civic A.I. Lab, Northeastern University, Boston, Massachusetts, United States*

*** Corresponding author:** *haziel.alvarez@fi.unam.edu*

Abstract

Mental health care in Mexico is structurally scarce: 21 of its 32 states, home to 70% of the population, have one public-sector psychiatrist or fewer per 100,000 inhabitants, and rural schools have no counselor to receive a positive screen. Language models make conversational support inexpensive, but a model asked to assess psychological distress returns a judgment that can be neither reproduced nor inspected, and attaching an explanation to that judgment does not make it reliable. We present Nehnemi, a screening and companion system for students at rural upper-secondary and university campuses that separates classification from conversation. The General Health Questionnaire-12, scored by its published binary rule in application code, holds sole classification authority; a conversational companion offering emotional support and subject tutoring operates inside a result it can neither compute nor revise; and a printed card read by a low-cost kiosk extends the same result to schools without connectivity. Because the threshold belongs to the instrument rather than to the model, the system's response to a positive screen is a guarantee stated in advance and readable in a single function afterwards. The system has not been deployed with students and no therapeutic benefit is claimed. We name the underlying design pattern, bounded delegation with a visible authority boundary, and state its limits.

Keywords: Psychological distress screening; Mental Health; Rural education; Accessibility

Resumen

La atención en salud mental en México es estructuralmente escasa: 21 de sus 32 entidades, donde vive el 70% de la población, cuentan con un psiquiatra del sector público o menos por cada 100,000 habitantes, y en las escuelas rurales no hay personal de orientación que reciba un resultado positivo. Los modelos de lenguaje abaratan el acompañamiento conversacional, pero un modelo al que se le pide evaluar el malestar psicológico devuelve un juicio que no puede reproducirse ni inspeccionarse, y acompañarlo de una explicación no lo vuelve confiable. Presentamos Nehnemi, un sistema de detección y acompañamiento para estudiantes de bachillerato y universidad en zonas rurales que separa la clasificación de la conversación. El Cuestionario de Salud General de 12 reactivos, calificado con su regla binaria publicada directamente en el código, es el único con autoridad para clasificar; un acompañante conversacional que ofrece apoyo emocional y tutoría opera dentro de un resultado que no puede calcular ni modificar; y una tarjeta impresa, leída por un kiosco de bajo costo, extiende ese mismo resultado a escuelas sin conectividad. Como el umbral pertenece al instrumento y no al modelo, la respuesta del sistema ante un resultado positivo es una garantía enunciada de antemano y legible después en una sola función. El sistema no ha sido desplegado con estudiantes y no se reclama beneficio terapéutico. Nombramos el patrón de diseño subyacente, delegación acotada con una frontera de autoridad visible, y declaramos sus límites.

Palabras clave: Detección de malestar psicológico; Salud mental; Educación rural; Accesibilidad

 CC BY-NC-ND 4.0

1 Introduction

The shortage of mental health care in rural populations is documented and durable. A review of 30 studies across 11 countries finds the same barriers recurring from the health system down to the individual (Jiang et al., 2026), and a Cochrane review reports only limited reliable evidence that any intervention redistributes health professionals toward underserved areas (Grobler et al., 2015). Mexico is an acute instance: 21 of its 32 states, together covering 70% of the population, employ one public-sector psychiatrist or fewer per 100,000 inhabitants, with providers concentrated in cities (Barrón-Velázquez et al., 2024). We focus on Mexican rural students in high schools and universities, for whom a referral to a mental health professional means travel they cannot afford.

Language models have made it inexpensive to place some form of psychological support in front of a student who would otherwise have none. The difficulty appears as soon as the model is asked to judge a person's mental health rather than converse with them. A model prompted to assess psychological distress returns an assessment that can be neither reproduced nor inspected: queried twice, it may differ.

Explainable AI offers a familiar answer to problems of this shape: attach an explanation to the model's output. But conventional techniques, such as saliency maps, attention weights, and post-hoc natural-language rationales, describe a model's internal computation rather than guarantee that its output is correct or stable (Rudin, 2019), and a language model's self-narrated rationale can diverge from the process that produced its answer (Turpin et al., 2023). Explaining an unreliable judgment does not make the judgment reliable; it only makes the unreliability easier to read. For populations with no fallback if the judgment is wrong, that is not enough.

Research on interfaces for underserved communities in Latin America points to a related requirement: people acting through an automated system need to see the workflow that operates on them. Chiang et al. (2017) found that mobile money adoption depended on interfaces that allow users to visualize and understand the workflow through which their money travels to recipients, and Chiang et al. (2018) report that Mexican immigrants collaborating with their government required a comparable sense of control. That requirement was developed for financial flows, where the quantity in transit belongs to the user; it has not been extended to psychometric classification, where what the system computes is a claim about the user's mind. That is the gap: a student who needs distress flagged and acted on is asked to trust a judgment they cannot check.

Two further bodies of work frame the design. Human-computer interaction in rural Mexico and Latin America has repeatedly shown that methods developed elsewhere transfer poorly without adaptation: personas built with rural participants rather than for them (Cabrero et al., 2016), cookstoves co-designed with their users (Angel et al., 2015), Indigenous data practices treated as a foundation for technological self-determination (Guerrero Millán & Nissen, 2024; Guerrero Millán & Radio Tosepan Limakxtum, 2026), design research asked to return value to participants from the outset (Hanrahan et al., 2020), and barriers documented as specific to rural workers (Flores-Saviaga et al., 2020). That work rests on sustained fieldwork; ours does not, and we return to this among the limitations. Second, a review of artificial intelligence in psychotherapy counts emotional accompaniment and early symptom detection among its viable applications but warns that over-delegation produces an algorithm that minimizes urgency or exaggerates risk (Sotelo-Martín, 2025). The norm it defends is a division of labour: such a system may accompany, but the assessment remains with a clinician. What that literature does not settle is who assesses when no clinician is present.

The contribution is framed in terms of interaction rather than of the instrument. First, legibility: the student can reconstruct any classification from their own twelve answers, so the limit of the model's authority is something they can open on a screen rather than a property asserted in prose. Second, agency: the colour that leaves the building is chosen by the student from the set the classification makes available, so what reaches a public space is not a verdict assigned to them. Third, adaptation to infrastructure: the same guarantee holds when the connection fails, through hardware a school can maintain without specialised staff. These are the dimensions on which the design should be judged, and we return to them in the discussion.

1.1 Objective or Research Questions

The objective is to design and build a distress-screening system for rural students in which the classification is produced by a published, inspectable instrument rather than by a language model, and in which that boundary of authority is visible to the student and preserved when connectivity fails. The article addresses two questions: how the authority to classify can be assigned to a citable instrument while a conversational model operates inside the resulting classification, and how that same classification can reach a school without connectivity without disclosing the student's result to bystanders.

2 Method

This work reports the design and construction of a system, not an empirical study with users. No data were collected from students and no deployment was carried out. The design decisions rest on published sources: the validation literature of the screening instrument, prior work on interface legibility in Latin America, and the critique of post-hoc explanation in high-stakes decisions. The central property claimed here, that the classification is produced by the instrument and never by the language model, is verified by inspection of the source code rather than by measurement, because prompt construction is confined to a single function in a single service.

2.1 Participants, Materials, or Dataset

There were no human participants and no dataset. The materials are the Spanish version of the General Health Questionnaire-12 (Rocha et al., 2011), its published binary scoring rule (D. Goldberg & Williams, 1988), and the software built around them: a web service and database, an

interface for students, an integration with a commercial language model, and the specification of an offline component based on a printed card and a kiosk with a colour sensor. The kiosk is specified but not yet fabricated.

2.1.1 Instruments and Procedure

The system administers the twelve items of the questionnaire. Six are worded positively and six negatively, and each polarity carries its own response options, so the interface presents each item with the options belonging to it; a shared option list would score half the instrument backwards. Table 1 shows both option sets and the score assigned to each position. Scoring follows the binary GHQ method: the first two options of an item score zero and the last two score one, giving a total from 0 to 12, and three or more marks a probable case. The threshold is not ours. Ruiz et al. (2017) derived it by ROC analysis in Colombian samples, with a sensitivity of .80 and a specificity of .64, and it matches what D. P. Goldberg et al. (1997) found across the fifteen countries of the World Health Organization study.

The resulting classification reaches the conversational companion as part of its input, carried in the system role where the student's own turns cannot overwrite it. Above the threshold, the instruction to encourage help-seeking and to name the national helpline is appended by a conditional on the score, not by the model's reading of the conversation. Nothing the companion produces is written back as a classification: the stored result changes only when a new questionnaire is scored.

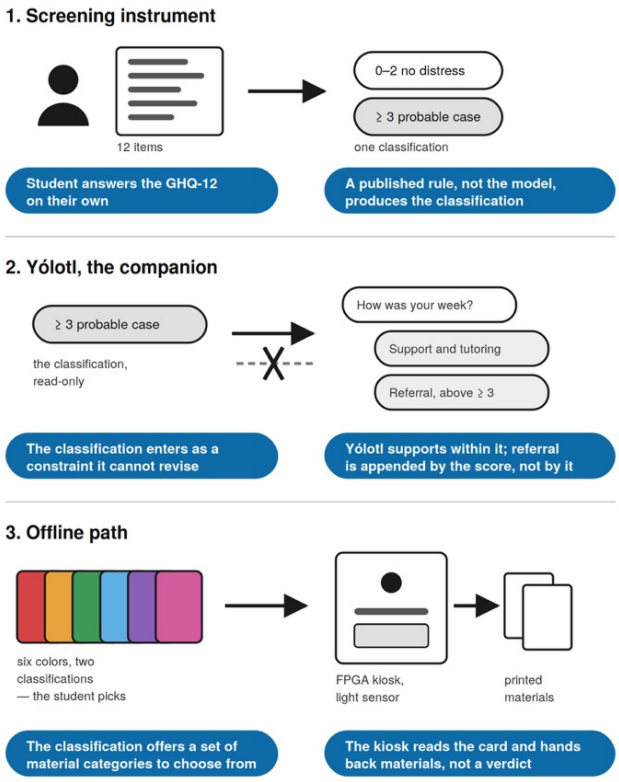


Figure 1. How the classification moves through Nehnemi.

Note. A published rule, not the model, produces the classification. The companion works inside it and never writes back, as the struck-through arrow marks. Offline, a coloured card and a kiosk carry the same classification into printed materials. Own elaboration.

Table 1. Response options and scores by item polarity.

Option	Score	Positively worded items (1, 3, 4, 7, 8, 12)	Negatively worded items (2, 5, 6, 9, 10, 11)
1	0	Better than usual	Not at all
2	0	Same as usual	No more than usual
3	1	Less than usual	Rather more than usual
4	1	Much less than usual	Much more than usual

Note. Binary GHQ scoring (0-0-1-1) over the twelve items, total 0 to 12. A shared option list across polarities would score half the instrument in reverse. Adapted from D. Goldberg and Williams (1988) and Rocha et al. (2011).

3 Results

The system was built and runs. A student consents, answers the twelve items, and receives a classification produced by a deterministic function of those twelve answers: the binary scoring rule described above, implemented in application code. The result is shown in words beside the total it came from, and every past evaluation expands to the

item answers behind its own total, so the student can reconstruct any classification they were given without asking the model to account for it.

The first question asked how classification authority could be assigned to a citable instrument while a conversational model operated inside the result. In the built system, the companion receives the classification in the system role, where the student's own turns cannot overwrite it, and nothing the companion produces is ever written back: the stored result changes only when a new questionnaire is scored. Above the threshold, the instruction to encourage help-seeking and to name the national helpline is appended by a conditional on the score rather than by the model's reading of the conversation. Because prompt construction is confined to a single function in a single service, this property is established by reading that function, not by sampling the model's behaviour. Two constraints operate here and they are not equally strong: the presence of the escalation instruction above the threshold is verifiable by inspection, whereas the tone rules carried in the same prompt, no diagnosis, no clinical vocabulary and no guidance on medication, are a request to a probabilistic system and are not guaranteed.

The second question asked how the same classification could reach a school without connectivity without disclosing the student's result to bystanders. The answer is a card whose colour is drawn from the set of material categories the classification makes available, with the student choosing among them. Six colours span two classifications, so the mapping cannot be run backwards: two students in different classifications may carry the same colour, and one student may carry different colours on different visits. A kiosk built on a field-programmable gate array reads the colour with a light sensor and exchanges the card for printed well-being and tutoring materials prepared in advance for that category. Distinguishing a handful of fixed colours requires only sensor circuitry and a lookup table in hardware, which avoids the cost, power draw and maintenance of an operating system in a school with limited technical support. This component is specified and has not been fabricated.

Regarding data handling, the system asks for consent before the first screening and states there that what the student writes is sent to an external model provider. Item-level answers and conversations are visible to the student and to the staff who operate the system; no teacher or school official is given access. A positive screen encourages help-seeking and names the national helpline. It does not book an appointment, notify an adult, or hand the result to a professional, because no professional is available to receive it.

4 Discussion

Nehnemi does not make an AI judgment explainable. It removes the judgment from the AI and gives it to something that was already legible, then keeps the model on the other side of that line. This answers the critique of post-hoc explanation from a direction the critique itself does not take. Rudin (2019) argues that for high-stakes decisions the right move is to decide with an interpretable model rather than to approximate an opaque one, and Turpin et al. (2023) show that a language model's stated reasoning can be steered by input features it never mentions. Both point away from explaining the model; neither specifies what should hold the decision instead. Here a published psychometric instrument holds it, and the model keeps the conversation.

The contribution to Human-Computer Interaction is not the instrument, which is decades old, but three properties of the interaction around it: the legibility of the classification to the person classified, the agency left to that person over what becomes visible to others, and the preservation of both when the infrastructure the design cannot assume is absent. Work on interfaces for underserved communities established that people acting through an automated system need to see the workflow that operates on them (Chiang et al., 2017, 2018), and that automatic techniques earn a bounded role beside social ones rather than in place of them (Johnson et al., 2021). In that literature the object computed is a resource whose ownership lies outside the system, such as money or bandwidth. Nehnemi carries the same requirement into a setting where the computed object is a claim about the user's own mind, and turns the boundary into something a student can open on a screen rather than something enforced only in code.

The design also inverts a common reading of a weak instrument. The specificity of .64 that Ruiz et al. (2017) report at this threshold usually counts against deployment. Here it is why the boundary exists: because the instrument is fallible, the system reports a probable case and offers a conversation rather than asserting a condition. A more accurate instrument would not change the architecture, only how often it is exercised.

The pattern generalizes past this instrument and past mental health. Wherever a decision carries an authority that can be cited, that authority can hold the decision while the model holds the conversation, with the seam between them in front of the person rather than buried in a service. The precondition is worth naming: it requires a published rule. Where none exists, where the judgment genuinely is the model's to make, the pattern offers nothing, and the honest conclusion is that such a system should not be making that judgment.

4.1 Limitations

Four limitations bear stating. First, the system has not been deployed with students, no usage data were collected, and no therapeutic benefit is claimed; the offline component is specified and has not been fabricated. Second, only part of the guarantee is verifiable by inspection: the escalation above the threshold is deterministic, but the tone rules carried in the same prompt are a request to a probabilistic system. Third, the design draws on published sources and expert guidance rather than on the sustained fieldwork that the literature it builds on rests upon (Cabrero et al., 2016; Guerrero Millán & Nissen, 2024; Hanrahan et al., 2020), and that literature is unanimous that instruments and interfaces developed elsewhere transfer poorly without adaptation. Fourth, the instrument's factor structure has been examined in Mexican adults, where a unidimensional

model fit best (Solís-Cámara Reséndiz et al., 2016), but the threshold used here was established on Colombian data and has not been re-derived in Mexico, nor in a student population of the age addressed. A further cost, accepted but not considered settled, is that the item-level answers of a minor reach an external model provider along with the conversation.

5 Conclusions

This article presented Nehnemi, a distress-screening and companion system for students at rural upper-secondary and university campuses in Mexico. The objective was to build a system in which the classification is produced by a published, inspectable instrument rather than by a language model, and in which that boundary of authority is visible to the student and survives the loss of connectivity. Both questions were answered by construction: a deterministic scoring rule implemented in application code holds sole classification authority and the conversational companion operates inside a result it can neither compute nor revise, while a card whose colour is chosen by the student from the categories the classification makes available carries the same result into a school without connection, without exposing it to bystanders.

The contribution is a pattern rather than an artefact: where a decision carries an authority that can be cited, that authority can hold the decision while the model holds the conversation, and the seam between them can be placed in front of the person instead of being enforced only in code. Its precondition is a published rule, and where none exists the pattern offers nothing. Installing the system in a school, fabricating the offline component, and studying their use with the communities addressed remain the necessary next steps. Each requires the participatory work that this design did not undertake, and it is there that the interaction claims made here become testable: whether students open the record of their own answers, and whether the choice of colour is experienced as control rather than as one more instruction.

Declarations

Author Contributions (CRediT)

Conceptualization: HAG, NECR, SS; **Investigation:** HAG; **Methodology:** HAG, SS; **Software:** HAG; **Visualization:** HAG; **Supervision:** NECR, SS; **Writing - original draft:** HAG; **Writing - review and editing:** HAG, NECR, SS.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Data, Code, and Materials Availability

No new data were generated or analyzed. No data were collected from students. The source code of the system is available from the corresponding author upon reasonable request.

Ethics Approval and Informed Consent

Not applicable. This work reports the design and construction of a system and did not involve human participants, data collection, or any form of study with students.

Consent for Publication

Not applicable. The interface images reproduce test data generated by the authors and do not contain information from any student.

Statement on the Use of Generative Artificial Intelligence

Generative artificial intelligence tools (Claude, Anthropic) were used during the preparation of this manuscript for drafting and revising text, structuring sections, locating candidate references, producing figures, and preparing the typesetting. They were not used to generate data, results, or the design decisions reported here. Every factual claim, figure, and reference was checked by the authors against its primary source; this verification corrected substantive errors during preparation, including an incorrect attribution of the instrument's validation population. The authors reviewed the full text and take full responsibility for its content.

References

Angel, W., Flores-Saviaga, C., & Savage, S. (2015). Participatory stoves: Designing renewable energy technologies for the rural sector. *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work and Social Computing (CSCW '15 Companion)*. https://doi.org/10.1145/2685553.2699021

Barrón-Velázquez, E., Mendoza-Velásquez, J. J., Mercado-Lara, A., Quijada-Gaytán, J. M., & Flores-Vázquez, J. F. (2024). The mental health provider shortage in the Mexican public sector: 2023 estimates of psychiatrists and psychologists. *Salud Mental*, *47*(4), 179–187. https://doi.org/10.17711/SM.0185-3325.2024.023

Cabrero, D. G., Prieto Álvarez, C. G., Moreno Rocha, M. A., & Abdelnour-Nocera, J. (2016). User-created personas in rural Mexico and in rural Spain: Approaches neither from the North nor from the South. *Avances en Interacción Humano-Computadora*, *1*(1), 13–17. https://doi.org/10.23427/avi.1.1

Chiang, C.-W., Anderson, C., Flores-Saviaga, C., Arenas Jr., E., Colin, F., Romero, M., Rivera-Loaiza, C., Chávez, N. E., & Savage, S. (2017). Understanding interface design and mobile money perceptions in Latin America. *Proceedings of the 8th Latin American Conference on Human-Computer Interaction (CLIHC '17)*. https://doi.org/10.1145/3151470.3151473

Chiang, C.-W., Betanzos, E., & Savage, S. (2018). Blockchain for trustful collaborations between immigrants and governments.

 CC BY-NC-ND 4.0

Flores-Saviaga, C., Li, Y., Hanrahan, B., Bigham, J., & Savage, S. (2020). The challenges of crowd workers in rural and urban America. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*(1), 159–162. https://doi.org/10.1609/hcomp.v8i1.7475

Goldberg, D., & Williams, P. (1988). *A user's guide to the general health questionnaire*. NFER-Nelson.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, *27*(1), 191–197.

Grobler, L., Marais, B. J., & Mabunda, S. (2015). Interventions for increasing the proportion of health professionals practising in rural and other underserved areas. *Cochrane Database of Systematic Reviews*, (6), CD005314. https://doi.org/10.1002/14651858.CD005314.pub3

Guerrero Millán, C., & Nissen, B. (2024). Cosmovision of data: An Indigenous approach to technologies for self-determination. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. https://doi.org/10.1145/3613904.3642598

Guerrero Millán, C., & Radio Tosepan Limakxtum. (2026). Digital territories: Entwining technology with Indigenous knowledges. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '26)*. https://doi.org/10.1145/3772318.3791011

Hanrahan, B. V., Ma, N. F., Betanzos, E., & Savage, S. (2020). Reciprocal research: Providing value in design research from the outset in the rural United States. *Proceedings of the 2020 International Conference on Information and Communication Technologies and Development (ICTD '20)*, 1–5. https://doi.org/10.1145/3392561.3397585

Jiang, Y., Lin, K., Dalton, H., Buys, N., & Sun, J. (2026). Assessment of barriers to accessing mental health services in rural or remote areas using a socio-ecological resilience framework: A scoping review. *International Journal of Mental Health Systems*, *20*, 3. https://doi.org/10.1186/s13033-026-00694-0

Johnson, M. W., Jang, E. H. B., Rodriguez, F., Chen, J. L., & Sammet, S. (2021). Network capacity as common pool resource: Community-based congestion management in a community network. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 61. https://doi.org/10.1145/3449135

Rocha, K. B., Pérez, K., Rodríguez-Sanz, M., Borrell, C., & Obiols, J. E. (2011). Propiedades psicométricas y valores normativos del General Health Questionnaire (GHQ-12) en población española. *International Journal of Clinical and Health Psychology*, *11*(1), 125–139.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ruiz, F. J., García-Beltrán, D. M., & Suárez-Falcón, J. C. (2017). General Health Questionnaire-12 validity in Colombia and factorial equivalence between clinical and nonclinical participants. *Psychiatry Research*, *256*, 53–58. https://doi.org/10.1016/j.psychres.2017.06.020

Solís-Cámara Reséndiz, P., Meda Lara, R. M., Moreno Jiménez, B., & Juárez Rodríguez, P. (2016). Estructura factorial del cuestionario de salud general GHQ-12 en población general de México. *Salud & Sociedad*, *7*(1), 62–76.

Sotelo-Martín, J.-A. (2025). Psicoterapia e inteligencia artificial: Diálogos clínicos potenciados por algoritmos artificiales para humanizar el cuidado emocional. *Revista Latinoamericana de Difusión Científica*, *7*(13), 138–161. https://doi.org/10.5281/zenodo.15843790

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965.

 CC BY-NC-ND 4.0



© by the authors. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.