PROGRAMA DE EMPRESARIADO SOCIAL

SÍNTESIS DEL PROYECTO

1. **País:** Paraguay
2. **Nº Proyecto**: PR-T1373
3. **Nombre Proyecto:** GuaranIA: integrando el idioma guaraní en el ámbito digital para la inclusión de poblaciones rurales y vulnerables
4. **Agencia Ejecutora:** Centro de Ingeniería para la Investigación, Desarrollo e Innovación Tecnológica (en adelante, CIDIT).
5. **Unidad del BID**: Fondo Multilateral de Inversiones (FOMIN) – IDB Lab

6. **Montos de Financiamiento**

	BID US$	Local US$	Total US$
Cooperación Técnica No Reembolsable	632.400	303.600	936.000

7. **Objetivo y propósito del proyecto:**

El objetivo del proyecto es promover el desarrollo de servicios basados en Inteligencia Artificial generativa en lengua guaraní orientados a poblaciones rurales/vulnerables, mediante el desarrollo de datos, el entrenamiento de modelos y la innovación abierta.

8. **Componentes del proyecto:**

El proyecto se estructura en cuatro Componentes:

Componente I: Diagnóstico para identificación de casos de uso y áreas prioritarias: El objetivo del componente es identificar potenciales casos de uso en los cuales la habilitación de tecnologías orientadas a la comunicación en guaraní permita a poblaciones en situación de vulnerabilidad, el acceso a productos o servicios que mejoren su calidad de vida.

En este sentido se propone realizar un diagnóstico, con la aplicación de técnicas de relevamiento de datos en campo como entrevistas, encuestas, grupos focales, talleres para identificar, a partir de la experiencia de las comunidades locales, problemáticas donde la carencia de herramientas tecnológicas con soporte para el idioma guaraní representa una barrera de acceso a oportunidades que puedan mejorar las condiciones de vida de las poblaciones de estas comunidades. Adicionalmente, se buscará que estos sectores estén alineados con las temáticas de otras operaciones del Banco relacionadas con procesos de transformación digital que beneficien a la población vulnerable en Paraguay. Como resultado del proceso de diagnóstico, se identificarán los sectores de mayor impacto sobre la calidad de vida de la población, en los que la implementación de casos de uso con recursos lingüísticos computacionales en el idioma guaraní puede generar mejoras en su acceso a productos y servicios en sectores de alto impacto. Preliminarmente se ha identificado oportunidades en ámbitos de educación, salud e inclusión financiera

El diagnóstico se realizará en tres (3) de los departamentos con mayor incidencia en pobreza

multidimensional en el país: Caazapá, Caaguazú y Alto Paraná, buscando incluir territorios adyacentes donde habitan comunidades indígenas guaraní hablantes, tales como la comunidad Guaira. El resultado de esta intervención será la de aportar conocimientos para entender el contexto de poblaciones rurales, indígenas y otros actores que oriente el desarrollo de la solución tecnológica propuesta.

En el marco del componente se buscará también vincular a la población beneficiaria, en sectores directamente relacionados con su calidad de vida, con el uso de herramientas de IA, de manera de familiarizarlos con la tecnología, mejorando de esta forma su percepción respecto de la IA y la usabilidad de las herramientas a ser desarrolladas en el marco del Componente II. Para ello, se generarán alianzas con: (i) centros educativos para la formación de docentes en el uso de herramientas de IA; (ii) unidades de salud familiar y otros centros de salud para la formación de personal de salud y usuarios; (iii) entidades financieras; (iv) gobiernos locales; (v) entidades académicas; (vi) empresas que puedan facilitar la adopción, despliegue de las herramientas y programas de formación.

Los productos esperados del componente son: (i) Diagnostico para la definición de sectores/áreas prioritarias para el desarrollo de pilotos de IA, terminado; (ii) Estrategia de comunicación del proyecto implementada; (iii) Al menos 12 alianzas formalizadas con entidades del sector público, privado, academia y sociedad civil para la colaboración en la implementación del proyecto; (iv) Documento de sistematización del proyecto y de análisis de resultados, completado.

Componente II: Desarrollo de herramientas tecnológicas basadas en IA para el procesamiento de lenguaje guaraní: El objetivo del componente es desarrollar recursos lingüísticos computacionales que faciliten la inclusión de la población guaraní hablante a la economía digital.

Como primera acción se trabajará en desarrollar un corpus digital de texto en guaraní, el cual será documentado y publicado a través de herramientas como datasheets for datasets propuestas por algunos investigadores. Para ello, se buscará: (i) generar alianzas con entidades públicas, académicas y privadas abordando la temática lingüística / guaraní; (ii) recolectar documentos de texto escritos en idioma guaraní mediante técnicas de webscraping, consultas a APIs de plataformas digitales y material digital escrito y publicado en Internet; (iii) digitalizar bibliografía escrita y transcripción de audios en guaraní; (iv) aplicación de técnicas de participatory AI para la recolección de datos, siguiendo las mejores prácticas de experiencias similares, en los que se involucrará a los mismos miembros de las comunidades identificadas en el trabajo de diagnóstico y relevamiento del componente I, de manera de validar y limpiar los datos recolectados en guaraní; (v) En caso de ser necesario, se valorará también la posibilidad de recurrir a la traducción de textos en otros idiomas, la generación de datos sintéticos a partir de textos reales, y la reutilización de base de datos de texto existentes y de dominio público como, por ejemplo, Jojajovai; (vi) revisión continua por parte de lingüistas expertos en el idioma para la anotación y validación del material procesado.

El corpus digital servirá como base para el entrenamiento de modelos de procesamiento de lenguaje natural (PLN) adaptados al guaraní, utilizando arquitecturas avanzadas y modelos fundacionales de IA que ya tienen una comprensión sólida del castellano y otros idiomas, resultando beneficioso para el guaraní paraguayo. Como resultado, los modelos serán ajustados para que sigan instrucciones en guaraní y Jopará (variante mixta con castellano), optimizando así su capacidad para realizar tareas específicas como la traducción, generación de texto o clasificación y su interacción dentro de diversas aplicaciones, como chatbots o asistentes virtuales.

También se evaluará el desempeño del modelo en entornos de código mixto (guaranícastellano) con el objetivo de manejar de manera eficiente el Jopará. Finalmente, todos los modelos y herramientas desarrolladas se liberarán en repositorios de código abierto bajo licencia GPLv3, facilitando su acceso para desarrolladores e investigadores locales e internacionales comprometidos con la inclusión digital

de la comunidad guaraní parlante. Los modelos serán publicados siguiendo prácticas comunes en el área como, por ejemplo, el uso de model cards para reportar su documentación.

Tanto durante la construcción del corpus como en las fases de entrenamiento y evaluación se aplicarán técnicas del estado del arte, recomendaciones y buenas prácticas que promuevan el uso responsable de las herramientas y ayuden evitar resultados inapropiados por parte de los modelos, incluyendo alucinaciones, así como también la producción de contenidos racistas, misóginos, discriminatorios, u ofensivo que amplifique o refuerce sesgos sociales. Asimismo, se buscará mitigar el impacto ambiental del desarrollo y operación de la tecnología incorporando, toda vez que sea posible, prácticas computacionales energéticamente eficientes, el uso de fuentes de energía renovable y desarrollo modular, evitando consumo energético innecesario.

Los productos esperados del componente son: (i) Corpus digital en texto guaraní validado, documentado y publicado bajo licencia GPLv3; (ii) Modelos de PLN entrenados en base al corpus desarrollado; (iii) Reporte de desarrollo responsable de corpus y modelos PLN documentado y validado.

Componente III: Implementación del uso de la herramienta IA a través de pruebas piloto en las comunidades: El objetivo de este componente es implementar en productos de software los recursos lingüísticos computacionales desarrollados en el componente II, ejecutando casos de uso demostrativos que beneficien la calidad de vida de las poblaciones vulnerables y población indígena identificadas en el componente I.

Sobre la base del corpus general creado en el componente II, se desarrollará, documentará y publicará corpus de texto específicos para cada uno de los casos de uso identificados, incorporando textos en guaraní relacionados a las temáticas y, más específicamente, a los casos de uso. Como resultado, se crearán modelos de generación de lenguaje “especializados” en los dominios relacionados a los casos de uso identificados en el componente I. Estos modelos especializados serán publicados bajo la misma licencia y siguiendo las mismas prácticas de documentación del modelo creado en el componente II.

Los modelos especializados serán integrados a módulos de software que faciliten su aplicación por terceros, pero principalmente su uso en los casos de uso identificados en el componente I. Estos módulos serán desplegados, mantenidos, y actualizados en servidores accesibles a través de Internet, con alta capacidad computacional y de almacenamiento. El código fuente de los módulos será disponibilizado en repositorios de código abierto bajo licencia GPLv3.

También los trabajos de implementación de este componente incluirán la integración de los módulos de software en el desarrollo de aplicaciones móviles orientadas al uso en dispositivos de baja gama en términos de capacidad de procesamiento, almacenamiento, resolución, y versión del sistema operativo. Así también se requerirá que estas aplicaciones funcionen con mínima dependencia de Internet, y que puedan ser operadas con mínimo entrenamiento por usuarios no expertos en tecnología. Los requerimientos funcionales de las aplicaciones dependerán de los casos de uso identificados en el componente I. El desarrollo seguirá un enfoque ágil como el Lean Software Development mientras que el código fuente desarrollado será disponibilizado en repositorios de código abierto bajo licencia MIT, permitiendo tanto el uso, mejoramiento, y evaluación sin costo por parte de la comunidad técnica-científica, así como también la explotación en propiedad por parte del sector privado.

El despliegue de las aplicaciones se realizará en forma gradual en tres fases. En una primera fase se realizarán pruebas técnicas/funcionales en entornos controlados, esto es dentro del equipo de desarrollo. Las aplicaciones serán corregidas y mejoradas en base a los resultados de esta fase.

En la segunda fase se llevarán adelante pruebas de usabilidad en entornos reales siguiendo la metodología moderada híbrida, incluyendo sesiones presenciales y a distancia. Las correspondientes

mejoras de usabilidad se aplicarán al finalizar esta etapa.

La tercera fase se enfocará en estudios piloto acotados donde se buscará entender el impacto potencial de la solución en la población objetivo, el cual será medido por medio de estudios experimentales. Para el efecto, se podrán realizar colaboraciones con actores de las áreas prioritarias identificadas en el Componente I y con presencia en el territorio acompañando potenciales poblaciones objetivo. La definición de los elementos del estudio de impacto como el grupo de prueba y control o línea de base28, las hipótesis, el alcance, las condiciones experimentales, las variables dependientes e independientes, y las formas de medirlas, dependerá del tipo de intervención que se realice dentro de cada uno de los casos de uso y poblaciones beneficiarias.

Los productos esperados del componente son: (i) Al menos 3 acuerdos para la implementación de casos de uso formalizados; (ii) Tres modelos de generación de lenguaje "especializados" en los dominios relacionados a los casos de uso identificados en el componente, desarrollados; (iii) Tres casos de uso piloto implementados y evaluados.

Componente IV: Estrategias de innovación abierta y capacitación para el desarrollo de productos y/o servicios: El objetivo del componente es identificar instituciones adicionales, públicas y privadas, que adopten las herramientas desarrolladas por el proyecto, aumentando el alcance en términos de productos, servicios y aplicaciones que aprovechan la tecnología, y ampliando el número de personas y comunidades que acceden a los recursos lingüísticos de GuaranIA.

En el marco de la ejecución del componente se organizará una o más convocatorias para la integración de las soluciones tecnológicas desarrolladas en el marco de GuaranIA a servicios y aplicaciones en los que la inclusión del idioma guaraní represente una oportunidad para que la población guaraní hablante pueda acceder, en igualdad de condiciones, a productos, servicios y nuevos casos de uso ofrecidos tanto por el sector público como el sector privado.

Las convocatorias de innovación abierta considerarán tres líneas de trabajo para la adopción de las herramientas de GuaranIA: (i) Línea A: Desarrollo de nuevos casos de uso o nueva aplicación, ejecución de experimento para su adopción y retroalimentación por parte de la población objetivo; (ii) Línea B: Integración de una aplicación GuaranIA existente adaptándola para la provisión de un servicio, ejecución de experimento para su adopción y retroalimentación por parte de la población objetivo; (iii) Línea C: Otros casos de uso. Línea abierta a propuestas que no se ajusten a las líneas A o B, pero que sean relevantes para el alcance de la convocatoria. Las convocatorias estarán orientadas a startups, empresas, sector público, academia, centros de investigación y organizaciones de la sociedad civil. Preliminarmente se han identificado algunos criterios de selección de las entidades participantes de la(s) convocatorias tales como: viabilidad, porcentaje de cofinanciamiento de los proponentes, potencial de impacto social, perspectivas de sostenibilidad y escala, relevancia del producto o servicio en la calidad de vida de la población vulnerable guaraní hablante. En el marco del componente también se trabajará una estrategia de escalamiento para la sostenibilidad de los productos tecnológicos que surjan del proyecto GuaranIA en el tiempo.

Finalmente, se buscará generar capacidades en actores del sector público y privado para el desarrollo y adopción responsable de la IA, con módulos de profundización para agentes de innovación, investigación y desarrollo y ecosistema de startups de impacto. Para ello, se implementarán ciclos de formación en temáticas tales como: (i) principios de desarrollo responsable en proyectos de IA; (ii) evaluación de riesgos en proyectos de ciencia de datos e IA; (iii) integración tecnológica de herramientas de IA. Como resultado indirecto de los ciclos formativos, se buscará conformar un grupo impulsor multidisciplinar sobre el uso responsable de la IA en el país.

Los productos esperados del componente son: (i) Convocatoria de innovación abierta para la inclusión del idioma guaraní en la economía digital, implementada; (ii) Al menos tres de integraciones

de la tecnología desarrolladas, implementadas y evaluadas; (iii) 60 personas capacitadas para la adopción responsable de herramientas desarrolladas e integradas; (iv) Estrategia de sostenibilidad y escala de uso e integración de los productos desarrollados

9. **Beneficiarios del proyecto**:

Serán beneficiarios del proyecto, al menos 500 las personas vulnerables usuarias de la implementación de los casos de uso desarrollados en el marco del proyecto, en sectores que impactan directamente en su calidad de vida, tales como educación, salud, servicios financieros, etc.

10. Resultados esperados y captura de beneficios:

Los resultados esperados a partir de este proyecto serán: (a) Al menos 500 personas vulnerables que se benefician de la implementación de los casos de uso desarrollados en el marco del proyecto; (b) herramienta tecnológica del lenguaje e inteligencia artificial para el idioma guaraní, desarrollada, publicada y disponibilizada; (c) al menos tres nuevos servicios ofrecidos a poblaciones bajo la Incidencia de Pobreza Multidimensional (IPM) ubicados en las zonas rurales y periurbanas del Paraguay; (d) al menos 3 iniciativas en las que las soluciones tecnológicas desarrolladas en el marco de GuaranIA se integran a nuevos servicios y aplicaciones para la inclusión del idioma guaraní en la economía digital.